International Journal on Advances in Intelligent Systems

2021 vol. 14 nr. 1&2

The *International Journal on Advances in Intelligent Systems* is Published by IARIA.
ISSN: 1942-2679
journals site: http://www.iariajournals.org
contact: petre@iaria.org

Sponsored by IARIA
www.iaria.org

International Journal on Advances in Intelligent Systems
Volume 14, Number 1 & 2, 2021

Karsten Böhm, FH Kufstein Tirol - University of Applied Sciences, Austria
Pierre Borne, Ecole Centrale de Lille, France
Christos Bouras, University of Patras, Greece
Anne Boyer, LORIA - Nancy Université / KIWI Research team, France
Stainam Brandao, COPPE/Federal University of Rio de Janeiro, Brazil
Stefano Bromuri, University of Applied Sciences Western Switzerland, Switzerland
Vít Bršlica, University of Defence - Brno, Czech Republic
Dumitru Burdescu, University of Craiova, Romania
Diletta Romana Cacciagrano, University of Camerino, Italy
Kenneth P. Camilleri, University of Malta - Msida, Malta
Paolo Campegiani, University of Rome Tor Vergata , Italy
Marcelino Campos Oliveira Silva, Chemtech - A Siemens Business / Federal University of Rio de Janeiro, Brazil
Ozgu Can, Ege University, Turkey
José Manuel Cantera Fonseca, Telefónica Investigación y Desarrollo (R&D), Spain
Juan-Vicente Capella-Hernández, Universitat Politècnica de València, Spain
Miriam A. M. Capretz, The University of Western Ontario, Canada
Massimiliano Caramia, University of Rome "Tor Vergata", Italy
Davide Carboni, CRS4 Research Center - Sardinia, Italy
Luis Carriço, University of Lisbon, Portugal
Rafael Casado Gonzalez, Universidad de Castilla - La Mancha, Spain
Michelangelo Ceci, University of Bari, Italy
Fernando Cerdan, Polytechnic University of Cartagena, Spain
Alexandra Suzana Cernian, University "Politehnica" of Bucharest, Romania
Sukalpa Chanda, Gjøvik University College, Norway
David Chen, University Bordeaux 1, France
Dickson Chiu, Dickson Computer Systems, Hong Kong
Sunil Choenni, Research & Documentation Centre, Ministry of Security and Justice / Rotterdam University of Applied Sciences, The Netherlands
Ryszard S. Choras, University of Technology & Life Sciences, Poland
Smitashree Choudhury, Knowledge Media Institute, The UK Open University, UK
William Cheng-Chung Chu, Tunghai University, Taiwan
Christophe Claramunt, Naval Academy Research Institute, France
Cesar A. Collazos, Universidad del Cauca, Colombia
Phan Cong-Vinh, NTT University, Vietnam
Christophe Cruz, University of Bourgogne, France
Beata Czarnacka-Chrobot, Warsaw School of Economics, Department of Business Informatics, Poland
Claudia d'Amato, University of Bari, Italy
Mirela Danubianu, "Stefan cel Mare" University of Suceava, Romania
Antonio De Nicola, ENEA, Italy
Claudio de Castro Monteiro, Federal Institute of Education, Science and Technology of Tocantins, Brazil
Noel De Palma, Joseph Fourier University, France
Zhi-Hong Deng, Peking University, China
Stojan Denic, Toshiba Research Europe Limited, UK
Vivek S. Deshpande, MIT College of Engineering - Pune, India
Sotirios Ch. Diamantas, Pusan National University, South Korea
Leandro Dias da Silva, Universidade Federal de Alagoas, Brazil
Jerome Dinet, Univeristé Paul Verlaine - Metz, France
Jianguo Ding, University of Luxembourg, Luxembourg
Yulin Ding, Defence Science & Technology Organisation Edinburgh, Australia
Mihaela Dinsoreanu, Technical University of Cluj-Napoca, Romania
Ioanna Dionysiou, University of Nicosia, Cyprus
Roland Dodd, CQUniversity, Australia
Suzana Dragicevic, Simon Fraser University- Burnaby, Canada

Mauro Dragone, University College Dublin (UCD), Ireland
Marek J. Druzdzel, University of Pittsburgh, USA
Carlos Duarte, University of Lisbon, Portugal
Raimund K. Ege, Northern Illinois University, USA
Jorge Ejarque, Barcelona Supercomputing Center, Spain
Larbi Esmahi, Athabasca University, Canada
Simon G. Fabri, University of Malta, Malta
Umar Farooq, Amazon.com, USA
Mehdi Farshbaf-Sahih-Sorkhabi, Azad University - Tehran / Fanavaran co., Tehran, Iran
Anna Fensel, Semantic Technology Institute (STI) Innsbruck and FTW Forschungszentrum Telekommunikation Wien, Austria
Stenio Fernandes, Federal University of Pernambuco (CIn/UFPE), Brazil
Oscar Ferrandez Escamez, University of Utah, USA
Agata Filipowska, Poznan University of Economics, Poland
Ziny Flikop, Scientist, USA
Adina Magda Florea, University "Politehnica" of Bucharest, Romania
Francesco Fontanella, University of Cassino and Southern Lazio, Italy
Panagiotis Fotaris, University of Macedonia, Greece
Enrico Francesconi, ITTIG - CNR / Institute of Legal Information Theory and Techniques / Italian National Research Council, Italy
Rita Francese, Università di Salerno - Fisciano, Italy
Bernhard Freudenthaler, Software Competence Center Hagenberg GmbH, Austria
Sören Frey, Daimler TSS GmbH, Germany
Steffen Fries, Siemens AG, Corporate Technology - Munich, Germany
Somchart Fugkeaw, Thai Digital ID Co., Ltd., Thailand
Naoki Fukuta, Shizuoka University, Japan
Mathias Funk, Eindhoven University of Technology, The Netherlands
Adam M. Gadomski, Università degli Studi di Roma La Sapienza, Italy
Alex Galis, University College London (UCL), UK
Crescenzio Gallo, Department of Clinical and Experimental Medicine - University of Foggia, Italy
Matjaz Gams, Jozef Stefan Institute-Ljubljana, Slovenia
Raúl García Castro, Universidad Politécnica de Madrid, Spain
Fabio Gasparetti, Roma Tre University - Artificial Intelligence Lab, Italy
Joseph A. Giampapa, Carnegie Mellon University, USA
George Giannakopoulos, NCSR Demokritos, Greece
David Gil, University of Alicante, Spain
Harald Gjermundrod, University of Nicosia, Cyprus
Angelantonio Gnazzo, Telecom Italia - Torino, Italy
Luis Gomes, Universidade Nova Lisboa, Portugal
Nan-Wei Gong, MIT Media Laboratory, USA
Francisco Alejandro Gonzale-Horta, National Institute for Astrophysics, Optics, and Electronics (INAOE), Mexico
Sotirios K. Goudos, Aristotle University of Thessaloniki, Greece
Victor Govindaswamy, Concordia University - Chicago, USA
Gregor Grambow, AristaFlow GmbH, Germany
Fabio Grandi, University of Bologna, Italy
Andrina Granić, University of Split, Croatia
Carmine Gravino, Università degli Studi di Salerno, Italy
Michael Grottke, University of Erlangen-Nuremberg, Germany
Maik Günther, Stadtwerke München GmbH, Germany
Francesco Guerra, University of Modena and Reggio Emilia, Italy
Alessio Gugliotta, Innova SPA, Italy
Richard Gunstone, Bournemouth University, UK
Fikret Gurgen, Bogazici University, Turkey

Maki Habib, The American University in Cairo, Egypt
Till Halbach, Norwegian Computing Center, Norway
Jameleddine Hassine, King Fahd University of Petroleum & Mineral (KFUPM), Saudi Arabia
Ourania Hatzi, Harokopio University of Athens, Greece
Yulan He, Aston University, UK
Kari Heikkinen, Lappeenranta University of Technology, Finland
Cory Henson, Wright State University / Kno.e.sis Center, USA
Arthur Herzog, Technische Universität Darmstadt, Germany
Rattikorn Hewett, Whitacre College of Engineering, Texas Tech University, USA
Celso Massaki Hirata, Instituto Tecnológico de Aeronáutica - São José dos Campos, Brazil
Jochen Hirth, University of Kaiserslautern, Germany
Bernhard Hollunder, Hochschule Furtwangen University, Germany
Thomas Holz, University College Dublin, Ireland
Władysław Homenda, Warsaw University of Technology, Poland
Carolina Howard Felicíssimo, Schlumberger Brazil Research and Geoengineering Center, Brazil
Weidong (Tony) Huang, CSIRO ICT Centre, Australia
Xiaodi Huang, Charles Sturt University - Albury, Australia
Eduardo Huedo, Universidad Complutense de Madrid, Spain
Marc-Philippe Huget, University of Savoie, France
Chi Hung, Tsinghua University, China
Chih-Cheng Hung, Southern Polytechnic State University - Marietta, USA
Edward Hung, Hong Kong Polytechnic University, Hong Kong
Muhammad Iftikhar, Universiti Malaysia Sabah (UMS), Malaysia
Prateek Jain, Ohio Center of Excellence in Knowledge-enabled Computing, Kno.e.sis, USA
Wassim Jaziri, Miracl Laboratory, ISIM Sfax, Tunisia
Hoyoung Jeung, SAP Research Brisbane, Australia
Yiming Ji, University of South Carolina Beaufort, USA
Jinlei Jiang, Department of Computer Science and Technology, Tsinghua University, China
Weirong Jiang, Juniper Networks Inc., USA
Hanmin Jung, Korea Institute of Science & Technology Information, Korea
Hermann Kaindl, Vienna University of Technology, Austria
Ahmed Kamel, Concordia College, Moorhead, Minnesota, USA
Rajkumar Kannan, Bishop Heber College(Autonomous), India
Fazal Wahab Karam, Norwegian University of Science and Technology (NTNU), Norway
Dimitrios A. Karras, Chalkis Institute of Technology, Hellas
Koji Kashihara, The University of Tokushima, Japan
Nittaya Kerdprasop, Suranaree University of Technology, Thailand
Katia Kermanidis, Ionian University, Greece
Serge Kernbach, University of Stuttgart, Germany
Nhien An Le Khac, University College Dublin, Ireland
Reinhard Klemm, Avaya Labs Research, USA
Ah-Lian Kor, Leeds Metropolitan University, UK
Arne Koschel, Applied University of Sciences and Arts, Hannover, Germany
George Kousiouris, NTUA, Greece
Philipp Kremer, German Aerospace Center (DLR), Germany
Dalia Kriksciuniene, Vilnius University, Lithuania
Markus Kunde, German Aerospace Center, Germany
Dharmender Singh Kushwaha, Motilal Nehru National Institute of Technology, India
Andrew Kusiak, The University of Iowa, USA
Dimosthenis Kyriazis, National Technical University of Athens, Greece
Vitaveska Lanfranchi, Research Fellow, OAK Group, University of Sheffield, UK
Mikel Larrea, University of the Basque Country UPV/EHU, Spain
Philippe Le Parc, University of Brest, France

Gyu Myoung Lee, Liverpool John Moores University, UK
Kyu-Chul Lee, Chungnam National University, South Korea
Tracey Kah Mein Lee, Singapore Polytechnic, Republic of Singapore
Daniel Lemire, LICEF Research Center, Canada
Haim Levkowitz, University of Massachusetts Lowell, USA
Kuan-Ching Li, Providence University, Taiwan
Tsai-Yen Li, National Chengchi University, Taiwan
Yangmin Li, University of Macau, Macao SAR
Jian Liang, Nimbus Centre, Cork Institute of Technology, Ireland
Haibin Liu, China Aerospace Science and Technology Corporation, China
Lu Liu, University of Derby, UK
Qing Liu, The Commonwealth Scientific and Industrial Research Organisation (CSIRO), Australia
Shih-Hsi "Alex" Liu, California State University - Fresno, USA
Xiaoqing (Frank) Liu, Missouri University of Science and Technology, USA
David Lizcano, Universidad a Distancia de Madrid, Spain
Henrique Lopes Cardoso, LIACC / Faculty of Engineering, University of Porto, Portugal
Sandra Lovrencic, University of Zagreb, Croatia
Jun Luo, Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, China
Prabhat K. Mahanti, University of New Brunswick, Canada
Jacek Mandziuk, Warsaw University of Technology, Poland
Herwig Mannaert, University of Antwerp, Belgium
Yannis Manolopoulos, Aristotle University of Thessaloniki, Greece
Antonio Maria Rinaldi, Università di Napoli Federico II, Italy
Ali Masoudi-Nejad, University of Tehran, Iran
Constandinos Mavromoustakis, University of Nicosia, Cyprus
Zulfiqar Ali Memon, Sukkur Institute of Business Administration, Pakistan
Andreas Merentitis, AGT Group (R&D) GmbH, Germany
Jose Merseguer, Universidad de Zaragoza, Spain
Frederic Migeon, IRIT/Toulouse University, France
Harald Milchrahm, Technical University Graz, Institute for Software Technology, Austria
Les Miller, Iowa State University, USA
Marius Minea, University POLITEHNICA of Bucharest, Romania
Yasser F. O. Mohammad, Assiut University, Egypt
Shahab Mokarizadeh, Royal Institute of Technology (KTH) - Stockholm, Sweden
Martin Molhanec, Czech Technical University in Prague, Czech Republic
Charalampos Moschopoulos, KU Leuven, Belgium
Mary Luz Mouronte López, Ericsson S.A., Spain
Henning Müller, University of Applied Sciences Western Switzerland - Sierre (HES SO), Switzerland
Susana Munoz Hernández, Universidad Politécnica de Madrid, Spain
Bela Mutschler, Hochschule Ravensburg-Weingarten, Germany
Deok Hee Nam, Wilberforce University, USA
Fazel Naghdy, University of Wollongong, Australia
Joan Navarro, Research Group in Distributed Systems (La Salle - Ramon Llull University), Spain
Rui Neves Madeira, Instituto Politécnico de Setúbal / Universidade Nova de Lisboa, Portugal
Andrzej Niesler, Institute of Business Informatics, Wroclaw University of Economics, Poland
Kouzou Ohara, Aoyama Gakuin University, Japan
Jonice Oliveira, Universidade Federal do Rio de Janeiro, Brazil
Ian Oliver, Nokia Location & Commerce, Finland / University of Brighton, UK
Michael Adeyeye Oluwasegun, University of Cape Town, South Africa
Sascha Opletal, University of Stuttgart, Germany
Fakri Othman, Cardiff Metropolitan University, UK
Enn Õunapuu, Tallinn University of Technology, Estonia
Jeffrey Junfeng Pan, Facebook Inc., USA

Hervé Panetto, University of Lorraine, France
Malgorzata Pankowska, University of Economics, Poland
Harris Papadopoulos, Frederick University, Cyprus
Laura Papaleo, ICT Department - Province of Genoa & University of Genoa, Italy
Agis Papantoniou, National Technical University of Athens, Greece
Thanasis G. Papaioannou, École Polytechnique Fédérale de Lausanne (EPFL), Switzerland
Andreas Papasalouros, University of the Aegean, Greece
Eric Paquet, National Research Council / University of Ottawa, Canada
Kunal Patel, Ingenuity Systems, USA
Carlos Pedrinaci, Knowledge Media Institute, The Open University, UK
Yoseba Penya, University of Deusto - DeustoTech (Basque Country), Spain
Cathryn Peoples, Ulster University, UK
Asier Perallos, University of Deusto, Spain
Christian Percebois, Université Paul Sabatier - IRIT, France
Andrea Perego, European Commission, Joint Research Centre, Italy
Mark Perry, University of Western Ontario/Faculty of Law/ Faculty of Science - London, Canada
Willy Picard, Poznań University of Economics, Poland
Agostino Poggi, Università degli Studi di Parma, Italy
R. Ponnusamy, Madha Engineering College-Anna University, India
Jerzy Prekurat, Canadian Bank Note Co. Ltd., Canada
Didier Puzenat, Université des Antilles et de la Guyane, France
Sita Ramakrishnan, Monash University, Australia
Elmano Ramalho Cavalcanti, Federal University of Campina Grande, Brazil
Juwel Rana, Luleå University of Technology, Sweden
Martin Randles, School of Computing and Mathematical Sciences, Liverpool John Moores University, UK
Christoph Rasche, University of Paderborn, Germany
Ann Reddipogu, ManyWorlds UK Ltd, UK
Ramana Reddy, West Virginia University, USA
René Reiners, Fraunhofer FIT - Sankt Augustin, Germany
Paolo Remagnino, Kingston University - Surrey, UK
Sebastian Rieger, University of Applied Sciences Fulda, Germany
Andreas Riener, Johannes Kepler University Linz, Austria
Ivan Rodero, NSF Center for Autonomic Computing, Rutgers University - Piscataway, USA
Alejandro Rodríguez González, University Carlos III of Madrid, Spain
Paolo Romano, INESC-ID Lisbon, Portugal
Agostinho Rosa, Instituto de Sistemas e Robótica, Portugal
José Rouillard, University of Lille, France
Paweł Różycki, University of Information Technology and Management (UITM) in Rzeszów, Poland
Igor Ruiz-Agundez, DeustoTech, University of Deusto, Spain
Michele Ruta, Politecnico di Bari, Italy
Melike Sah, Trinity College Dublin, Ireland
Francesc Saigi Rubió, Universitat Oberta de Catalunya, Spain
Abdel-Badeeh M. Salem, Ain Shams University, Egypt
Yacine Sam, Université François-Rabelais Tours, France
Ismael Sanz, Universitat Jaume I, Spain
Ricardo Sanz, Universidad Politecnica de Madrid, Spain
Marcello Sarini, Università degli Studi Milano-Bicocca - Milano, Italy
Munehiko Sasajima, I.S.I.R., Osaka University, Japan
Minoru Sasaki, Ibaraki University, Japan
Hiroyuki Sato, University of Tokyo, Japan
Jürgen Sauer, Universität Oldenburg, Germany
Patrick Sayd, CEA List, France
Dominique Scapin, INRIA - Le Chesnay, France

Kenneth Scerri, University of Malta, Malta
Rainer Schmidt, Austrian Institute of Technology, Austria
Bruno Schulze, National Laboratory for Scientific Computing - LNCC, Brazil
Ingo Schwab, University of Applied Sciences Karlsruhe, Germany
Wieland Schwinger, Johannes Kepler University Linz, Austria
Hans-Werner Sehring, Tallence AG, Germany
Paulo Jorge Sequeira Gonçalves, Polytechnic Institute of Castelo Branco, Portugal
Kewei Sha, Oklahoma City University, USA
Roman Y. Shtykh, Rakuten, Inc., Japan
Robin JS Sloan, University of Abertay Dundee, UK
Vasco N. G. J. Soares, Instituto de Telecomunicações / University of Beira Interior / Polytechnic Institute of Castelo Branco, Portugal
Don Sofge, Naval Research Laboratory, USA
Christoph Sondermann-Woelke, Universitaet Paderborn, Germany
George Spanoudakis, City University London, UK
Vladimir Stantchev, SRH University Berlin, Germany
Cristian Stanciu, University Politehnica of Bucharest, Romania
Claudius Stern, University of Paderborn, Germany
Mari Carmen Suárez-Figueroa, Universidad Politécnica de Madrid (UPM), Spain
Kåre Synnes, Luleå University of Technology, Sweden
Ryszard Tadeusiewicz, AGH University of Science and Technology, Poland
Yehia Taher, ERISS - Tilburg University, The Netherlands
Yutaka Takahashi, Senshu University, Japan
Dan Tamir, Texas State University, USA
Jinhui Tang, Nanjing University of Science and Technology, P.R. China
Yi Tang, Chinese Academy of Sciences, China
John Terzakis, Intel, USA
Sotirios Terzis, University of Strathclyde, UK
Vagan Terziyan, University of Jyvaskyla, Finland
Lucio Tommaso De Paolis, Department of Innovation Engineering - University of Salento, Italy
Davide Tosi, Università degli Studi dell'Insubria, Italy
Raquel Trillo Lado, University of Zaragoza, Spain
Tuan Anh Trinh, Budapest University of Technology and Economics, Hungary
Simon Tsang, Applied Communication Sciences, USA
Theodore Tsiligiridis, Agricultural University of Athens, Greece
Antonios Tsourdos, Cranfield University, UK
José Valente de Oliveira, University of Algarve, Portugal
Eugen Volk, University of Stuttgart, Germany
Mihaela Vranić, University of Zagreb, Croatia
Chieh-Yih Wan, Intel Labs, Intel Corporation, USA
Jue Wang, Washington University in St. Louis, USA
Shenghui Wang, OCLC Leiden, The Netherlands
Zhonglei Wang, Karlsruhe Institute of Technology (KIT), Germany
Laurent Wendling, University Descartes (Paris 5), France
Maarten Weyn, University of Antwerp, Belgium
Nancy Wiegand, University of Wisconsin-Madison, USA
Alexander Wijesinha, Towson University, USA
Eric B. Wolf, US Geological Survey, Center for Excellence in GIScience, USA
Ouri Wolfson, University of Illinois at Chicago, USA
Yingcai Xiao, The University of Akron, USA
Reuven Yagel, The Jerusalem College of Engineering, Israel
Fan Yang, Nuance Communications, Inc., USA
Zhenzhen Ye, Systems & Technology Group, IBM, US A

International Journal on Advances in Intelligent Systems
Volume 14, Numbers 1 & 2, 2021

CONTENTS

Backtracking (the) Algorithms on the Hamiltonian Cycle Problem

Joeri Sleegers

Mice & Man Software and A.I. Development
Amsterdam
The Netherlands
jsleegers@hotmail.com

Daan van den Berg

Yamasan Science & Education
Amsterdam
The Netherlands
daan@yamasan.nl

***Abstract*—Even though the Hamiltonian cycle problem is NP-complete, many of its problem instances are not. In fact, almost all the hard instances reside in one area: near the Komlós-Szemerédi bound, where randomly generated graphs have an approximate 50% chance of being Hamiltonian. If the number of edges is either much higher or much lower, the problem is not hard – most backtracking algorithms decide such instances in (near) polynomial time. Recently however, targeted search efforts have identified very hard Hamiltonian cycle problem instances *very far away* from the Komlós-Szemerédi bound. In that study, the used backtracking algorithm was Vandegriend-Culberson's, which was supposedly the most efficient of all Hamiltonian backtracking algorithms. In this paper, we make a unified large scale quantitative comparison for the best known backtracking algorithms described between 1877 and 2016. We confirm the suspicion that the Komlós-Szemerédi bound is a hard area for all backtracking algorithms, but also that Vandegriend-Culberson is indeed the most efficient algorithm, when expressed in consumed computing time. When measured in recursive effectiveness however, the algorithm by Frank Rubin, almost half a century old, performs best. In a more general algorithmic assessment, we conjecture that edge pruning and non-Hamiltonicity checks might be largely responsible for these recursive savings. When expressed in system time however, denser problem instances require much more time per recursion. This is most likely due to the costliness of the extra search pruning procedures, which are relatively elaborate. We supply large amounts of experimental data, and a unified single-program implementation for all six algorithms. All data and algorithmic source code is made public for further use by our colleagues.**

***Keywords*—Hamiltonian Cycle; exact algorithm; exhaustive algorithm; heuristic; phase transition; order parameter; data analytics; instance hardness; replication.**

I. Preamble

Traversing a crack in the fabric of the scientific spacetime-continuum, this paper finds itself in the unusual position that its designated conclusions have already been overthrown. Following a replication study [1], these extended results should have been published earlier, but as history unfolded, they simply were not. In any case, the study by Cheeseman, Kanefsky & Taylor (henceforth: 'Cetal') was first [2]. Sharpening the resolution of the $P \stackrel{?}{=} NP$ problem, they showed that for various NP-complete problems, Hamiltonian cycle, graph colouring, satisfiability and the asymmetric traveling salesman problem[1], instances vary greatly in their computational hardness. Sporting over 1400 citations, the paper became a landmark in the field.

It took nearly 30 years for Cetal's results on traveling salesman to be overthrown, appearing to have been flawed by an overlooked roundoff error [3] (also see the accompanying videos: [4] [5]). Their results on the Hamiltonian cycle problem however, were succesfully replicated and published at IARIA's Data Analytics 2018 conference, and brought the data, sourcecode and figures alive in online interactively publicly accessible resources [1] [6]. The results of that extended study showed that for Cetal's algorithm, and two others, the hardest instances of the Hamiltonian cycle were located along an area known as the "Komlós-Szemerédi bound". So far so good.

But the lingering question was how large the influence of the solvers was. Were the found hard instances hard specifically for the solving algorithms used? For the pruning methods? For the branching heuristic? And to make matters worse, a recent follow-up study showed that for the allegedly most efficient Hamiltonian cycle backtracker, the Vandegriend-Culberson (henceforth: 'Vacul')-algorithm, (which still requires exponential time in the worst case), the hardest instances were located *very far away* from the Komlós-Szemerédi bound [7] [8]. Only findable by sophisticated evolutionary algorithms and a whole lot of computing power, these very hard instances never showed up in earlier studies.

So is there an insurmountable contradiction here? Probably not. The authors conjecture three reasons for their counterintuitive results: "A first explanation for these surprising results is that these results are specific for the backtracking algorithm we used. This is unlikely however, as the algorithm minimizes most other backtracking algorithms found in literature (yet unpublished results)". Put differently: chances are very high that these graphs are *also* hard for other complete backtracking algorithms but evidence pending, it remained unconfirmed as yet. This study will at the very least post a very serious sidenote to that hypothesis. A second explanation given by these authors might be that in most studies on Hamiltonian cycle backtracking algorithms, runs are cutoff after a preset

[1]Even though the authors themselves dubbed traveling salesman as "NP-complete", they solved the NP-hard version of the problem.

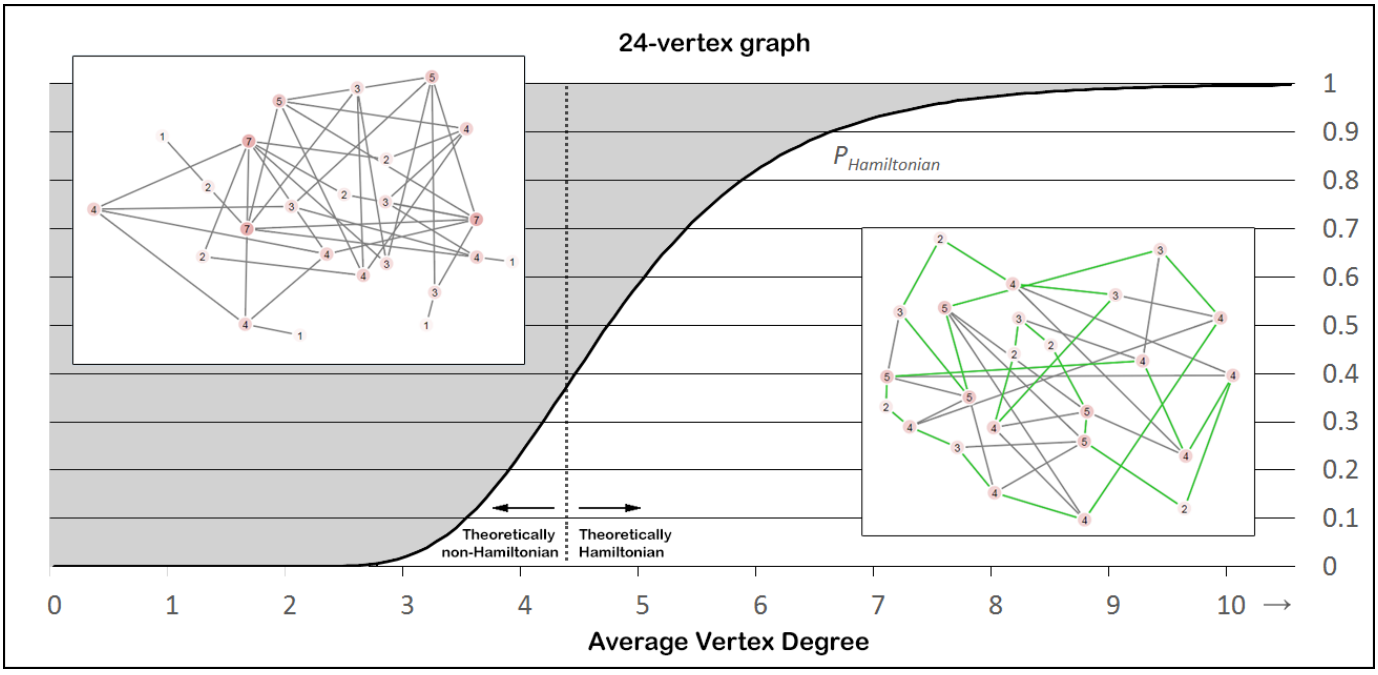


Fig. 1. The probability of a randomly generated graph being Hamiltonian depends on the average vertex degree, and is sigmoidally shaped around the 'threshold point' of $ln(V) + ln(ln(V))$. Top-left inset is a non-Hamiltonian random graph, bottom-right inset is a Hamiltonian graph with the Hamiltonian cycle itself being highlighted.

number of recursions, as even small graphs can take up significant decision time. Although theoretically feasible, in practice these cutoff points are usually situated near the Koml´os-Szemer´edi bound, and not in edge-dense regions far away where the extremely hard instances were located.

The third and most compelling explanation might therefore come from their obsevation that on first glance, these very hard instances might have low Kolmogorov complexity – they are in some sense *structured* graphs. And as unstructured objects in randomly generated ensemble vastly outnumber the structured objects, the chances of being created by a stochastic process (which is the case in most large-scale comparative studies) are extremely small. Put differently: one would simply not find the very hard instances unless knowing exactly where to look. Or as the authors more poetically phrased: "These graphs are an isolated island of structured hardness in a ocean of unstructured easiness. Whether more such islands exist, and what they look like, awaits further exploration."[2] [7].

So to understand their results, a couple of things need to happen. First, Vacul's algorithm must *evidently* minimize the other backtracking algorithms. Then, a global evolutionary search algorithm should look for the hardest instances for *all* these backtracking algorithms. Third, a hardness hierarchy should be made for all six algorithms, another large quantitative study. Fourth, an aggregated theory should explain *why* some instances of the Hamiltonian cycle problem are harder than others, and what their relation to the Komlós-Szemerédi bound is (or is not!). It is the authors' belief that such an aggregated theory exists. Beyond that, as a fifth point, might lie implications for computability, and the $P \stackrel{?}{=} NP$ problem itself, but developments in this area will depend on foregoing results, and are as yet too close to call.

In this paper, which is an extension of the IARIA'18 paper mentioned in the second paragraph [1], we will address step one: show that in large random ensembles, the hardest instances of the Hamiltonian cycle problem lie around the Komlós-Szemerédi bound for *all* well-known general backtracking algorithms found in literature. We will make exact calculations on large ensembles, and perform a rigorous comparative analysis. While reading the paper, the reader should keep two things in mind: first, other backtracking algorithms than those in this study are well existable and second: it is possible that much harder instances exist for any backtracker in this study or elsewhere. As shown earlier, extremely hard instances are likely extremely rare, extremely hard to find, but also extremely important for the $P \stackrel{?}{=} NP$ problem, as it is these instances that etch upper bounds on these algorithms' runtimes. Finding these, possibly with targeted evolutionary algorithms, is the second step and will hopefully be done in the near future. For now, we will work with completely unbiased random ensembles of instances to map out the gross features of the complexity landscape.

II. Introduction

The "Great Divide" between P and NP has haunted computer science and related disciplines for over half a century. Problems in P are problems for which the runtime of the best

[2]The literate reader might remind Aldous Huxley's famous quote *"Consider the horse. They considered it." – A Brave New World, 1932.* [9]

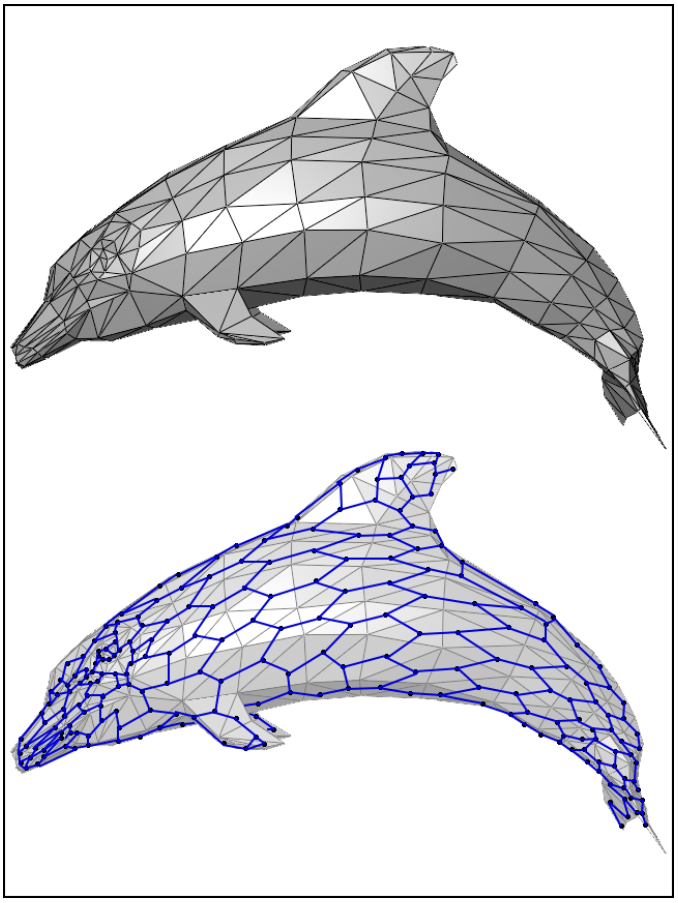


Fig. 2. Fast rendering of triangle mesh 3D images critically depends on finding Hamiltonian cycles through the corresponding 'cubic' graphs, in which every vertex has a maximum degree of three.

known algorithm increases polynomially with the problem size, for example, calculating the average of an array of numbers. If the array doubles in size, so does the runtime of the best known algorithm - a polynomial increase. A problem in NP however, has no such polynomial-time algorithm and it is an open question whether one will ever be found. An example hereof is 'satisfiability' (sometimes abbreviated to SAT), in which an algorithm assigns values 'true' or 'false' to variables in Boolean formulas like $(a \vee \neg b \vee d) \wedge (b \vee c \vee \neg d)$. The task is to choose the variable assignment so that the formula as a whole is satisfied (becomes 'true'), *and returning that assignment*, or making sure that no such assignment exists. Algorithms that do this, algorithms that are guaranteed to give a solution whenever it exists and return 'no' otherwise, are called *exact* algorithms[3]

Being exact is a great virtue for an algorithm, but it comes at a hefty price. Often, these algorithms operate by *brute-force* procedures: exhaustively trying all combinations for all variables until a solution is found, which usually takes vast amounts of time. Depth-first search is the keystone example of this study; it is exhaustively exact and indeed consumes harrowing amounts of computational power (see Figures 4 and 5). Smarter algorithms exist too; clever search pruning can speed things up by excluding large sections of state space at the cost of some extra computational instructions, an investment that usually pays off. Heuristic algorithms are also fast but not necessarily exact - so it is not guaranteed a solution is found if one exists. After decades of research, the runtimes of even the most efficient complete SAT-algorithm known today still increases exponentially with the number of variables – much worse than polynomial, even for low exponents. Therefore, SAT is in NP, a class of 'Notorious Problems' that rapidly become unsolvable as their size increases. In practice, this means that satisfiability problems (and other problems in NP) with only a few hundred variables are practically unsolvable, whereas industries such as chip manufacture or program verification in software engineering could typically employ millions ˜ [10]˜ [11].

The problem class NP might be considered "the class of dashed hopes and idle dreams" [12], but nonetheless scientists managed to pry loose a few bricks in the great wall that separates P from NP. Most notably, the seminal work "Where the *Really* Hard Problems Are" by Cheeseman, Kanefsky and Taylor (henceforth abbreviated to 'Cetal'), showed that although runtime increases non-polynomially for problems in NP, some *instances* of these hard problems might actually be easy to solve ˜ [2]. Not every formula in SAT is hard – easily satisfiable formulas exist too, even with many variables, but the hard ones keep the problem as a whole in NP. But Cetal's great contribution was not only to expose the huge differences in instance hardness within a single NP-problem, they also showed *where* those really hard instances are – and how to get there. Their findings were followed up numerous times and truly exposed some of the intricate inner anatomy of instance hardness, and problem class hardness as a whole.

So where *are* these hard problem instances then? According to Cetal, they are hiding in the phase transition. For the problems in their study, instances suddenly jump from 'having many solutions' to 'having no solutions' when their constrainedness changes. For an example in satisfiability, most randomly generated SAT-formulas of two clauses and four variables such as our formula $(a \vee \neg b \vee d) \wedge (b \vee c \vee \neg d)$ are easily satisfiable; they have many assignments that make them true. But as soon as the order parameter, the ratio of clauses versus variables α, passes ≈ 4.26, (almost) no satisfiable formulas exist [13]˜ [14]. So if we randomly generate a formula with 20 or more clauses on these same four variables, it is almost certainly unsatisfiable and those rare formulas that *are* satisfiable beyond the phase transition have very few solutions – which counterintuitively enough makes them easy again. So, for most exact algorithms, both extremes are quickly decided: for the highly satisfiable formulas in $\alpha << 4.26$, a solution is quickly found, and unsatisfiable formulas in $\alpha >> 4.26$ are quickly proven as such. But in between, just around $\alpha = 4.26$, where the transition from satisfiable to unsatisfiable takes

[3]Technically speaking, exact algorithms for decision problems such as the Hamiltonian cycle problem should be called *complete*. But since 'exact' seems more in schwung recently, we will stick to that.

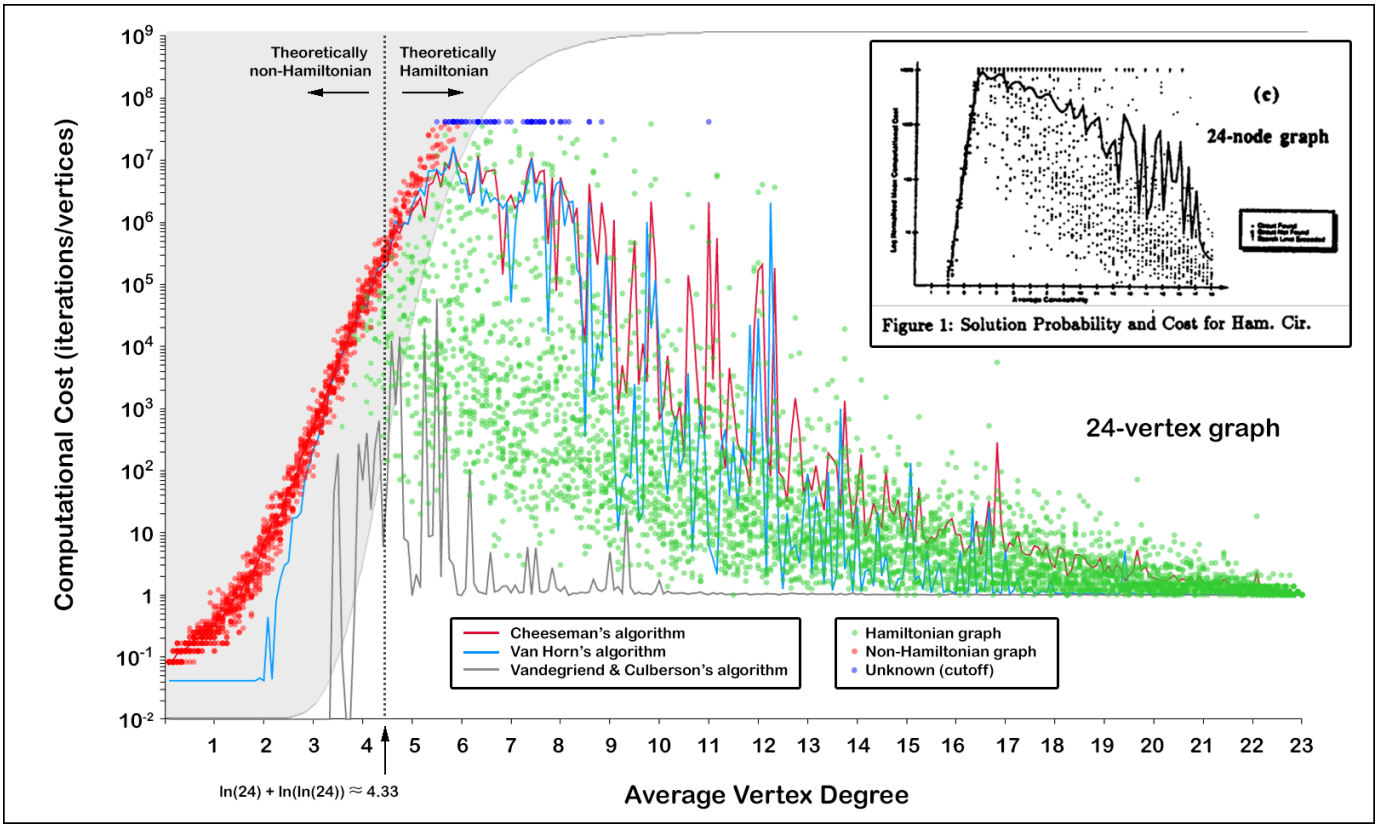


Fig. 3. Results from an earlier replication of Cetal's seminal work on Hamiltonian cycle hardness, extended with algorithms by Van Horn and Vacul. The top-right inset is Cetal's original figure, and it covers no data points. Note how the computational cost is highest along the Komlós-Szemerédi bound.

place, are formulas that take the longest to decide upon. This is where the really hard problem instances are: hiding in the phase transition that separates the solvable from the unsolvable region.

Cetal identify this order parameter not only for SAT; the Hamiltonian cycle problem has one too, and so does k-colorability. For the Hamiltonian cycle problem though, one should understand that the phase transition is somewhat inverse to SAT: as one adds edges to a graph, it becomes *solvable* rather than unsolvable. But the general principle still holds: both extremes are easy, and the phase transition is where the really hard problem instances are. And although Cetal's seminal results are relatively coarse, they have been followed up in more detail, and they are solid ˜ [14]–[17]. To put it in a quote by Ian Gent and Toby Walsh: "[Indeed, we have yet to find an NP-complete problem that *lacks* a phase transition]" ˜ [18].

The ubiquity of phase transitions throughout the class is not a complete surprise. Satisfiability, k-colorability and the Hamiltonian cycle problem are *NP-complete* problems; a subset of problems in NP that with more or less effort can be transformed into each other ˜ [19]. This means a lot. This means that if someone finds a polynomial exact algorithm for just one of these problems, all of them become easy and the whole hardness class will simply evaporate. That person would also be an instant millionaire thanks to the Clay Mathematics Institute that listed the P$\stackrel{?}{=}$NP-question as one of their 'Millenium Problems' ˜ [20]. But the intricate relations inside NP-completeness might also stretch into the properties of phase transitions and instance hardness. Or, to pour it into another fluid expression by Ian Gent and Toby Walsh "[Although any NP-complete problem can be transformed into any other NP-complete problem, this mapping does *not* map the problem space uniformly]" ˜ [18]. So, a phase transition in say, satisfiability, does *not* guarantee the existence of a phase transition in Hamiltonian Cycle or in Vertex Coloring. The fact is though, that Cetal do find them for all three, and their results are solid.

III. The Hamiltonian Phase Transition

The Hamiltonian cycle problem comes in many different shapes and forms, but in its most elementary formulation involves finding a path (a sequence of distinct edges) in an undirected and unweighted graph that visits every vertex exactly once, and forms a closed loop. The probability of a random graph being Hamiltonian (i.e., having a Hamiltonian Cycle), has been thoroughly studied ˜ [21]–[23]. In the limit, it is a smooth function of vertex degree and therefore the probability for a random graph of v Vertices and e edges being

Hamiltonian can be calculated analytically[4]:

$$P_{Hamiltonian}(v,e) = e^{-e^{-2c}} \quad (1)$$

in which

$$e = \frac{1}{2}v \cdot ln(v) + \frac{1}{2}v \cdot ln(ln(v)) + c \cdot v \quad (2)$$

Like the phase transition around α in SAT, the Hamiltonian phase transition is also sigmoidally shaped[5] across a 'threshold point', the average degree of $ln(v) + ln(ln(v))$ for a graph of v vertices (Figure 1). The phase transition gets ever steeper for larger graphs and becomes instantaneous at the threshold point as v goes to infinity. For this (theoretical) reason, the probability of being Hamiltonian at the threshold point is somewhat below 0.5 at $e^{-1} \approx 0.368$.

The probability of being Hamiltonian is one thing, deciding whether a given graph actually *has* a Hamiltonian cycle is quite another. A great number of exact algorithms have been developed through the years, the earliest being exhaustive methods that could run in $O(v!)$ time [24]. A dynamic programming approach, quite advanced for the time, running in $O(n^2 2^n)$ was built by by Michael Held & Richard Karp, and by Richard Bellman independently [25] [26]. Some early edge pruning efforts and check routines can be found in the work of Silvano Martello and Frank Rubin whose algorithms still run in $O(v!)$ but are in practice much faster (as we will show in Figures 4 and 5) [27] [28]. Many of their techniques eventually ended up in the algorithm by Vandegriend & Culberson [29]. Algorithms by Bollobás and Björklund run faster than Bellman–Held–Karp, but are technically speaking not exact for finite graphs [30] [31]. The 2007 algorithm by Iwama & Nakashima [32] runs in $O(2^{1.251v})$ time on cubic graphs, thereby improving Eppstein's 2003 algorithm that runs in $O(2^{1.260v})$. While these kind of marginal improvements on specific instances are typical for the progress in the field, these two actually deserve some extra attention.

The cubic graph, in which every vertex has a maximum degree of three, is of special importance in the generation of 3D computer images. Many such images are built up from triangle meshes, and as specialized hardware render and shade triangles at low latencies, the performance bottleneck is actually in feeding the triangular structure into the hardware. A significant speedup can be achieved by not feeding every triangle by itself, but by combining them into triangle strips. An adjacent triangle can be defined by only one new point from the previously fed triangle, and therefore adjacent triangles combined in a single strip can speedup the feeding procedure by a maximum factor three for each 3D object. Finding a single strip that incorporates all triangles in the mesh is equivalent to finding a Hamiltonian cycle through the corresponding cubic graph in which every triangle is a vertex, which makes both Eppstein's and Iwama&Nakashima's result of crucial importance for the 3D imagery business (see Figure 2).

So concludingly, none of the exact algorithms for the Hamiltonian cycle problem runs faster than exponential on all instances (with vertices of any degree), and the Bellman–Held–Karp "is still the strongest known", as Andreas Björklund states on the first page of his 2010-paper [31]. Summarizingly, all backtracking algorithms still run in $O(v!)$ time. In the next sections, we will have a closer look at the algorithmic details of all *exact* algorithms from previous pargraphs. We will first make a structural and historical comparison, and after that bang out a lot of data to make a rigorous quantitative comparison as well. Finally, we will draw some conclusions, discuss the impact of the work, and project a trajectory for future research.

IV. ALGORITHMICS

For this extended investigation, we generated large numbers of problem random instances for the Hamiltonian cycle problem, varying in numbers of vertices and edges. We then solved these using almost every authoritative exact algorithm we could find: **Depth-first search** invented far before any modern day computer in 1882, **Rubin's algorithm** from 1974, **Martello's algorithm '595'** published in 1983, **Cetal's** algorithm, used in their 1991 seminal work on instance hardness, Vandegriend & Culberson's, abbreviated to **Vacul's algorithm**, an elaborate machinery published in 1998 and finally **Van Horn's algorithm**, directly derived from Cetal's, published in 2018 [33] [28] [27] [2] [29] [1].

There are at least several more, some of which are very old and without Google Scholar index or available pdfs. One prominent candidate missing in our list is the dynamic programming implementation built by Michael Held & Richard Karp, and by Richard Bellman independently [25] [26]. While the algorithm is complete, and with a time complexity of $O(v^2 2^v)$ has a better worst case performance than our six algorithms which all have $O(v!)$, it has significant memory requirements. Furthermore, it is not a backtracker and therefore does not perform 'recursions' as such, making a direct comparison to the other six slightly more difficult. It might be a viable candidate for a future comparison but for now, we will stick to backtracking algorithms, and canalize towards a generalized approach.

As it turns out, these six algorithms have several similarities and differences. But even though the respective papers have significant numbers of citations, some of the (especially earlier) authors of these particular algorithms seemed to be unaware of each others' progress. The main exception is Vacul, who include a reference to both Martello and Cetal, but missed Frank Rubin's work, which might simply be due to geographical dispersity and the lack of internet. Or conversely: the relatively recent development of computational resources, and proliferation of global communication enables us *now* to make a large structural comparison between them relatively easily. In

[4] An unfortunate convention: please note that the e in the right hand side of Eq. 1 is the base of the natural logarithm, whereas in the left hand side, e is the number of edges

[5] Remember this transition is *invertalized* with respect to SAT: it goes 'from no to yes' when getting denser, where SAT goes 'from yes to no'.

any case, there appear to be some globally emerging algorithmic design patterns for this problem, which are dispersedly found accross algorithms. We will structurally compare all, and assess their effectiveness regarding the Hamiltonian cycle problem.

In this study we generalize the approach, unifying similar procedural subroutines to the same piece of source code. The only feature that was removed is the random restart option from Vacul's algorithm. Surely a good way of shortening the average runtime on many NP-complete problems [34], it also makes one-off comparisons a lot harder on large randomized instance ensembles such as ours. As most algorithms predate the internet, none came with readily implementable source code. We like to emphasize our belief that text such as you are reading now is an inferior medium for communicating algorithmics. As the smallest of details can make the largest differences when it comes to runtimes. Therefore, one should *always supply publicly accessible source code* when writing about algorithms. We will try to set the example by supplying ours [35]. It is quite possible that algorithmic details as we chose them are different from the original authors' implementations. It is also thinkable that even where we are precise, further improvements are possible. In any case, let us go forward with public source code along our publications on algorithms. It is better than text.

In the next section, we will describe the six backtrack algorithms used, but also the subroutines, many of which are common to multiple algorithms. The subjects are somewhat 'interwoven' between the algorithmic explanation, simply because it seems to make the most sense storywise. Operationally speaking, it makes more sense to compartmentalize these subroutines, so that is how the reader will find them in the source code [35]. For a general overview, please refer to Table I.

V. Three Basic Algorithms

Depth-first search can in every way be considered as the basis for all algorithms in this study. It is surely the oldest, gaining widespread popularity from Tarjan's paper [36], but was designed roughly a century earlier by Frenchman Charles Pierre Trémaux, and mentioned in a publication from 1882 as a strategy for solving mazes [37]. As (planar) mazes are in many ways equivalent to (planar) graphs, its success as an algorithm for graph traversal is hardly surprising.

In modern-day computers, depth-first search is readily implemented either by a recursive function or via a stack data structure, the latter of which is usually a constant factor slower. Depth-first search is an exact algorithm in optimization problems, meaning it will always return the best possible answer (e.g. the shortest route, or the optimal timetable). For decision problems like the Hamiltonian cycle problem or satisfiability, it will always return a solution if the problem instance has one, or ensure it does not exist. For finding a Hamiltonian cycle (or any other particular vertex order) in a graph, its ominous runtime complexity is $O(v!)$ in the number of vertices v, which makes it practically unusable for any number of v over two digits. The algorithm can just as well be deployed for solving SAT-formulas, running in slightly less daunting exponential complexity of $O(2^n)$ in n, the number of Boolean variables.

Depth-first search is a constructive algorithm that starts at the first vertex in the data structure that holds the randomly generated graph, thereby not having any specified preference for degree. From there, its recursive step is to add the first adjacent vertex (again, in the order of the data structure) that is not already in the path. This step is repeated either until no adjacent vertices are available. This can mean two things: either all vertices are in the path, and a final check for closure confirms the existence of a Hamiltonian Cycle after which the algorithm halts, or the algorithm backtracks, removing the last vertex from the path and adding the next adjacent vertex in its place. If none such vertex exists it backtracks again. If this happens at the root level then all possibilities are exhausted, and the algorithm halts and returns "no Hamiltonian Cycle". As a typical property of (uninformed) exhaustive exact algorithms, it thereby tries all permutations of vertices if necessary and has a time complexity, or worst-case runtime, of $O(v!)$.

A big *practical* difference however, materializes by not just adding the next vertex from the data structure, but preferring either vertices of high degree or low degree. This can be done runtime, or by sorting the data structure upon reading the graph ahead of recusing. In our implementation, we always chose the former option, resorting during the run. As this next-vertex-preference is a somewhat rule-of-the-thumb, we will call it the **branching heuristic** and can be instantiated with categorical[6] parameter values$\{none, high, low\}$ (see Table I). In its simplest implementation, it involves only changing a '$>$' to a '$<$' in the source code of the algorithm. Nonetheless the impact on the algorithm's performance for large ensembles of instances such as ours can be enormous [1].

Cetal's algorithm follows the exact same paradigmic lines as plain depth-first search, but prefers higher degree vertices when branching. Ties are unaddressed, and whether a more sophisticated order of preference has any significant impact on the algorithm's runtime remains an open question, especially for larger graphs. Cetal's algorithm still runs in $O(v!)$, but its somewhat more fine-grained time complexity might be $O(v \cdot log(v))!$, with the additional term accounting for sorting the vertices to descending degree numbers. Cetal's paper was published with four experiments, among which the Hamiltonian cycle problem, and like this study, contained experimental results on an ensemble of random graphs in regular degree intervals.

Van Horn's algorithm is in many ways the opposite of Cetal's algorithm, starting at the vertex with the lowest degree and preferring lower degree vertices over higher degree vertices when recursing. It therefore also runs in $O(v!)$ time,

[6]Sometimes referred to as 'symbolic parameter' or 'qualitative parameter'.

but its fine-grained time complexity is $v \cdot log(v) + v!$. Its only algorithmic difference is that it traverses the data structure backwardly, or alternatively, that its data structure is sorted in inverse order. Van Horn's algorithm was introduced during an extended replication of Cetal's Hamiltonian study, and thereby contained experimental results on random graphs of various average degree. The ensemble was likely larger than Cetal's[7], and made publicly available, along with the source code of the algorithms used[8].

VI. Search Pruning

A. Search Pruning: Edge Pruning

One of the most prolific enhancements of depth-first search on NP-complete problems is **search pruning**: cutting off branches from the search tree that cannot hold a solution. If a contradictive pair of clauses is found from some variable assignment in CNF3SAT, the search backtracks immediately, departing from further assignments in the subtree. If a perfect rectangle packing problem contains an unfillable gap, it cannot be solved regardless of the rest of the configuration, so depth-first will halt and backtrack [38]. As such, search pruning can be a valuable way of saving recursions, but the (sometimes huge) drawback is that the pruning procedure *itself* takes computational resources too. Even though most pruning procedures are subexponential routines, they can theoretically be invoked in every recursion. The art therefore, is to balance an investment in pruning so that it pays enough time dividend. In many cases, this is well possible, or, abiding by Steven Skiena's famous words: "Clever pruning can make short work of surprisingly hard problems" [39].

For the Hamiltonian cycle problem, search pruning comes in two forms. The first, **edge pruning**, involves removing edges from the graph, which can be done either in the preprocessing stage or during a recursive step. By removing edges that *cannot possibly be in any Hamiltonian cycles*, the nodes in the recursive search tree get a lower degree, leading to fewer recursions. Interestingly enough, the prunative removal of edges is always related to the existence of *required edges* that are adjacent to a vertex with degree two and therefore *must* be in any Hamiltonian cycle that might exist in the graph. It can be slightly confusing to keep search pruning and edge pruning apart, but it can be remembered like this: edge pruning means cutting away edges, search pruning involves speeding up the search process. Edge pruning therefore, partially instantiates search pruning.

From literature, we find exactly three edge pruning methods which are all implemented in this study. The first, **neighbour pruning**, finds a vertex with two required edges, and removes all others. Second, the **path pruning** method looks for paths of required edges, which might eventually become a Hamiltonian cycle, and removes edges that would prematurely close it. Third, **solution pruning** removes all edges from the second-last vertex of the partial Hamiltonian path so far. Since path pruning and neighbour pruning might reciprocally facilitate each other's operation, both are repeated until no further pruning occurs. Finally, edges that are removed during preprocessing before recursing are definitely gone, but edges pruned in the recursive step *have* to be put back when the algorithm backtracks.

B. Search Pruning: Non-Hamiltonicity Checks

The second category of search pruning does not involve the removal of edges, but checking whether a Hamiltonian cycle is achievable at all. There are four such *non-Hamiltonicity checks* and it should be noted that indeed all of them are negative: they can only decide graphs to be either 'surely non-Hamiltonian' or 'undecided'. Although checks that give 'surely Hamiltonian' or 'undecided' are certainly imaginable, they are not described in the literature regarding the six general backtrack algorithms used in this study.

The first is a plain and simple **degree check** which verifies whether any vertex in the graph has edge degree one or zero. If this is the case, the graph cannot be Hamiltonian, and the algorithm needs to backtrack. The second graph configuration that cannot contain a Hamilton Cycle, is a graph where required edges form a closed loop smaller than v. This graph configuration can be checked by iterating over all required edges and thus required $O(v)$ computational effort. We call this check the **premature closure** check. The third check, the **disconnectedness check** determines whether a graph is broken up into two or more pieces. The subroutine involves picking a vertex and adding it to an empty list of found vertices. From that list, it picks the next item and adds all its adjacent vertices to end of the list, unless they are already added. When the subroutine reaches the end of the list, it counts the number

[7]Cetal's ensemble size is unknown, but an estimate is given by Van Horn et al.

[8]We provide no reference to their source code to avoid confusion. Our implementation is in many ways a generalization of theirs. Of course, they do own the credit for open sourcing their implementation

Algorithm	Pruning	Heuristic	Non-Hamiltonicity Checks
Depth-First	None	None	None
Cetal's	None	High	None
Van Horn's	None	Low	None
Martello's	Solution, Path	Low	Degree
Rubin's	Solution, Path, Neighbour	None	Degree, One-Connectedness, Disconnectedness, Premature Closure
Vacul's	Solution, Path, Neighbour	Low	Degree, One-Connectedness, Disconnectedness

TABLE I. Overview of all the techniques used by the algorithms that are examined in this research. For the pruning methods, it is displayed if an algorithm uses that method. For the branching heuristic it is presented if an algorithm uses one and if so which one.

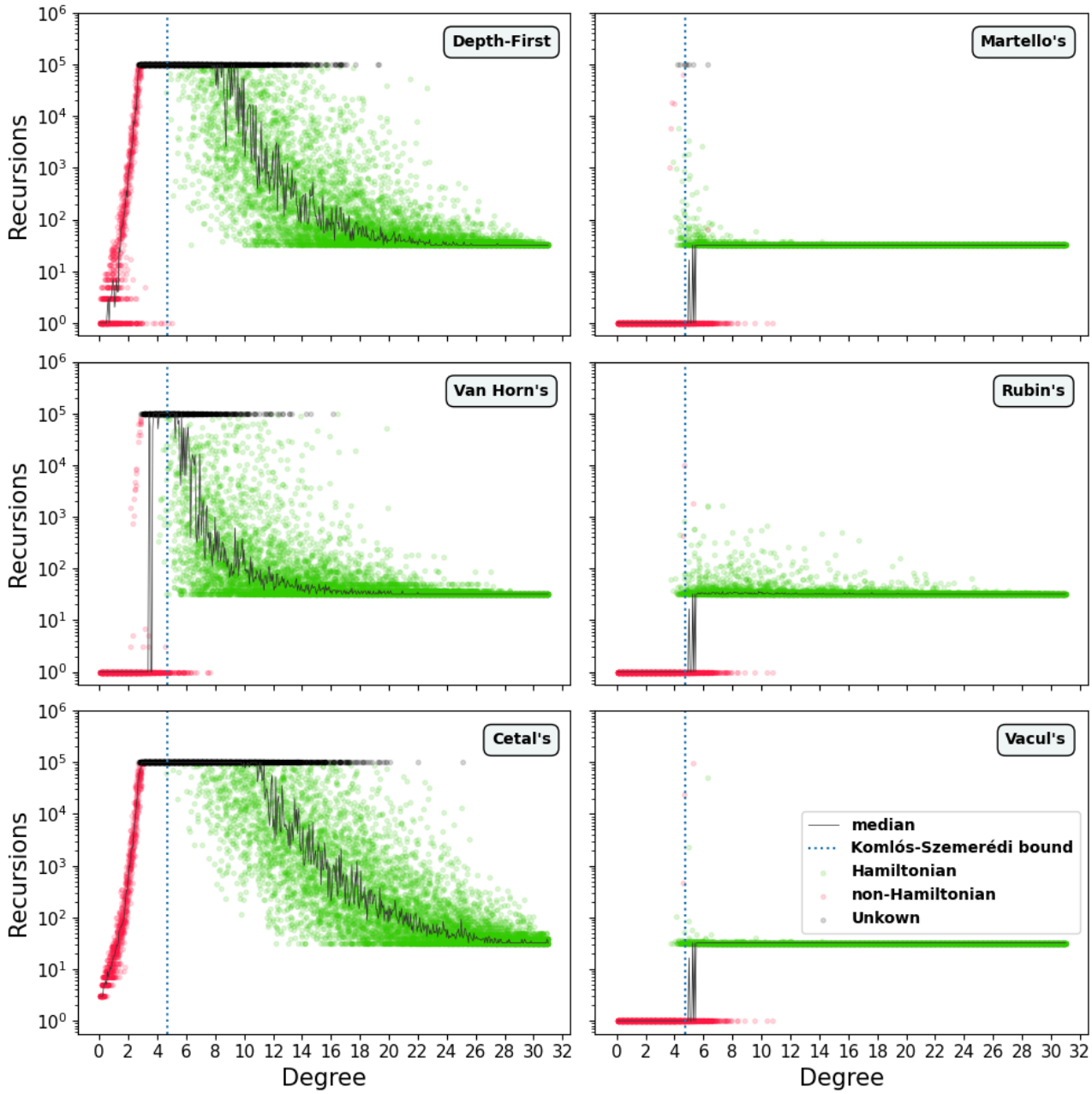


Fig. 4. The number of recursions required for solving the 9920 random graphs of $v = 32$. For all algorithms, the hardest graphs are situated close to the Komlós-Szemerédi bound of $ln(32) + ln(ln(32)) \approx 4.71$ where the probability of being Hamiltonian transitions from zero to one. But while the choice of branching heuristic clearly makes a difference (Depth-first, Van Horn's, Cetal's) the improvement is most dramatic when the algorithm also implements edge pruning and non-Hamiltonicity check procedures (Martello's, Vacul's, Rubin's).

of items in the list. Only if it is equal to v, the graph is connected. Note that in some cases, graphs filtered out by this subroutine can also be filtered out by the degree check; it could be considered a hint that the order in which search pruning subroutines are applied also matters for the computation time. The fourth and final check is the **one-connectedness check**. A graph is one-connected if it contains a vertex that if removed, breaks up the graph into two or more disconnected parts. Such a vertex is commonly dubbed an *articulation point*, and Tarjan's algorithm finds all articulation points in $O(|v| + |e|)$ time [36]. To construct a Hamilton Cycle, there need to be at least two edge-disjoint paths between any two non-adjacent vertices. Therefore, there cannot be a Hamilton cycle in a 1-connected graph. Rubin's is the only algorithm that deploys this technique, we do not know how, but we assumed it could have been done with Tarjan's algorithm, which was published two years earlier. In any case, that is how *we* implemented the check, as literature gave us no definitive answer.

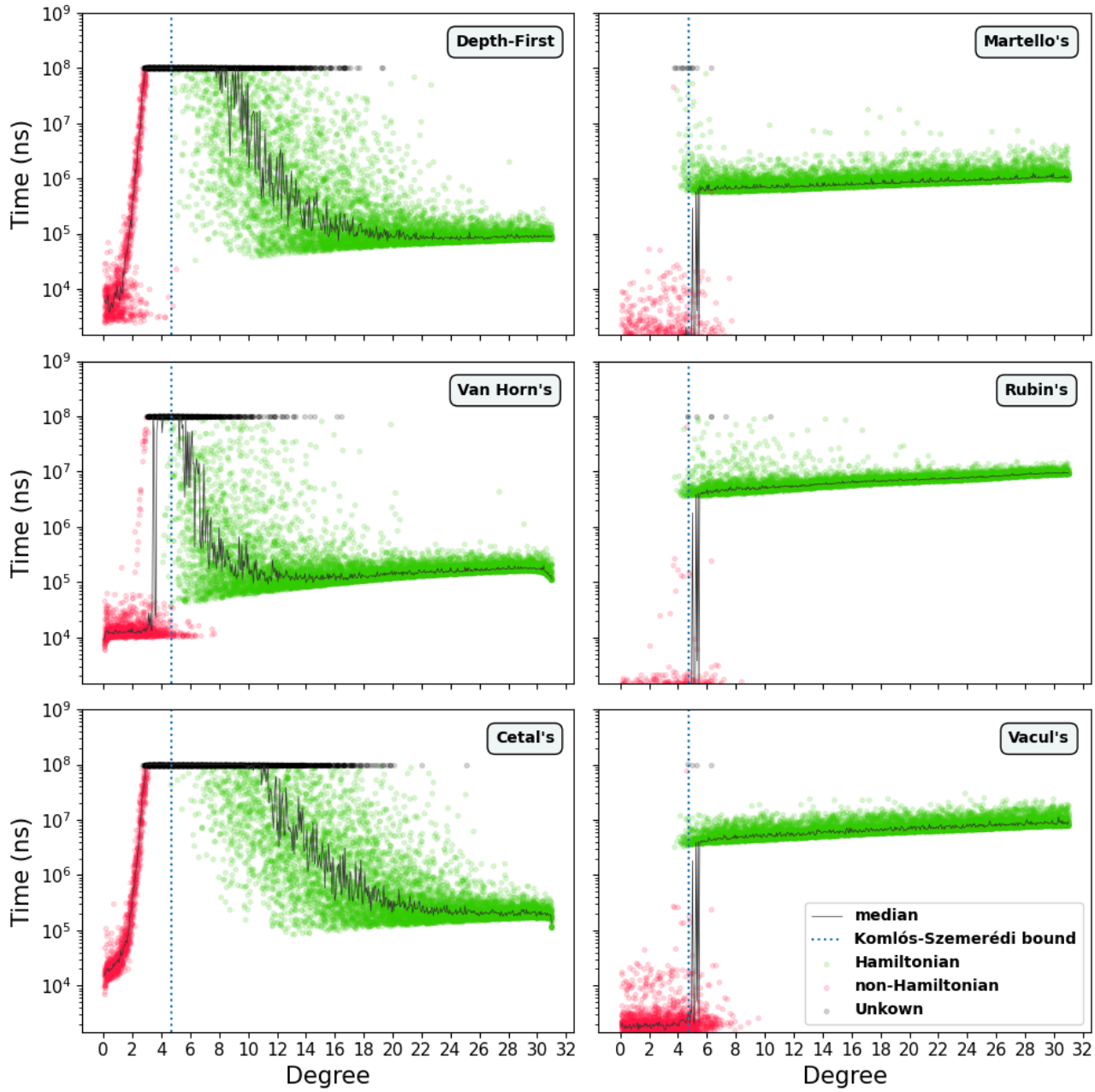


Fig. 5. Procedures for pruning and non-Hamiltonicity checks (Martello's, Vacul's, Rubin's) clearly take 'real' time, but even on graphs as small as these, they pay off – the right hand side triplets have close to zero cutoffs. In these algorithms, the dramatic difference in time consumption between the red and green dots around the Komlós-Szemerédi bound show that these procedures save time especially on *non-Hamiltonian* graphs.

VII. Three Advanced Algorithms

Martello's algorithm *'595'* is a rare example of an algorithm from the 80's that actually came with written source code in FORTRAN when published. The practice of supplying 'open source' code along scientific experiments is a practice that is much valued today, but was definitely ahead of time in 1983. Martello's algorithm was designed with the purpose of finding *all* Hamiltonian cycles in *directed* graphs. We made the smallest possible changes, adapting the algorithm to undirected graphs, and made it halt after the first solution. Martello's algorithm has both solution pruning and path pruning, the latter of which is repeatedly applied in each recursion. Furthermore, it uses the *low* degree preference in its branching heuristic, preferring sparselier connected vertices over denslier connected vertices.

Like Silvano Martello, Frank Rubin was an influential researcher on early algorithms for NP-complete problems and like Silvano Martello, he designed an algorithm for the *directed* Hamiltonian cycle problem, which we adapted to undirected graphs. **Rubin's algorithm** was quite sophisticated

for the time (1974) and in retrospect would have outperformed most other algorithms. It deploys solution, path and neighbour pruning exhaustively in each recursion, and additionally performs all four checks for non-Hamiltonicity.

The premature closure check in this algorithm is worth giving some thought, as it might be redundant in combination with the check for disconnectness. We think that if a set of required edges forms a closed cycle smaller than v, the graph is automatically disconnected. Furthermore, this algorithm is the only one that performs a check for one-connectedness. In the original paper it is not specified how this check is done, but as stated earlier, it is well possible that this was Tarjan's algorithm and in any case, we implemented it as such. Finally, Rubin does not fully specify his disconnectedness check, but does say it runs in quadratic time. So does ours, and chances are they are nearly identical.

These combined features make it a very advanced algorithm for the time. In fact, no other algorithm in our ensemble deploys so many different subroutines. The only thing it does not employ however, is a branching heuristic. It assumedly just picks the next vertex from the data structure. We think that if Frank Rubin would have just implemented the low-degree branching heuristic too, it would have been the most efficient backtracking algorithm to date. And given that it was contrived nearly half a century ago, these minute details could have profoundly changed the course of history for this problem.

The final algorithm included in this research is the algorithm by Vandegriend and Culberson, abbreviated to **Vacul's algorithm** [29]. It was designed for the undirected Hamiltonian cycle problem and uses solution pruning, path pruning, and neighbor pruning, the last two of which are run exhaustively each recursion. It also uses the degree check, the disconnectedness check and the one-connectedness check, 24 after Frank Rubin introduced them in his algorithm. Since Rubin's study is not referenced by VaCul, it is possible they invented these routines themselves. We can only guess the reasons, but the lack of widely accessible internet papers at the time is at least one likely culprit.

It is sad to see that history missed so many beats, but again, these were the days from before internet and the synchronisation of information; now is the time to make up for it. Important to note is that Vacul's algorithm has an additional feature, a *random restart*, which we left out. Even though stochastically speaking, random restarts on long backtrack runs can save computation time, it makes the algorithm a lot harder to compare to the other algorithms. Vacul's paper comes with a quantitative test on a set of random graphs of variable average degree.

VIII. Experiment and results

Analogous to previous studies by Cetal, Vacul and Van Horn et al., three sets of randomly generated undirected graphs were created: one with $v = 16$ vertices, one with $v = 24$ vertices and one with $v = 32$ vertices (we will only show and discuss $v = 32$, results are comparable for other values of v). We created 20 graphs for every number of edges $e = \{1, 2, 3, ..., \frac{1}{2}v(v-1)\}$, resulting in 2400 graphs to solve for $v = 16$. For $v = 24$, the procedure resulted in 5520 graphs, and 9920 random graphs were generated for the 32-vertex set. This amounts to a subtotal of 17,840 graphs, each of which has been solved by all six algorithms *twice* - once for recursions and once for system time. This means the whole investigation comprises 214,080 runs and therefore, to keep things a little insightful, we will show and discuss results of the 9920-piece ensemble for $v = 32$ only. It should be clearly understood though, that the results obtained from the different algorithms in both the time subexperiment and the recursion subexperiment, came from the *same* 17,840 source graphs, facilitating a direct comparison.

We first solve (i.e.: decide) all graphs for Hamiltonicity for all six algorithms implemented on a general code base which is publicly accessible [35]. Programming the generalized algorithm required some interpretation, as none of the algorithms was published with directly usable source code,

Algorithm	#cutoff	#finished	avg. recs	stddev. recs	total (x1000)
Depth-First	1959	7961	2992	11661	219,716
Van Horn's	814	9106	844	5890	8,908
Cetal's	2571	7349	4551	13980	29,054
Martello's	8	9912	40	686	1,193
Rubin's	0	9920	31	106	307
Vacul's	0	9920	44	1136	438

TABLE II. Recursions required by all six algorithms on the entire ensemble of random graphs. Note that 'average recursions' and 'stddev recursions' apply to the *finished* problem instances only.

Algorithm	#cutoff	#finished	avg. ns	stddev. ns	total (x 1M)
Depth-First	1916	8004	3054626	11341555	216,049
Van Horn's	813	9107	913606	5524056	89,620
Cetal's	2583	7337	4695537	13464862	292,751
Martello's	14	9906	864532	1502152	9,964
Rubin's	7	9913	6080907	4439895	60,980
Vacul's	4	9916	6125890	3718473	61,144

TABLE III. Runtime required by all six algorithms on the entire ensemble of random graphs. Average and stddev apply to the *finished* problem instances only.

but incorporates the three options for branching heuristic, the three edge pruning routines and the four non-Hamiltonicity checks, all in subroutines that can be toggled individually. As such, there is no difference between the branching heuristic in Depth-first and Rubin's, or the the degree-check in Martello's and Vacul's, even if historically speaking, they might not have been identical. The only clearly different feature is the random restart option, which is removed from Vacul's algorithm. There is a lot to say about this procedure, mostly that it *can* be very useful for decision problem instances that *do* have a solution. A somewhat smaller adaptation might be the disconnectedness check and the one-connectedness check, both from Rubin's algorithm. We suspect our quadratic implementation might be conceptually equal to Rubin's, which is also reported as a quadratic order, but eludes further specification. Finally, the sorting procedure for the branching heuristic can be done in various ways whose theoretical and practical runtimes may differ, though not to extend of becoming superpolynomial. We suspect that this routine too, is equal to all in literature, but we can not be sure. Summarizing, all algorithmic components are uniform, deterministic and adapted to undirected graphs.

We ran two experiments: one recording the number of recursions, and the other recording system time. Traditionally speaking, number of recursions is the way to go. Measuring recursions is immune to the choice of programming language, compiler used, processor speed, parallelization or incoming resource expenditures such as downloads or updates that might interfere with runtimes. Second, the number of recursions computationally speaking largely outweighs pruning and check procedures, which are usually of lower order complexities - typically polynomial versus factorial. Finally, and following previous two reasons, the order complexity of the algorithm also rest with the number of recursions. As, in a broader view, it is the order complexity of algorithms for NP-complete problems that hinge the $P \stackrel{?}{=} NP$ problem, any results which might influence the view on this problem must also be expressed in recursions. However, as we wanted to extend beyond the reach of theoretical computer science and simply get an estimate of time it takes *in the real world* to mobilize the alleged 'optimization procedures', and immediately question the 'real' impact. A cutoff for the maximum number of a run was 10^5 recursions, and 10^8 nanoseconds[9]

In the recursion experiment, the cutoff was often reached for the three basic algorithms (Fig. 4), leaving 20% of the instances unsolved for depth-first, 8% for Van Horn's algorithm and a toe-curling 26% for Cetal's algorithm (for exact values, see Table II). Of the more advanced algorithms, Martello's was unable to solve 0.1% of the instances, but Rubin's and Vacul's sucessfully solved them all. The average number of recursions needed for the solved instances further confirms the hierarchy in the three basic algorithms, and the success of the search pruning efforts. For the advanced algorithms, the ranking slightly changes, with Martello's having the second lowest computational effort within the solved graphs. This changes nothing however for the hierarchy of algorithmic performance when measured in recursions: **1.Rubin's**, **2.Vacul's**, **3.Martello's**, **4.Van Horn's**, **5.Depth-first**, **6.Cetal's**.

For the time experiment, the cutoff value was again reached often for the three basic algorithms (Fig. 5). These results were largely proportional to the recursions-experiment, leaving 19% of the instances unsolved for depth-first, 8% for Van Horn's algorithm and a painful 26% for Cetal's algorithm (for exact values, see Table III). The picture changes slightly for the more advanced algorithms however, where Martello's, Rubin's and Vacul's algorithms left 14, 7 and 4 graphs unsolved, still well below 0.2% for all three algorithms. Still, it should be noted that *relatively* speaking, this is a huge increase in failure (75%

[9]We could have chosen microseconds, milliseconds or even plain seconds here, but this unit holds the best balance between explanation and visualization such as seen in Fig. 5.

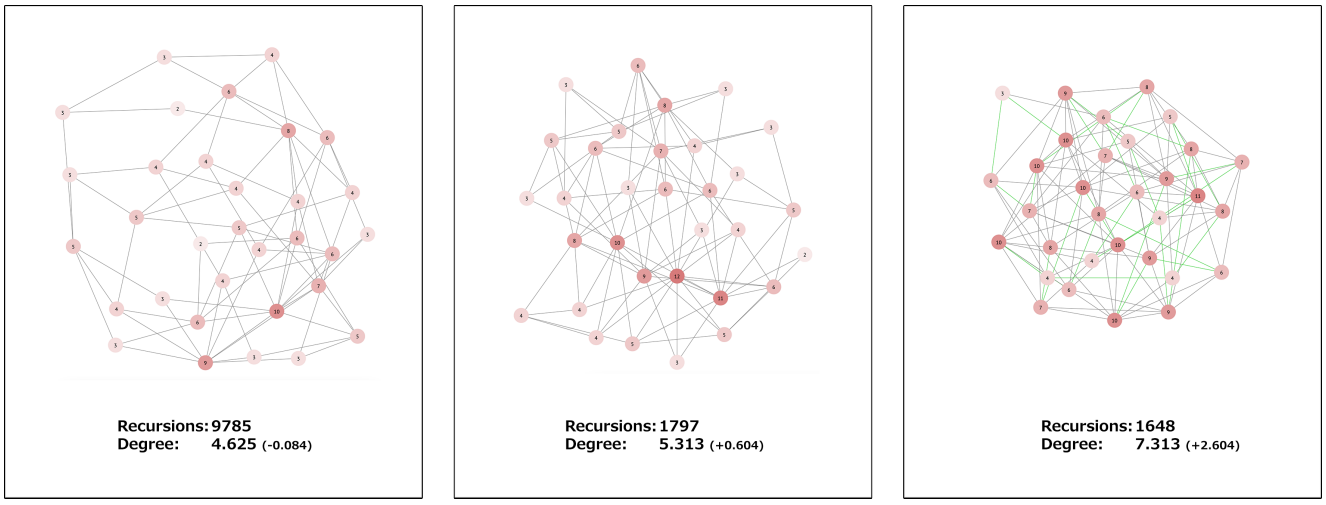


Fig. 6. The three hardest graphs for Rubin's algorithm, the best performing algorithm in this study, support the generalized conclusion that hard instances reside near the Komlós-Szemerédi bound of $ln(32) + ln(ln(32)) \approx 4.71$ (the 'distance' in edge degree is in brackets). The suspicion lingers however, that this is only true for *randomly generated* instances ensembles, and targeted results find harder, denser graphs.

more failures for Martello's alone). It also swaps the top two contestors in the hierarchy of algorithmic performance relative to the recursion-experiment. The best performing algorithms when measured in system time are: **1.Vacul's**, **2.Rubin's**, **3.Martello's**, **4.Van Horn's**, **5.Depth-first**, **6.Cetal's**.

IX. Conclusion and Discussion

It is a fresh awakening that even for graphs of $v = 32$, which is relatively small, and the theoretical computational effort required relatively futile, the optimization procedures still make such a large difference. Generally speaking, these procedures appear to pay off in the number of recursions required to decide a graph, especially further away from the Komlós-Szemerédi bound. When measured in system time however, a different view emerges. The average time to decide a graph hugely increases, especially in denser regions away from the Komlós-Szemerédi bound.

The big question is how these effects scale up, and whether the increase in search pruning effort really pays off on a larger scale, and whether that counts for recursions, for system time, or for both. Furthermore, the distribution of edge pruning techniques, branching heuristics and non-Hamiltonicity checks among the algorithms is largely haphazard. This is because for now, we tried to stick as closely as possible to historical conventions handed down to us from literature, but a more structured approach might be desirable. Furthermore, it is quite surprising to see the Rubin's algorithm, which is the second oldest algorithm, overaging Van Horn's by more than four decades, contained the most sophisticated procedures, but also performed best. On the other side, it is a little desillusioning to find Cetal's algorithm, so oftenly cited, finishing last in this test. Finally, some thought must be given to whether other edge pruning techniques, branching heuristics or checks could be devised, and in what order such procedures should be applied. Last but not least, it *might* still pay off to dynamically rearrange the data structure in some way as to make the degree preference in branching dynamically applicable. It should be noted though that this might be programmistically tough, as backtracking would the require un-sorting the dynamically rearranged data structure.

For this study, we made 20 graphs for every possible edge density. Although this procedure is rigourously systematic, it also allows for some skewness in the results: in the end, there is structurally speaking, just one possible graph with one edge. On the other end of the density spectrum, we have 20 'random' graphs with with 496 edges, all being fully connected, and thereby isomorphic. As the number of different graphs for v vertices and e edges is equal to $\binom{\frac{1}{2}v(v-1)}{e}$, the maximum number of different graphs can be found at 248 edges for 32 vertices, which is exactly halfway the Figures 4 and 5. So if every existible graph of for $v = 32$ was equally likely, most of the graphs would a) be Hamiltonian and b) be easy, at least for the best three of our algorithms. For larger graphs, this effect grows stronger as the Komlós-Szemerédi bound grows (double) logarithmically in v, whereas the peak existence grows (half) quadratically. A really weird but inescapable conclusion therefore is, that for larger values of v, nearly all instances are Hamiltonian, and many of those might be easy. This does not hold for our results, or that of Cetal, Vacul and Van Horn et al., who all 'columnized' their graphs into different degree categories. But combined, the combinatorial assessment and the columnized results show that the prediction of instance hardness for the Hamiltonian cycle problem critically relies on the a priori availability of information of the graph(s) to be solved, even if one future algorithm turns out to be superior.

And this brings us to the last point, tying the discussion back to the preamble. It has been shown that for at least one algorithm, Vacul's in this case, the hardest instances are non-Hamiltonian, very far away from the Komlós-Szemerédi bound, in a very dense region of the combinatorial space [8] [7] (also see [52]). This observation apparently contradicts nearly everything that was written in the previous paragraph, but can be explained from the randomness in large ensembles of problem instances such as used in this study. The found extremely hard instances were *structured*, and abiding by the teachings of A. N. Kolmogorov, structured objects in large randomized ensembles are rare [40]. So finding these instances paradoxically means knowing where to look, and one way to do that is to use a parameter-unsensitive evolutionary algorithm such as Plant Propagation [41] [42] [43] [44] [45] [50] [51]. For very large graphs, one could better resort to a single-individual search heuristic, such as HillClimbing or simulated annealing [46] [47] [48]. On the other end of the scale, the use of large randomized ensembles for algorithmic performance raises some questions. Although proper benchmarking is becoming a hot topic [49], the issue of randomness herein is still seldomly discussed and needs more attention.

X. Future Work

After the results of this study, the road ahead took a few sudden turns. What needs to happen next, is to see if one further generalized backtracking algorithm can convincingly dominate all others. Such a study should include combinations of heuristics and search pruning that are not tested as yet, and it is our belief that such an algorithm exists. After that, its hardest instances should be found, possibly by a global evolutionary search algorithm. Considering earlier results, it could resemble the graph found by Sleegers & Van den Berg in 2020, as its structure suggests it might be hard for backtracking algorithms in general. Only after that, hardness hierarchies, and possible implications for computability, and the $P \stackrel{?}{=} NP$ problem itself might be found, but progress in this direction will depend on future developments.

XI. Acknowledgements

Thanks to Gijs van Horn for maintaining the nice interactive Hamiltonian cycle problem up to date (https://hamiltoncycle.gijsvanhorn.nl/).

REFERENCES

[1] G. van Horn, R. Olij, J. Sleegers, and D. van den Berg, "A predictive data analytic for the hardness of hamiltonian cycle problem instances," DATA ANALYTICS 2018, p. 101, 2018.

[2] P. C. Cheeseman, B. Kanefsky, and W. M. Taylor, "Where the really hard problems are." in IJCAI, vol. 91, 1991, pp. 331–340.

[3] J. Sleegers, R. Olij, G. van Horn, and D. van den Berg, "Where the really hard problems aren't," Operations Research Perspectives, vol. 7, p. 100160, 2020.

[4] D. van den Berg, "Video showing where cetal's results on atsp are flawed (part1)," https://www.youtube.com/watch?v=zfqmiLkbCRc, 2020.

[5] ——, "Video showing where cetal's results on atsp are flawed (part2)," https://www.youtube.com/watch?v=l8W_GVzqDnk, 2020.

[6] G. van Horn, "Interactively viewable hamiltonian cycle hardness," https://hamiltoncycle.gijsvanhorn.nl/, 2020.

[7] J. Sleegers and D. van den Berg, "Plant propagation & hard hamiltonian graphs," Evo* 2020, p. 10, 2020.

[8] ——, "Looking for the hardest hamiltonian cycle problem instances," in IJCCI, 2020, p. 40–48.

[9] A. Huxley, Brave new world. Ernst Klett Sprachen, 2007.

[10] R. E. Bryant, "On the complexity of vlsi implementations and graph representations of boolean functions with application to integer multiplication," IEEE transactions on Computers, vol. 40, no. 2, pp. 205–213, 1991.

[11] F. Ivančić, Z. Yang, M. K. Ganai, A. Gupta, and P. Ashar, "Efficient sat-based bounded model checking for software verification," Theoretical Computer Science, vol. 404, no. 3, pp. 256–274, 2008.

[12] M. Krötzsch, "Complexity theory, lecture 6: Nondeterministic polynomial time," https://iccl.inf.tu-dresden.de/w/images/0/08/CT2019-Lecture-06-overlay.pdf, 2019.

[13] T. Larrabee and Y. Tsuji, Evidence for a satisfiability threshold for random 3CNF formulas. Citeseer, 1992.

[14] S. Kirkpatrick and B. Selman, "Critical behavior in the satisfiability of random boolean expressions," Science, vol. 264, no. 5163, pp. 1297–1301, 1994.

[15] I. P. Gent and T. Walsh, "Easy problems are sometimes hard," Artificial Intelligence, vol. 70, no. 1-2, pp. 335–345, 1994.

[16] T. Hogg and C. P. Williams, "The hardest constraint problems: A double phase transition," Artificial Intelligence, vol. 69, no. 1-2, pp. 359–377, 1994.

[17] T. Hogg, "Refining the phase transition in combinatorial search," Artificial Intelligence, vol. 81, no. 1-2, pp. 127–154, 1996.

[18] I. P. Gent and T. Walsh, "The tsp phase transition," Artificial Intelligence, vol. 88, no. 1-2, pp. 349–358, 1996.

[19] M. R. Garey and D. S. Johnson, Computers and intractability. wh freeman New York, 2002, vol. 29.

[20] A. M. Jaffe, "The millennium grand challenge in mathematics," Notices of the AMS, vol. 53, no. 6, pp. 652–660, 2006.

[21] P. Erdos and A. Rényi, "On the evolution of random graphs," Publ. Math. Inst. Hung. Acad. Sci, vol. 5, no. 1, pp. 17–60, 1960.

[22] L. Pósa, "Hamiltonian circuits in random graphs," Discrete Mathematics, vol. 14, no. 4, pp. 359–364, 1976.

[23] J. Komlós and E. Szemerédi, "Limit distribution for the existence of Hamiltonian cycles in a random graph," Discrete Mathematics, vol. 43, no. 1, pp. 55–63, 1983.

[24] S. Roberts and B. Flores, "Systematic generation of Hamiltonian circuits," Communications of the ACM, vol. 9, no. 9, pp. 690–694, 1966.

[25] M. Held and R. M. Karp, "A dynamic programming approach to sequencing problems," Journal of the Society for Industrial and Applied Mathematics, vol. 10, no. 1, pp. 196–210, 1962.

[26] R. Bellman, "Dynamic programming treatment of the travelling salesman problem," Journal of the ACM (JACM), vol. 9, no. 1, pp. 61–63, 1962.

[27] S. Martello, "Algorithm 595: An enumerative algorithm for finding Hamiltonian circuits in a directed graph," ACM Transactions on Mathematical Software (TOMS), vol. 9, no. 1, pp. 131–138, 1983.

[28] F. Rubin, "A search procedure for Hamilton paths and circuits," Journal of the ACM (JACM), vol. 21, no. 4, pp. 576–580, 1974.

[29] B. Vandegriend and J. Culberson, "The gn, m phase transition is not hard for the Hamiltonian cycle problem," Journal of Artificial Intelligence Research, vol. 9, pp. 219–245, 1998.

[30] B. Bollobas, T. I. Fenner, and A. M. Frieze, "An algorithm for finding Hamilton paths and cycles in random graphs," Combinatorica, vol. 7, no. 4, pp. 327–341, 1987.

[31] A. Bjorklund, "Determinant sums for undirected Hamiltonicity," SIAM Journal on Computing, vol. 43, no. 1, pp. 280–299, 2014.

[32] K. Iwama and T. Nakashima, "An improved exact algorithm for cubic graph tsp," in International Computing and Combinatorics Conference. Springer, 2007, pp. 108–117.

[33] S. Even, "Graph algorithms," 1979.

[34] C. Gomes and B. Selman, "On the fine structure of large search spaces," in Proceedings 11th International Conference on Tools with Artificial Intelligence. IEEE, 1999, pp. 197–201.

[35] J. Sleegers, "Source code," https://github.com/Joeri1324/What-s-Difficult-About-the-Hamilton-Cycle-Poblem-, 2020.

[36] R. Tarjan, "Depth-first search and linear graph algorithms," SIAM journal on computing, vol. 1, no. 2, pp. 146–160, 1972.

[37] E. Lucas, "Recreations mathematiques, four volumes: Gautheir–villars," Paris, France (1882/1894), pp. 161–197, 1882.

[38] D. van den Berg, F. Braam, M. Moes, E. Suilen, S. Bhulai et al., "Almost squares in almost squares: solving the final instance," 2016.

[39] S. S. Skiena, "The algorithm design manual," p. 247, 1998.

[40] M. Li, P. Vitányi et al., An introduction to Kolmogorov complexity and its applications. Springer, 2008, vol. 3.

[41] A. Salhi and E. S. Fraga, "Nature-inspired optimisation approaches and the new plant propagation algorithm," 2011.

[42] M. De Jonge and D. van den Berg, "Parameter sensitivity patterns in the plant propagation algorithm." in IJCCI, 2020, p. 92–99.

[43] M. de Jonge and D. van den Berg, "Plant propagation parameterization: Offspring & population size," Evo* 2020, p. 19, 2020.

[44] M. Paauw and D. Van den Berg, "Paintings, polygons and plant propagation," in International Conference on Computational Intelligence in Music, Sound, Art and Design (Part of EvoStar). Springer, 2019, pp. 84–97.

[45] W. Vrielink and D. van den Berg, "Fireworks algorithm versus plant propagation algorithm," 2019.

[46] R. Geleijn, M. van der Meer, Q. van der Post, and D. van den Berg, "The plant propagation algorithm on timetables: First results," EVO* 2019, p. 2, 2019.

[47] R. Dahmani, S. Boogmans, A. Meijs, and D. van den Berg, "Paintings-from-polygons: simulated annealing," in International Conference on Computational Creativity (ICCC'20), 2020.

[48] S. Kirkpatrick, C. D. Gelatt, and M. P. Vecchi, "Optimization by simulated annealing," science, vol. 220, no. 4598, pp. 671–680, 1983.

[49] T. Bartz-Beielstein, C. Doerr, J. Bossek, S. Chandrasekaran, T. Eftimov, A. Fischbach, P. Kerschke, M. Lopez-Ibanez, K. M. Malan, J. H. Moore et al., "Benchmarking in optimization: Best practice and open issues," arXiv e-prints, pp. arXiv–2007, 2020.

[50] W. Vrielink and D. van den Berg, "A Dynamic Parameter for the Plant Propagation Algorithm," EVO* 2021, p. 5, 2021.

[51] W. Vrielink and D. van den Berg, "Parameter control for the Plant Propagation Algorithm," EVO* 2021, p. 1, 2021.

[52] D. van den Berg and P. Adriaans, "Subset Sum and the Distribution of Information." in IJCCI, 2021, p. 134-140

Severe Weather-based Fire Department Incident Forecasting

Guido Legemaate*, Jeffrey de Deijn†, Sandjai Bhulai‡ and Rob van der Mei‡§

*Fire Department Amsterdam-Amstelland, Ringdijk 98, 1097 AH Amsterdam
Email: g.legemaate@brandweeraa.nl
†Vrije Universiteit Amsterdam, Faculty of Science, De Boelelaan 1081, 1081 HV Amsterdam
Email: jeffreydeijn@hotmail.com
‡Vrije Universiteit Amsterdam, Department of Mathematics, De Boelelaan 1081, 1081 HV Amsterdam
Email: s.bhulai@vu.nl
§Centrum Wiskunde en Informatica, Science Park 123, 1098 XG Amsterdam
Email: mei@cwi.nl

***Abstract*—For fire departments, having enough firefighters available during a shift is obviously an important requirement. Nevertheless, just like in any organization, having too many firefighters standby is not desirable from a financial point of view. Despite the fact that fire departments can and should not be run like production companies, at least for staffing purposes, forecasting the number of incidents that each fire station has to handle is highly relevant. In this paper, we develop models to create a forecast for the number of incidents that each fire station in the Dutch safety region Amsterdam-Amstelland has to handle for specific incident types and deal with major and small incidents. Previous studies mainly focused on multiplicative models containing correction factors for the weekday and time of year. Our main contribution is to incorporate the influence of different weather conditions in the categories of wind, temperature, rain, and visibility. Rain and wind typically have a strong linear influence, while temperature mainly has a non-linear influence. We show that an ensemble model has the best predictive performance.**

Keywords–incident forecasting; fire department planning; generalized linear models; ensemble models; severe weather conditions.

I. Introduction

As for most organizations, the ability to accurately forecast demand is of "paramount importance" for emergency services, fire departments included [1][2]. In the 1970s, the Fire Department of the City of New York and The New York City-Research And Development (RAND) Institute jointly conducted various groundbreaking studies [3]. More recent academic interest seems to be focused more on ambulance services. While there are obvious similarities between emergency service providers, they differ in (the number of) incident types, demand characteristics, and operational logistics.

Nevertheless, the problems that fire departments have to deal with, like loss of coverage and the degradation of response times, are similar. The same is true for possible gains. At a strategic and tactical level, improved forecasting of workload leads to a better placement of base stations, and improved staffing and scheduling. At an operational level, one may pro-actively relocate units to maximize coverage and minimize response times during major incidents [4]. All things considered, efficient planning of emergency service resources is crucial.

Demand is an important factor when models are being developed to improve the performance of emergency service providers. It is, however, not uncommon that, for instance, call arrival rates are estimated using ad-hoc or rudimentary methods such as averages based on historical data [5]. This may ultimately lead to a degradation of performance, or over- or under-staffing [6]. In most cases, reducing response times is an important performance measure since this increases the survival rate of victims [7][8].

Numerous papers have been written on forecasting forest or wildfire occurrences, many of those using weather variables and vegetation types as part of their model [9]. Forest fire forecasting is no longer a study in academia alone. In fact, in the United States, e.g., the National Interagency Coordination Center operates a predictive service which provides decision support to the United States Forest Service, which facilitates pro-active management and planning of fire assets on both operational and tactical levels [10].

Although the scale of wildfire occurrences in the Netherlands is smaller than in many other parts of the world, it is mainly the greater interrelationship of different types of infrastructure, i.e., the wildland-urban interface, that causes concern and even lead to surface fuel models for the Netherlands [11]. For a more urban environment, like the conurbation of Western Holland, which also includes Amsterdam, forest fire occurrences are not very common.

The occurrence of certain types of incidents which fire departments in urban settings typically respond to also correlate with weather conditions. As such, incorporating this information into the planning process of emergency services yields important advantages over current practice. Typical weather and storm-related incidents that fire departments in the Netherlands respond to are fallen trees, potentially falling debris that needs securing (roofs, construction work, scaffolding), and water damage. Another important factor is that the weather also impacts fire department operations by overwhelming available resources.

At least in the Netherlands, to the best of our knowledge, there are no known applications of forecasting algorithms that are used in practice at fire departments, being urban or specialized forest services. Given this, we aim to provide an easily applicable model that can be put to use for a general fire department when dealing with severe weather conditions. Therefore, we quantify and model the fact that - under these conditions - fire departments experience an increased amount of incidents, which in itself leads to an increased amount of deployments.

The organization of this paper is as follows. In Section II, we describe the data used to obtain the forecasts. Section III describes the models used for forecasting. In Section IV, we analyze the performance of the models and state the insights. Finally, in Section V, we conclude and address a number of topics for further research.

II. DATA

The available data contains one row for each incident that happened in the region Amsterdam-Amstelland from January 2008 up until April 2016. The most interesting information includes the incident's start- and end time, location, incident type, the concerned fire station, and the number of fire trucks used. Since the size of incidents matters for the number of people you need, the focus is on forecasting the number of trucks needed.

A. Major and small incidents

The vast majority of incidents require only one or otherwise just a few trucks. Therefore, it makes sense to distinguish between 'major' and 'small' incidents. Major incidents are mostly due to coincidences that are hard to predict. Specifically, they do not rely on bad weather conditions or a particular time of the year in the Netherlands, for example, as with forest fires in countries with a tropical climate. This arouses the expectation that the inter-incident times of major incidents can be modeled as a Poisson process.

To test the Poisson assumption, we apply the Kolmogorov-Smirnov (KS) test on the inter-incident times in cases when more than k trucks are needed for several values of k. The KS-test shows that if we define an incident as 'major' when at least $k = 6$ trucks are used, then the KS-test does not reject exponentially of the inter-incident times (approximate p-value $= 0.429$). However, for values of $k < 6$, the KS-test doubts (or rejects) this exponentially (approximate p-value $= 0.073$ and 0.002 when at least $k = 5$ and $k = 4$ trucks are used, respectively. Hence, according to this result, we define an incident to be major when at least six trucks are needed.

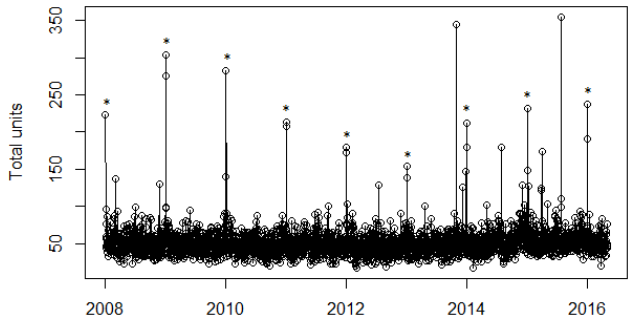


Figure 1. Total number of trucks used for small incidents per day. * *Peaks caused due to an increased amount of incidents around New Year's Eve.*

Next, we focus on the small incidents. Small incidents are probably easier to predict, since bad weather conditions often cause many *small* incidents to happen (like fallen trees, water damage, or police/ambulance assistance at traffic accidents). To study this, we first omit all incidents on December 31 and January 1. There are extremely many incidents around New Year's Eve as can be seen on Figure 1, mainly caused by fireworks-related incidents. These conditions do not occur in the rest of the year, therefore we model these days separately as described in the modeling section.

After elimination we find that not all outliers in Figure 1 are New Year's days. In fact, the only five days that, for the amount of trucks used per day (>138), on par with New Year's day are days with severe weather conditions as can be seen in Table I.

On these days with severe weather conditions only 0.46%, instead of an average 1.72%, of incidents are major incidents. Without a clear reason to assume that the frequency of major incidents on this particular type of days is lower, there must be another explanation rather than chance. If so certain circumstances cause many small incidents to happen, like those caused by severe

TABLE I. WEATHER CONDITIONS ON THE FIVE DAYS THAT COULD COMPETE WITH NEW YEAR'S DAYS IN TERMS OF AMOUNT OF TRUCKS USED.

Date (day)	Trucks	Highest windspeed (km/h) Overall	Highest windspeed (km/h) Worst hour	Total rainfall (mm) Overall	Total rainfall (mm) Worst hour
28/10/2013 (Mon.)	345	79.2	111.6	3	10.7
24/12/2013 (Tue.)	147	64.8	111.6	2.3	6.8
28/07/2014 (Mon.)	179	28.8	43.2	12.6	60.5
31/03/2015 (Tue.)	174	64.8	100.8	3.8	7.9
25/07/2015 (Sat.)	355	72	100.8	5.8	19.7

weather conditions. Data from the fire department on incident types that happened on days with severe weather conditions further support this finding.

B. Seasonal patterns

There are clear seasonal patterns in the data for the number of trucks needed throughout each year, week, and day. The plots in Figure 2 illustrate this. The pattern in Figure 2c depicts the activity cycle that an average person goes through every day of the week. The week pattern (Figure 2b) differs per type of incident and looks a little different throughout the year. The pattern in Figure 2a can be included in the model in a more subtle way than taking factors per month. The problem here is that, for instance, the differences between the beginning and end of January are considerable. We correct for this by using a Loess-smoothed function over the factors per week. We will include all these patterns in our model.

C. Weather variables

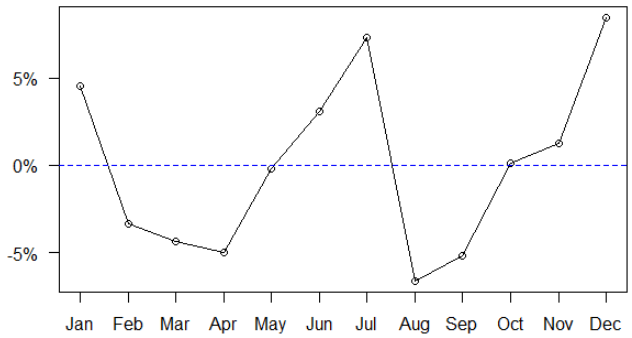
Besides the time-dependent components, we want to know which weather variables we must include in our model. Therefore, we use the Pearson correlation test to determine which weather conditions have a significant influence on the number of trucks we need. The results of these tests are summarized in Table II.

TABLE II. PEARSON'S PRODUCT-MOMENT CORRELATION TESTS BETWEEN SOME WEATHER VARIABLES AND THE NUMBER OF TRUCKS USED FOR SMALL INCIDENTS PER DAY.

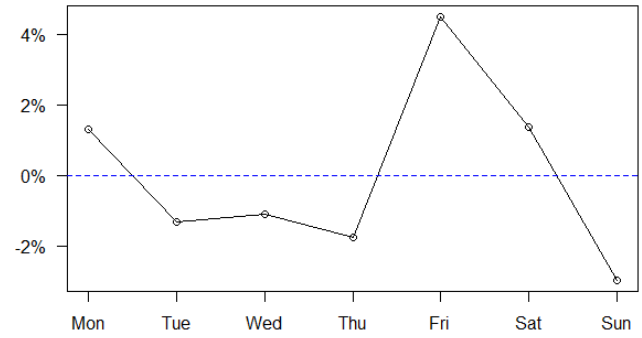
Category	Variable	p-value	Correlation
Wind	Average wind speed (FG)	$< 10^{-12}$	0.132
	Maximum hourly mean wind speed (FHX)	$< 10^{-15}$	0.177
	Maximum wind gust (FXX)	$< 10^{-15}$	0.189
Temperature	Average temperature (TG)	0.6897	0.007
	Boolean: 1 if average > 0 (TG>0)	$< 10^{-8}$	0.105
Rainfall *	Rainfall duration (DR)	0.0004	0.061
	Total rainfall (RH)	$< 10^{-15}$	0.151
	Maximum hourly rainfall (RHX)	$< 10^{-12}$	0.132
Visibility **	Minimum visibility (VVN)	0.2217	-0.014
	Boolean: 1 if minimum $<$ 200m (VVN<2)	0.2893	0.010

** In 0.1 mm and -1 for <0.05 mm; ** On 0-89 scale, where 0: <100 m, 89: >70 km.*

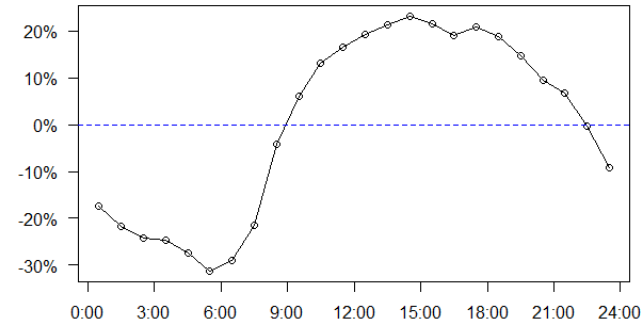
We can see from this that the minimum visibility and the average temperature both have no significant (direct) influence. However, if we consider a variable indicating whether it was on average freezing on that day, then this does have predictive value. Obviously, we also have to include some variables indicating the amount of rainfall and wind. However, the variables within these categories are highly correlated (sample correlation around 0.9) and, therefore, we may exclude some of them to simplify our model.



(a) Year pattern: higher during summer and winter.



(b) Week pattern: peak on Friday.



(c) Day pattern: low at night, high at midday.

Figure 2. Seasonal patterns: the given percentages represent relative differences with respect to the average (in blue).

D. Fireworks-related incidents

It is a tradition in the Netherlands to celebrate New Year's Eve with fireworks. Only then, the general public is allowed to light fireworks. Fireworks need to comply with legal standards, and may only be sold during the

last three days of the year at licensed shops.

Over the years, fire departments in the Netherlands have seen a slow but steady rise in fireworks-related incidents [12]. Most common incident types that fire departments respond to during New Year's Eve are dumpster fires, outside fires and vehicle fires. Also more serious incidents happen, like in 2020 a fire in an Arnhem flat that left two members of a family dead, and two other family members critically injured. The fire began in the ground floor hallway of a high-rise apartment building and was identified to be caused by fireworks. A family of four were found trapped in an elevator, which shut down as the building lost electricity due to the fire. Noteworthy, but not related to New Year's Eve, is the Enschede fireworks disaster of May 13, 2000. A catastrophic explosion in a fireworks depot, situated in a residential area of the eastern Dutch city of Enschede, essentially obliterated the neighborhood of Roombeek [13].

In 2014, in an attempt to mitigate the nuisance caused by fireworks-related incidents on New Year's Eve, the Dutch government reduced the time window in between the public was allowed to set off fireworks. This reduced time window was set to 6 pm on December 31 to 2 am on January 1, while before it was allowed starting from 10 am on December 31.

Figures 3 and 4 show the average number of trucks used for small incidents per hour around New Year's Eve before and after New Year's Eve 2014/2015, respectively. The reduced time window, and possible relation between fireworks and small incidents, seems to be reflected in the average number of trucks used per hour as well. Furthermore, it seems that the reduction has only compressed all incidents into a smaller time window, as the average total number of trucks per hour has increased at certain periods. These preliminary findings however need further research to find out whether this is a coincidence or not.

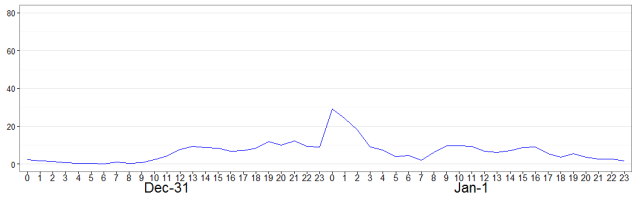


Figure 3. Average number of trucks used for small incidents per hour around New Year's Eve before 2014/2015.

III. MODELS

In this section, we will create a model that predicts directly the number of trucks that each fire station needs.

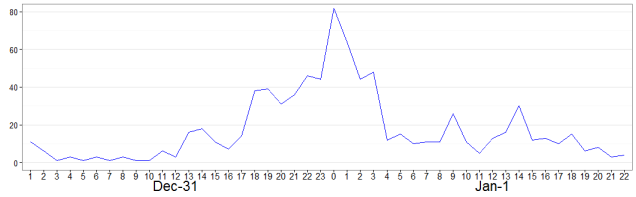


Figure 4. Average number of trucks used for small incidents per hour around New Year's Eve after 2014/2015.

In the previous section, we have shown that the major incidents (with at least six trucks needed) are very hard to predict and that we can best model them by an (inhomogeneous) Poisson process. We also showed that the daily pattern of the number of trucks used for small incidents is quite standard. So, if we know for some day how many trucks are needed in total, we can quite accurately extract from this how many trucks are needed per hour. Therefore, we will try to forecast the number of trucks needed per day per fire station.

Fire departments in general have a variety of incident types they respond to. Not all of them occur frequently enough to make a good forecast on. Since these in this aspect have little value, they are eliminated and the remaining incident types are clustered based on their correlation with certain weather variables.

TABLE III. INCIDENT CLUSTERS AND CORRELATION WITH RESPECT TO WIND SPEED, TEMPERATURE, RAINFALL, AND VISIBILITY.

Cluster	Type	Wind	Temp.	Rain	Visib.	# p/day
1	Outside fire	-0.135	0.09	-0.193	0.075	3.46
2	Animal in water Animal assistance Person in water Locked out	-0.088 -0.072 -0.041 -0.006	0.134 0.129 0.056 0.159	-0.058 -0.088 -0.023 -0.043	0.013 0.069 0.009 0.062	1.65
3	Contamination / nuisance	-	-0.228	0.038	-0.111	2.52
4	Locked in elevator Automated alarm	- -	-0.088 -0.069	0.021 0.051	-0.015 -0.037	8.16
5	Fire rumor Inside fire General assistance water	- - -	-0.103 -0.038 -0.019	- - -	- - -	3.57
6	Police assistance	0.048	-0.062	0.026	-	1.34
7	Ambulance assistance Vehicle in water Reanimation	- - -	-0.065 -0.042 -0.086	- - -	-0.039 -0.025 -0.008	8.55
8	General assistance	0.063	0.079	0.057	0.052	2.28
9	Storm- and water damages	0.319	0.028	0.279	-	2.10

In total, we now have nine different incident clusters in our dataset, some of which occur much more/less often than others. In Table III, we show the correlation with respect to one variable of each four weather categories. Looking at these correlations in detail, we can see that these are often in line with our expectations. For instance, high wind speed and rainfall obviously increase the number of incidents due to 'storm and water damage'

(type 9) and decrease the likelihood of 'outside fires' occurring (type 1).

We will estimate, for each incident type t, a model that predicts the number of trucks used for *small* incidents $y_{t,d}$ on date d, i.e.,

$$y_{t,d} = f_{t,d} \cdot g_{t,d} \cdot x_{t,d}.$$

Here, $f_{t,d}$ is a correction factor for the week number based on a Loess-smoothed function as in Figure 5, and $g_{t,d}$ is a weekday factor as in Figure 2b. Both are computed separately for each incident type. Finally, the term $x_{t,d}$ contains all remaining information. This includes the average level, dependencies on the weather, a possible trend and dependencies on all other variables that we are currently not considering, but which do exist in reality.

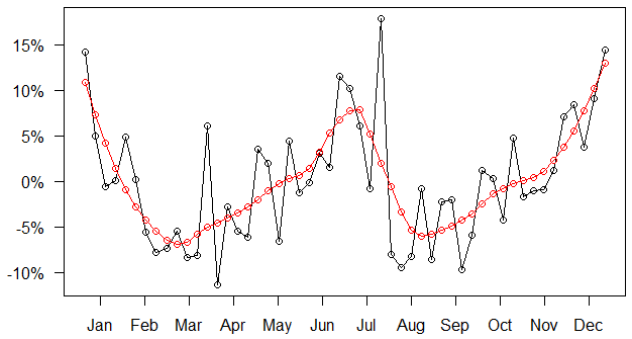


Figure 5. The year pattern per week (in black) together with its Loess-smoothed variant ($\alpha = 0.3$).

A. Linear regression model

The first attempt to model $x_{t,d}$ is by means of the linear regression model (LM)

$$\begin{aligned} x_{t,d} = \beta_0 &+ \beta_1 \cdot d + \beta_2 \cdot \text{windspeed}_d + \beta_3 \cdot \text{temperature}_d \\ &+ \beta_4 \cdot \text{rainfall}_d + \beta_5 \cdot \text{visibility}_d + \epsilon_{t,d}, \end{aligned}$$

where $\epsilon_{t,d}$ is assumed to have expectation zero and some finite variance. Note that this model includes an intercept (β_0), a linear trend ($\beta_1 \cdot d$) and (at most) four weather variables.

B. Generalized Linear Model

Our second model, a Generalized Linear Model (GLM) arises from an observation that the largest outlier neither has the highest wind speed nor the most rainfall. However, the *combination* of wind and rainfall might be the cause. It may, therefore, be a good idea to include also cross-effects in our model, i.e.,

$$\begin{aligned} x_{t,d} = \beta_0 &+ \beta_1 \cdot d + \beta_2 \cdot \text{windspeed}_d + \beta_3 \cdot \text{temperature}_d \\ &+ \beta_4 \cdot \text{rainfall}_d + \beta_5 \cdot \text{visibility}_d \\ &+ \beta_6 \cdot \text{windspeed}_d \cdot \text{temperature}_d \\ &+ \beta_7 \cdot \text{windspeed}_d \cdot \text{rainfall}_d \\ &+ \beta_8 \cdot \text{windspeed}_d \cdot \text{visibility}_d \\ &+ \beta_9 \cdot \text{temperature}_d \cdot \text{rainfall}_d \\ &+ \beta_{10} \cdot \text{temperature}_d \cdot \text{visibility}_d \\ &+ \beta_{11} \cdot \text{rainfall}_d \cdot \text{visibility}_d \\ &+ \epsilon_{t,d}. \end{aligned}$$

Here, $\epsilon_{t,d}$ is again a residual term with zero expectation and some finite variance. Note that this is not a GLM as one may know from the literature: the only feature that causes it to be generalized is that it now also handles the cross-term relations between the weather variables.

C. Random Forests

The Random Forest (RF) algorithm is a machine learning algorithm that can be used for both classification and regression tasks. Compared to LM and GLM it has a large computation time, but RF is often used in practice since it generally has great performance. It will, therefore, be worth a try to implement this algorithm for our regression problem.

As input, the algorithm needs a $T \times (K+1)$-matrix with K explanatory variables and one observation variable (in this case $x_{t,d}$), all of sample size T. In the first iteration of the algorithm, a sample of size T is drawn with replacement from the input matrix. On this sample, a decision tree (DT) algorithm is executed. This procedure is repeated N times, yielding N decision trees. When a new sample comes in, we can take all N predictions for this sample and average these to get the final prediction.

D. Performance measures

To evaluate the different models, we create a train and a test set. The train set contains all data up until 2015/06. The test set contains all data from 2015/07 onwards. This holds for all incident types, so all test sets contain exactly nine months of data and the quality of the forecasts can, therefore, be compared easily. We will measure the quality of a forecast on n samples using the Mean Absolute Percentage Error (MAPE), assuming $y_t > 0$,

$$\text{MAPE} = \frac{1}{n}\sum_{t=1}^{n}\left|\frac{y_t - \widehat{y}_t}{y_t}\right| = \frac{1}{n}\sum_{t=1}^{n}\frac{|y_t - \widehat{y}_t|}{y_t},$$

as well as its weighted version, i.e.,

$$\text{wMAPE} = \frac{\sum_{t=1}^{n}\frac{|y_t - \widehat{y}_t|}{y_t}y_t}{\sum_{t=1}^{n} y_t} = \frac{\sum_{t=1}^{n}|y_t - \widehat{y}_t|}{\sum_{t=1}^{n} y_t}.$$

Here, y_t is the true value in time period t and $\widehat{y}_t$ is the prediction.

E. Fireworks-related modeling

All models from the previous sections are based on data without both major incidents and all incidents on New Year's Eve. Since New Year's Eve, in terms of amount of small incidents, is far from normal, we can not just make a forecast for those days with the current models. Using an ensemble method, a forecast for all occurrences of New Year's Eve in our dataset was made. With the real amount of trucks used subtracted from this, we assume that this result approximates the number of fireworks-related incidents.

Table IV shows the correlation between some weather variables and fireworks-related incidents. These incidents occur more often on a cold New Year's Eve with little wind and rain.

TABLE IV. CORRELATION BETWEEN WEATHER VARIABLES AND FIREWORKS-RELATED INCIDENTS.

Variable	Correlation	p-value
Windspeed (FG)	-0.679	0.003
Temperature (TG)	-0.667	0.003
Rainfall (DR)	-0.575	0.016
Visibility (VVN)	-0.407	0.104

Besides the fact that our dataset only holds 17 New Year's Eves, we also found out that due to policy changes the time window for setting off fireworks has been changed. Therefore, only two New Year's Eves in our dataset are completely representative for future ones. Based on these limitations we implement a simple linear model with just an intercept and four weather variables.

The results of estimating the model on all New Year's Eves are given in Table V, including p-values of two-sided t-tests to test the null hypothesis that the true parameter equals zero.

If we estimate the model just on the first twelve New Year's Eves and leave the last five for testing, we get a MAPE of 0.349. Due to too little New Year's Eves in our dataset we are unable to make a very accurate forecast in this particular occasion. This may very well also be the reason for the lack of significant predictive power by the weather variables. Since New Year's Eve from many different perspectives is not a regular day, certainly agreed upon by the fire department, we chose to use this simple estimation thus not to spend more time trying to improve upon this model.

TABLE V. PARAMETER ESTIMATES OF A LINEAR MODEL FOR FIREWORKS-RELATED INCIDENTS.

Variable	Estimate	p-value
Intercept	219.158	0.000
Windspeed (FG)	-0.607	0.352
Temperature (TG)	-0.565	0.198
Rainfall (DR)	0.041	0.941
Visibility (VVN)	-0.405	0.519

IV. RESULTS

In this section, we will compare the performance of the different models and evaluate the insights derived from them. The results on the MAPE and wMAPE values are given in Table VI. These performance measures are based on the total daily number of trucks used for small incidents (over all fire stations and types). This enables us to compare all models through one value. It is also interesting to see how significant a parameter is on a 1 to 5 scale, as in Table VII for LM, Table VIII for GLM, and Table IX for RF. Here, we assign 1 when the p-value < 0.001 (very significant) until 5 when the p-value ≥ 0.1 (not significant).

TABLE VI. PERFORMANCE MEASURES OF THE MODELS.

Model	MAPE	wMAPE
LM	0.1886	0.1924
GLM	0.1865	0.1880
RF	0.2006	0.2019

A. Linear regression model

For the linear model, comparing Table VII to Table III, we observe that when a weather variable has significant predictive power for some type, then their mutual correlation is relatively high as well. This is a nice result, but unfortunately, the reverse is not true. For instance, type 3 is highly correlated with one of the temperature variables, but this variable does not have predictive power for this type, which is surprising.

If we look at Table VII in more detail, it stands out that several types have no weather variables with significant

TABLE VII. SIGNIFICANCE OF ESTIMATED PARAMETERS FOR LM.

Variable	Type cluster (see Table III)) 1	2	3	4	5	6	7	8	9	Avg
Intercept	1	1	1	1	1	4	1	1	1	1.33
Trend	1	5	1	4	3	5	5	5	5	3.78
Wind speed	1	5	5	3	5	5	5	5	1	3.89
Temperature	3	4	5	1	2	5	5	5	5	3.89
Rainfall	1	3	5	5	5	5	5	4	1	3.78
Visibility	5	4	5	4	5	5	5	5	5	4.78

Scaling: 1: $p < 0.001$, *2:* $p < 0.01$, *3:* $p < 0.05$, *4:* $p < 0.1$, *5:* $p < 1$

predictive power. Opposed to type 3, this is not surprising for types 6 and 7, since their correlations to the weather variables are relatively low as well. On the other hand, types 1 and 9 are well predicted by the amount of wind and rainfall, which is intuitively explainable as well.

Since the wMAPE is higher, we conclude that the LM is not very good at predicting relatively busy days (compared to predicting average days). However, the fire brigade is, of course, more interested in when they have busy days. They are prepared for average days anyway.

B. Generalized Linear Model

Recall that the GLM model is an expanded version of the linear model, so it could be at least as good. The question is how much value it adds to the linear model. Comparing the significance of the variables in Table VIII to that of LM in Table VII, we observe that, in general, the single weather variables have lost some importance in favor of cross-term variables they partition in. Type 1 is an excellent example of this. Here, the temperature had some predictive power in the LM, but now it turns out that it is mainly the *combination* with the amount of rainfall that matters. In addition, also wind speed and rainfall turn out to be less predictive on their own than the LM indicated. It is their cross-term effect that is important. Looking at the average column on the right, we also see that the intercept has lost some importance. Apparently, a bigger part can be modeled by the weather after adding some cross-term variables. Of all weather variables, it is even the case that two cross-term variables have the most predictive power.

Noting the influence of the cross-term variables, we expect that the performance of the GLM is better than that of the LM. If we compute the results for the totals per day, we still see that the wMAPE is somewhat higher than the MAPE, but compared to their equivalents of the LM, they are slightly better (about 2%).

TABLE VIII. SIGNIFICANCE OF ESTIMATED PARAMETERS FOR GLM.

Variable	Type cluster (see Table III)) 1	2	3	4	5	6	7	8	9	Avg
Intercept	1	2	1	1	1	5	1	2	3	1.89
Trend	1	5	1	4	3	5	5	5	5	3.78
Wind speed	3	5	5	5	5	5	5	5	1	4.33
Temperature	5	5	5	3	2	5	5	5	4	4.33
Rainfall	5	3	5	5	5	5	5	5	1	4.33
Visibility	5	5	4	5	5	5	5	5	5	4.89
Wind*Temp.	5	5	5	5	5	5	5	5	5	5.00
Wind*Rain	3	3	5	5	5	5	5	5	1	4.11
Wind*Visib.	5	3	5	4	5	5	5	5	5	4.67
Temp.*Rain	2	3	5	5	5	5	5	5	1	4.00
Temp.*Visib.	5	5	5	5	5	5	5	5	5	5.00
Rain*Visib.	5	5	5	5	5	5	5	3	5	4.78

Scaling: 1: $p < 0.001$, *2:* $p < 0.01$, *3:* $p < 0.05$, *4:* $p < 0.1$, *5:* $p < 1$

TABLE IX. IMPORTANCE W.R.T. TOTAL DECREASE IN RSS.

Variable	Type cluster (see Table III)) 1	2	3	4	5	6	7	8	9	Avg
Wind speed	4	2	4	1	3	2	2	4	1	2.56
Temperature	1	1	1	3	2	1	1	1	3	1.56
Rainfall	3	4	3	2	1	4	4	3	2	2.89
Visibility	2	3	2	4	4	3	3	2	4	3.00

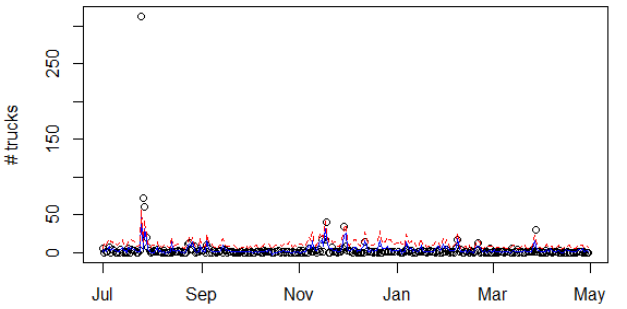
C. Random Forests

Different from the previous models, the RF algorithm does not estimate a parameter for each variable. We, therefore, have to find another measure for the importance of each variable. We will consider the 'RSS-ranking' for this purpose.

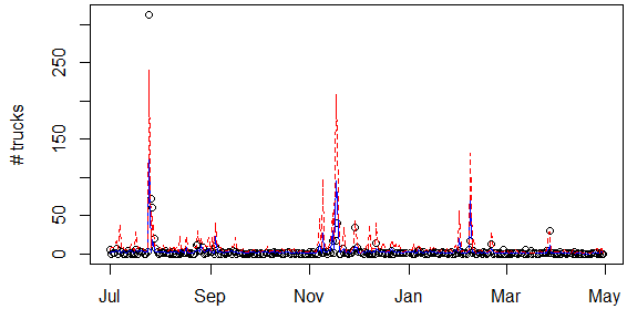
In the RF algorithm, in each decision node, the algorithm splits the remaining sample based on a decision rule on the variable that reduces the standard deviation most. In other words, it tries to improve the fit of the model to the training data as much as possible, i.e., the biggest decrease in *residual sum of squares* (RSS) between the fitted model and the observation data in the training set. Hence, we can measure the importance of a variable based on the total decrease in RSS from splitting on this variable. Table IX shows the results of the RSS ranking. As in the previous models, visibility is often the least important variable. However, the biggest difference is that in this case, the temperature is remarkably important.

When we compare the results of RF to the previous models, we see that, in general, RF gives the worst results. However, the effort for running this model is perhaps not in vain. When diving deeper into the results, we discover that the RF has the best wMAPE for type 9, which may be an indication that this algorithm is better in predicting busy days. This is confirmed by the plot of the predictions for type 9 of both GLM and RF

in Figure 6. Obviously, the RF algorithm recognizes much better than GLM when the weather conditions are risky and likely to cause many incidents to happen.



(a) Generalized Linear Model



(b) Random Forest

Figure 6. Forecasts (in blue) of the number of trucks used for small incidents of type 9, including the upper bound of its 95%-prediction interval (in red).

D. Ensemble model

From the previous discussion, we conclude that GLM gives the best results when we look at the totals per day, but it is worse in predicting busy days than RF. Motivated by this, we propose to use *ensemble averaging* (EA), defined by

$$\text{EA} = \gamma \cdot \text{RF} + (1 - \gamma) \cdot \text{GLM},$$

for some constant $\gamma \in [0, 1]$.

We have to determine the optimal value of γ to use in order to get the best results. Since GLM initially gives the best results, and we only need RF to be able to predict the busy days a bit better, we may expect that we have to put more weight on GLM, i.e., that $\gamma < 0.5$. When we vary γ from 0 to 1, both the MAPE = 0.1853 and the wMAPE = 0.1860 take their minimum in $\gamma^* = 0.2$ (which is better than GLM individually; when compared with $\gamma = 0$).

TABLE X. CAPACITY NEEDED PER DAY AND FIRE STATION WITH CERTAINTY THIS CAPACITY SUFFICES THAT DAY.

Fire station	Avg cap. needed 90%	95%	99%	% of days 2 needed 90%	95%	99%	Available cap. 1?
Aalsmeer	0.14	0.17	0.27	0.0%	0.0%	0.0%	No
Amstelveen	0.44	0.53	0.80	0.0%	0.3%	3.3%	No
Anton	0.40	0.48	0.73	0.0%	0.0%	0.3%	No
Diemen	0.12	0.15	0.25	0.0%	0.0%	0.0%	No
Dirk	0.34	0.41	0.64	0.0%	0.0%	0.7%	No
Driemond	0.04	0.05	0.10	0.0%	0.0%	0.0%	Yes
Duivendrecht	0.17	0.20	0.30	0.0%	0.0%	0.0%	No
Hendrik	0.59	0.71	1.07	0.7%	1.7%	67.7%	No
IJsbrand	0.19	0.24	0.38	0.0%	0.0%	0.0%	Yes
Landelijk Noord	0.04	0.06	0.11	0.0%	0.0%	0.0%	Yes
Nico	0.35	0.42	0.64	0.0%	0.0%	0.3%	No
Osdorp	0.42	0.51	0.77	0.0%	0.0%	1.0%	No
Ouderkerk a/d Amstel	0.06	0.08	0.13	0.0%	0.0%	0.0%	Yes
Pieter	0.41	0.50	0.75	0.0%	0.0%	1.7%	Yes
Teunis	0.28	0.34	0.53	0.0%	0.0%	0.0%	No
Uithoorn	0.12	0.15	0.25	0.0%	0.0%	0.0%	No
Victor	0.28	0.34	0.51	0.0%	0.0%	0.0%	No
Willem	0.30	0.36	0.55	0.0%	0.0%	0.0%	No
Zebra	0.23	0.28	0.44	0.0%	0.0%	0.0%	Yes

E. Practical implication

After the forecasts are complete, we extract from them the capacity we expect each fire station to need each day. For this, we want to have some certainty that the capacity is satisfying for that day. Different from a confidence interval, which only measures the uncertainty of the forecast, a prediction interval includes, in addition, the variability of the number of incidents in real life. We can, therefore, use the upper bound of the prediction interval to ensure that the predicted capacity will be satisfactory with, for instance, 95% certainty.

The $100(1 - \alpha)\%$-prediction interval for the GLM model $y = X^\intercal \beta + \epsilon$ for a future observation y_0 can be computed as [14]

$$\widehat{y}_0 \pm t_{n-k}^{(1-\alpha/2)} \widehat{\sigma} \sqrt{x_0^\intercal (X^\intercal X)^{-1} x_0 + 1},$$

where $\widehat{y}_0$ is the predicted value for y_0, $t_{n-k}^{(1-\alpha/2)}$ is the $(1-\alpha/2)$-quantile of the t-distribution with $n-k$ degrees of freedom, n is the number of samples in the training set, and k is the number of variables in the model.

For the RF algorithm, we have N decision trees, which all yield one prediction for each future observation. The variability of these N individual predictions captures the uncertainty of the final prediction (the average of the individuals). In order to capture the variability of the observations, we need again our assumption on the residuals. In this case, we will use this by adding to each of the N individual predictions a random value, drawn from the empirical distribution of the residuals in the training set. Then, the resulting N values include all the variation we need. Their $(\alpha/2)$- and $(1-\alpha/2)$-quantiles together directly form the desired prediction interval.

If we combine all these results, we get Table X that gives the needed capacity for each fire station. From this,

we can conclude that, on an average day, (almost) all fire stations only need a capacity of one truck. Only if we want to be 99% sure that the capacity suffices, we need a capacity of two trucks at station 'Hendrik' on an average day. Then 'Amstelveen' also needs a capacity of two on some days. Moreover, 'Pieter' does not have the required capacity in 1.7% of the days (see in red).

V. Conclusions and discussion

In this paper, we developed a model to create a forecast on the number of incidents that each fire station in Amsterdam-Amstelland has to handle. Here, special interest went to the influence of several weather conditions and to the issue of dealing with the low number of incidents.

The answer is split into two parts. The forecasts created for the small incidents can be done reasonably well by EA. Major incidents can be modeled by an inhomogeneous Poisson process. Concerning the weather, (the combination of) rain and wind on average had the most influence in the linear models and temperature appeared to contain mostly non-linear relations with the number of incidents. As expected beforehand, the visibility has the least predictive power among those four weather variables.

The current implementation computes a different model per type-cluster and subsequently divides the total prediction over the fire stations. One enhancement for further research that certainly seems logical is to make an estimate per region instead of per station. Incidents happen at a certain place in a certain region, not where a truck of a certain fire station happened to be in the vicinity of. An added benefit is that we can more accurately use the characteristics of separate regions. For example, a region with many big and/or old trees may be more at risk during days with severe weather conditions. To expand on that even further, this risk can subsequently be adjusted for seasons (e.g., spring vs. winter) and/or regions (e.g., different tree types) with more or less leaves on the trees, making them respectively more or less prone to falling over due to wind gusts.

Some of these characteristics can be captured by first dividing the prediction per type-cluster over all regions according to a certain weight. As a proof of concept we calculated the share of each region in the number of trucks used for small incidents per type-cluster, of which the results can be found in Table XII. Using the LM of Section III-A we calculated the results per region, as shown in Table XI. As expected, we find that - in terms of wMAPE - when fewer incidents happen, it gets harder to make a good forecast. When we calculate the totals per day we observe that MAPE(LM2) = 0.1887 and wMAPE(LM2) = 0.1919, which is in line with our previous finding for the results of the LM as shown in Table VI. Note that in both cases we used the same models for the type-clusters, which not surprisingly led to similar results. Future research should be able to generate a model which can be applied to each separate region, while still taking all different incident types into account, and come up with a way to divide the prediction over all fire stations.

TABLE XI. QUALITY OF LM FORECASTS PER REGION IN TERMS OF WMAPE AND AVERAGE NUMBER OF TRUCKS USED FOR SMALL INCIDENTS PER DAY.

Region	wMAPE	Avg # trucks
External	1.766	0.2
Center	0.321	16.2
Harbor area	0.371	10.2
North	0.414	7.3
East	0.634	4.2
South	0.576	5.4
Southeast	0.370	9.0
Average	0.636	7.5

References

[1] G. A. G. Legemaate, S. Bhulai, and R. D. van der Mei, "Applied urban fire department incident forecasting," in Proceedings of IARIA Data Analytics, Sep. 2019.

[2] J. B. Goldberg, "Operations Research Models for the Deployment of Emergency Services Vehicles; EMS Management Journal," EMS Management Journal, vol. 1, no. 1, 2004, pp. 20–39.

[3] J. M. Chaiken and J. E. Rolph, "Predicting the demand for fire service," RAND Corporation, P-4625, 1971.

[4] P. L. van den Berg, G. A. G. Legemaate, and R. D. van der Mei, "Increasing the responsiveness of firefighter services by relocating base stations in Amsterdam," Interfaces, vol. 47, no. 4, 2017, pp. 352–361.

[5] D. S. Matteson, M. W. McLean, D. B. Woodard, and S. G. Henderson, "Forecasting emergency medical service call arrival rates," The Annals of Applied Statistics, vol. 5, no. 2B, 2011, pp. 1379–1406.

[6] H. Setzler, C. Saydam, and S. Park, "EMS call volume predictions: A comparative study," Computers & Operations Research, vol. 36, no. 6, Jun. 2009, pp. 1843–1851.

[7] M. P. Larsen, M. S. Eisenberg, R. O. Cummins, and A. P. Hallstrom, "Predicting survival from out-of-hospital cardiac arrest: a graphic model." Annals of emergency medicine, vol. 22, no. 11, Nov. 1993, pp. 1652–8.

[8] M. Gendreau, G. Laporte, and F. Semet, "The Maximal Expected Coverage Relocation Problem for Emergency Vehicles," The Journal of the Operational Research Society, vol. 57, 2006, pp. 22–28.

[9] A. Ganteaume, A. Camia, M. Jappiot, J. San-Miguel-Ayanz, M. Long-Fournel, and C. Lampin, "A Review of the Main Driving Factors of Forest Fire Ignition Over Europe," Environmental Management, vol. 51, no. 3, Mar. 2013, pp. 651–662.

TABLE XII. SHARE OF EACH REGION IN THE NUMBER OF TRUCKS USED FOR SMALL INCIDENTS PER TYPE-CLUSTER.

Region\Type-cluster	1	2	3	4	5	6	7	8	9	Avg
External	0.8%	0.2%	0.0%	0.1%	0.1%	0.1%	0.3%	3.2%	0.2%	0.6%
Center	22.2%	28.9%	35.7%	21.1%	32.1%	29.8%	33.3%	26.5%	32.9%	29.2%
Harbor area	26.5%	15.5%	20.8%	14.7%	19.8%	20.6%	17.6%	20.0%	18.4%	19.3%
North	14.0%	22.1%	10.0%	11.8%	13.8%	14.1%	11.1%	12.3%	13.1%	13.6%
East	9.7%	8.8%	7.5%	14.9%	8.1%	7.0%	9.3%	8.4%	7.2%	9.0%
South	10.6%	9.9%	10.3%	19.2%	7.0%	12.9%	11.9%	16.2%	12.7%	12.3%
Southeast	16.1%	14.4%	15.7%	18.2%	19.1%	15.4%	16.4%	13.3%	15.6%	16.0%
Total	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%

[10] N. I. C. C. U.S.A. Predictive Services Program Overview. Last accessed on 11/30/2021. [Online]. Available: https://www.predictiveservices.nifc.gov

[11] B. P. Oswald, N. Brouwer, and E. Willemsen, "Initial Development of Surface Fuel Models for The Netherlands," Forest Research: Open Access, vol. 06, no. 02, 2017.

[12] Brandweer Nederland. Jaarwisseling: incidenten voor het tweede jaar op een rij toegenomen. Last accessed on 11/30/2021. [Online]. Available: https://www.binnenlandsbestuur.nl/bestuur-en-organisatie/nieuws/brandweer-moest-vaker-uitrukken.11915233.lynkx

[13] P. G. van der Velden, C. J. Yzermans, and L. Grievink. New York, NY, US: Cambridge University Press, 2012, ch. Enschede Fireworks disaster, pp. 473–496.

[14] J. J. Faraway, "Practical regression and ANOVA using R," University of Bath, 2002.

Applying Motivational Theories and Personalization in a Mobile Application within the Domain of Physiotherapy-Related Exercises

Marie Sjölinder
RISE
Stockholm, Sweden
marie.sjolinder@ri.se

Vasiliki Mylonopoulou
University of Gothenburg
Gothenburg, Sweden
vasiliki.mylonopoulou@ait.gu.se

Anneli Avatare Nöu
RISE
Stockholm, Sweden
anneli.nou@ri.se

Olli Korhonen
University of Oulu
Oulu, Finland
Olli.Korhonen@oulu.fi

***Abstract*—This paper describes motivational features and personalization in a mobile application for physiotherapy-related exercises. Motivational and personalization theories are discussed in terms of being relevant for the developed application. The motivational features that were applied supported goal setting, possibilities to follow progress, personalization and possibilities to compare one's own progress or performance with other users. During the iterative development of the application, an explorative study was conducted in which the participants were interviewed about the aspects related to motivation and personalization. In the study, the participants emphasized the importance of goal setting together with the physiotherapist and of being able to track progress. With respect to being able to compare performance or progress with other users, the outcome of our work is in line with previous research in which comparisons have been rejected. Based on the outcome of the study and on insights with respect to applying motivational theories, the implications and usefulness of the applied theories are presented and discussed.**

Keywords - movement-related disorders; mobile application for performing exercises; motivational theories; goal setting; social comparison; personalization

I. INTRODUCTION

This paper is an extension of the work described in Sjölinder et al. [1], where the development of a mobile application for physiotherapy-related exercises was described. This extended paper focuses on the usefulness and applicability of different motivational theories.

Movement-related disorders is one of the most common occupational hazards in the European Union, and workers in all sectors and occupations are affected [2]. This is an increasing problem and one of the key causes of long-term sickness leave. Early detection and early intervention could reduce the number of serious movement-related problems. By gathering and analyzing movement data from large groups of people over a long period of time, different movement-related patterns can be categorized. Based on this categorization, a person's movement pattern can be placed into one cluster and early signs of problems and movement-related disorders can be detected before they have started to cause problems or pain. Based on this knowledge, personalized support and exercises can be suggested using a smartphone application. However, the challenge is to motivate the users to perform the suggested exercises based on personalized recommendations from the physiotherapist, and to comply with training programs aimed at solving possible future problems.

In this study, motivational features and personalization were applied in a mobile application for physiotherapy-related exercises. The features were related to goal setting, providing support in tracking progress, personalization and possibilities to compare one's own performance with others. Conducting interviews and gathering feedback from users was a part of a larger process in which the application was developed in an iterative way with different user groups. The aim of the interviews was to gain a deeper understanding of how to apply motivational features and personalization when developing applications based on large amounts of aggregated movement-related data. Based on the outcome of the study and on previous work, the implications and usefulness of the applied theories are discussed. In the following text, Section 2 describes the project and the concept that the developed application was part of. Section 3 to Section 5 present previous research and the background to this work. Section 3 gives an overview of motivational theories, and Section 4 gives an overview of personalization theories and approaches. In Section 5, the theories and the central concepts are discussed in terms of possibilities to be applied in the context of the developed application. Section 6 describes the explorative study that was conducted, and it

presents the outcome of the study, which was a part of the iterative development. Based on the study, Section 7 discusses implications and usefulness of the applied theories. Finally, Section 8 discusses the work conducted and suggests possible future work.

II. An Application for Supporting Physiotherapy-Related Exercises

Physiotherapy is a profession in healthcare that aims at improving the functional ability and health of the healthcare user [3]. The core of physiotherapy is to involve the healthcare user in such a way that the user can participate in the physiotherapy process and the decisions that are made regarding their own health [4]. Although physiotherapists are positive towards technological applications, the adoption of technological applications in physiotherapy has remained low [5]. One of the major challenges in physiotherapy is that the healthcare users are not provided with information that allows them to actively participate in the care process. Because of this, there is a continuous need for technological applications that could provide both the physiotherapist and the healthcare user with easily interpretable personalized data [6].

As there is a niche in the market for such technological interventions, Qinematic, a small Swedish startup, developed a software service that records and analyzes body movements using 3D digital video. The users stand in front of a Kinect sensor and follow instructions about which movements to conduct. Based on these sessions, 3D data is gathered and stored. As an extension to this service, a research project with two aims was formulated. The first aim was to develop machine-learning algorithms to analyze gathered movement data, and the second aim was to develop user applications to provide information about dysfunctional movement patterns, facilitate contact with healthcare providers, make it possible for physiotherapists to suggest exercises, and allow the users to set goals and track their progress (Figure 1).

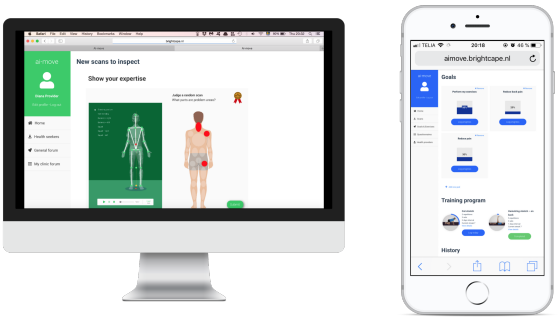


Figure 1. Application for health providers to the left, and for their clients to the right

Via the application, the healthcare provider had the possibility to gather further information by asking the healthcare users health-related questions, with the aim of providing better and more personalized care. The entire system consisted of several parts, including machine learning and categorization of dysfunctional movement patterns. The work presented in this paper focuses on the development of motivational features in the application targeted towards healthcare users with possible dysfunctional movement patterns. However, the larger concept surrounding the application, with a machine-learning module and a healthcare provider application, placed other demands related to how to apply motivational features than what are faced when developing applications that only support users to be more physically active or to perform exercises.

III. Motivation and Behavior Change Theories

This section describes some of the most important motivational and behavior change theories and their relation to the design of technological applications. These theories shaped the starting point of our design discussions and some of them played a central role in the final design of the application presented in this paper.

A. Models and theories focusing on the individual

The **Health Belief Model (HBM)** is one of the most used behavior change models [7]. The aim of the model is to provide explanations of behaviors related to health prevention [8]. The focus of the model is on the individual's beliefs and attitudes. It suggests that the individual's perception determines success in terms of behavior change [7]. For the health behavior to trigger, there must be an external stimulus or a cue prompting the appropriate action [7]. Basic individual variables are (1) perceived susceptibility, i.e., how possible it is to have the condition; (2) perceived seriousness, i.e., how severe the impact of the condition will be on the person's life; and (3) perceived benefits of and barriers to taking action (an example of a benefit is the belief that taking a preventive action will have a positive health outcome, and a barrier could be that the action the individual has to take is expensive). A fourth individual factor was added in 1982 [9]: self-efficacy from social cognitive theory (a social behavior change theory described below). Perceived self-efficacy is one's own belief in their ability to perform a task [10]. The figure below (Figure 2) visualizes the main components of the HBM.

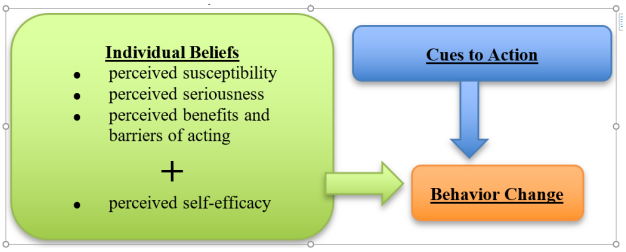


Figure 2. Main components of the Health Belief Model

In the field of technology and health, the HBM is used in combination with other behavior change and motivation theories to design technology for physical activity/wellbeing [11], and for more (medical) condition-focused applications [12][13]. The field of technology and health – apart from using the HBM – has extended it and combined it with other theories to better suit the technological context [14][15].

Self-determination theory (SDT) is a motivational theory focused on the types of motivation a person has towards different behaviors. It consists of concepts such as intrinsic motivation (motivation that comes from within) and extrinsic motivation (external driver). Further, the theory focuses on individual differences with respect to intrinsic and external motivation, basic needs, goals, and relatedness to others [16]. SDT has two important characteristics. First, it is more focused on the type of motivation than the amount of motivation. Second, it underlines the importance of three components: autonomy, competence, and relatedness [17]. Autonomy is related to the feeling of being in control of our own self and our actions. Competence is related to the ability to perform a task. Relatedness is related to our will to interact with others. SDT is a complex theory, and its detailed description is beyond the scope of this work.

In the field of technology and health/wellbeing, SDT has been used in the evaluation and design of technologies. Some examples of where it has been applied are to serve as a basis for the creation of heuristics for healthcare wearables [19], to get integrated in the evaluation process of wearable technology for physical activity [20], and to enrich commonly used design tools such as personas [21].

Stage models are different models that focus on people's readiness to change and categorize them based on that. These models have a clear definition of each stage, and clearly defined factors one must fulfill to move to the next stage [22]. One must pass through all the stages before reaching the final behavior. However, relapsing to previous stages is expected and this is not necessarily sequential [22]. Moreover, one can stay in a stage eternally [22]. The most well-known stage model is the Transtheoretical Model (TTM) first used for smoking cessation and addictions, and later expanded to physical activity and eating habits [23]. The TTM has six time-based stages, each of these including ten factors [23]. Table 1 presents the different stages of the TTM.

TABLE I. DESCRIPTION OF THE TTM STAGES

TTM Stages	Description
Pre-contemplation	The person is not intended to act soon (usually in the next 6 months)
Contemplation	The person is intended to act soon (usually in the next 6 months)
Preparation	The person intends to act in the next 30 days and may have taken actions in the past
Action	The person changed behavior but kept it for less than 6 months
Maintenance	The person changed behavior but kept it for less than a year
Termination	The person changed behavior and kept it for more than a year

The Precaution Adoption Process Model (PAPM) is another stage model [24][25] that takes a different approach than the TTM. The PAPM has seven stages of change that are based on the psychological state of the person rather than the time duration the person practices the new behavior; see Table 2 [26]. It has clearly defined and stage-specific factors for each transition between the stages [27]. The PAPM was created to meet the need for a qualitative approach to adopting new complex behaviors that cannot be fully described by cost-benefit individual models such as the HBM [26].

TABLE II. DESCRIPTION OF THE PAPM STAGES

PAPM Stages	Description
Stage 1	Unaware of an issue
Stage 2	Unengaged by the issue
Stage 3	Undecided about acting
Stage 4	Decided not to act
Stage 5	Decided to act
Stage 6	Acting
Stage 7	Maintenance

In the field of technology and health/wellbeing, the stage models have been used mainly to evaluate the effect of the technology on users' behavior change. An illustrative example of such usage is the well-cited study involving the Fish'n'Steps interactive computer game, in which users could track their own progress as visualized by the growth of a virtual character [28]. The stage models are used today in similar ways [29].

B. Motivational and behavior change theories focusing on social aspects

Social cognitive theory (SCT) focuses on the interplay between individual factors, behaviors, and the environment [30]. It is based on Bandura's social learning theory [31], which supports learning within a social context, i.e., that people bring with them knowledge and experience and that learning happens by imitating others. In SCT, concepts from cognitive psychology and social learning are merged [22] into five categories: psychological determinants of behavior (which include self-efficacy, goals, and outcome expectations) [30], observational learning, environmental determinants of behavior, self-regulation, and moral disengagement. Goals that people set for themselves can be both short-term and long-term goals [30]. Self-efficacy – the belief in one's own ability to conduct a task or take action – is a vital component of the theory [10].

In the field of technology, SCT has been used to support the design of applications related to health and wellbeing, such as physical activity [11][32]. Moreover, it has been used to develop applications that target behaviors by involving multiple actors, such as patients and caregivers. An example of such a case is the development of an application supporting children with asthma and their parents [33]. Another example is a breastfeeding application that was aimed at motivating fathers to support their partners in continuing to breastfeed [34].

Social comparison theory (SCT) suggests that people, in the lack of standard measurements, compare themselves to others for self-evaluation [35], self-enhancement [36], self-projection [37], and coping [38][29]. People often confuse comparison with competition due to the close relationship between these words; however, this relationship is rarely studied [40]. Regardless of the positive results from psychological studies on social comparison [38][39], people often refuse to engage in comparisons due to social norms [41], different perceptions of the term "comparison" [42], or a confusion of "comparison" with "competition" [1]. Social comparison has shown potential in the field of psychology, and it is often used in the field of technology.

In the field of technology, social comparison is often applied as a gamification feature for behavior change [1][43]. Its design is challenging as it needs special care if the designer wants to avoid promoting competition while still promoting one or more of the other aspects, such as social learning [44]. In general, the design field has designed comparison features without specifically referring to the theory, such as in the well-cited Fish'n'Steps study we referred to in the TTM model [23], involving a computer game in which the users can compare the states of their avatars and draw conclusions about each other's physical activities.

C. Goal setting

Goal-setting theory (GST) has proven to have a positive effect on behavior change [45] and has also been used in healthcare in relation to physical activity [46]. Moreover, goal setting is part of other behavior change theories such as SDT and SCT. GST focuses on the goals, as the name implies, which can be divided into sub-goals and into different levels of difficulty [47]. Locke and Latham [48] identified three types of goals assignment: (1) self-set, (2) assigned, and (3) participative-set. Self-set goals are those the individual sets and usually have personal significance. An assigned goal is set for the individual by someone else and a participatory-set goal is a goal that the individual has contributed to define. The GST has evolved since it was first defined, and the goals can now be defined as learning and performance goals [46], with the first being more relevant to people who are new in the particular behavior change as they learn the new behavior.

In the field of technology, GST has been the second most popular theory after the TTM stage model [49]. Goals have also been used in combination with gamification [50]. One example of goal-setting theory in persuasive technology is CALFIT, which shows the results of daily progress between users who have personalized goals and those who do not [51]. Another example is the design of applications that target physical activity for people with chronic obstructive pulmonary disease by implementing goals in different ways [52]. GST can be combined with other theories and applied in design, such as in the MS application where the user, a multiple sclerosis patient, learns to estimate the energy they will consume by doing various activities [53].

IV. PERSONALIZATION THEORIES AND APPROACHES

This section describes some of the most important personalization theories and approaches that can be considered in the design of technological applications. Personalization theories and approaches have their roots in service marketing [54], but information technology has increasingly become the main enabler for personalization [55]. Personalization approaches and theories introduced in this section provided a starting point for our design as they helped us to consider the design at the different levels.

Personalization at the level of the technological application. The first personalization approach focuses on the design of personalization at the level of the technological application itself. At this level, personalization can be defined as a process that changes the functionality, interface, information content, or distinctiveness of a technological application to increase its personal relevance to an individual [56]. Personalization theories and approaches at this level have often focused on classifying and describing personalization in different dimensions.

One of the most well-known and comprehensive classifications for personalization was provided by Fan and Poole [57], who classify personalization into three main dimensions: The dimensions are: 1) "What to personalize?," which refers to the aspect of the technological application that is adjusted to provide personalization, for instance, user interface (UI); 2) "To whom to personalize?", which refers to the target of personalization – whether the personalization is targeted at a single individual or a group of users; and 3) "Who personalizes?", which refers to the party that is providing personalization, meaning whether personalization is done by the system/service provider or by the user [57]. These dimensions and the aspects in these dimensions are intended to support the design of personalization at the level of the technological applications. When it comes to personalization, developers have often lacked the theoretical frameworks for personalization [58]. These dimensions have also been considered in the design of technological applications in the domain of healthcare [59].

Personalization through technological application. The second personalization approach considers personalization more broadly than in terms of just the technological application itself, focusing instead on personalization that can be mediated through the technological application. Personalization through technological application is referred to as technology-mediated personalization (TMP) [60]. Shen and Ball [60] classify personalization through technological application into three main categories: 1) "Interaction personalization", which refers to the use of technological applications to

address the user by name, for example, through personalized emails or greetings; 2) "Transaction outcome personalization", which refers to the use of technological applications to allow the user to personalize certain aspects of the product of service, for instance, in the form of a webpage layout in which the user is allowed to make adjustments based on his/her preferences; and 3) "Continuity personalization", which refers to continuous personalization based on the learning and knowledge of the user and where the gained expertise is used to provide even more personalized products and services to this individual user [60]. The TMP approach connects to personalization literature in the field of service marketing, where personalization takes place in the service interaction between the customer and the service provider [61][62], and where the role of the service provider is emphasized in personalization [63].

Role of the technological application(s) in personalization. The third personalization approach considers personalization even more holistically, at the level of the entire service process, where instead of a single technological application, several different technologies can support the service process to make the service pathway more personalized for the individual user. The focus at this level is on the role technological applications can play in the personalization of the entire service pathway. According to Korhonen and Isomursu [64], the role of technological applications in the personalization of the entire service pathway can be classified into three categories: 1) "Coercive personalization" in which personalization is provided automatically by technological applications without the involvement of the human actor (for instance, based on predefined personalization parameters); 2) "Data display personalization" in which personalization is provided automatically by technological applications, but is interpreted manually by the human actor; and 3) "Collaboration-based personalization" in which personalization is supported by technological applications, but the focus is on the interpretation and co-creation of personalized services between the human actors.

V. APPLYING MOTIVATIONAL FEATURES AND PERSONALIZATION IN A MOBILE APPLICATION FOR PHYSIOTHERAPY

This section describes how different motivational and personalization theories were applied as features in the developed application.

A. Applying motivation and behavior change theories

GST influenced the design of the application in relation to creating goals and different types of rewards (see Figures 3 and 4). Goals could be set by the physiotherapist or by the user him/herself. In the design, the goals were closely related to each exercise. Both proximal (short-term) and distal (long-term) goals could be set within the application. Although originating from GST, the distinction between short-term and long-term goals can also be found in the autonomy part of SDT. Short-term goals were set by the physiotherapist and were related to the exercises, and the long-term goals were set by the user. The users self-evaluated their progress on their long-term goals in terms of reduced pain (Figure 3). One social outcome was the evaluation that was conducted by the physiotherapist, which could be seen from the perspective of SDT's relatedness.

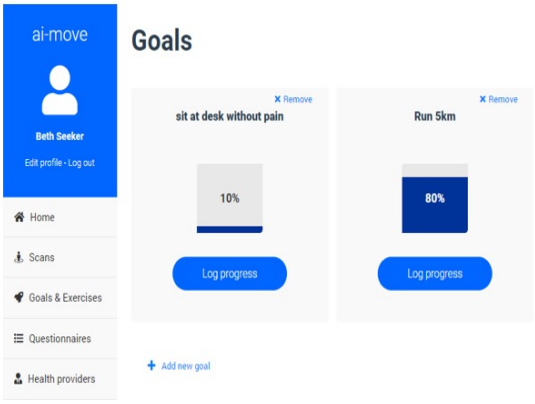


Figure 3. Possibilities to see progress in relation to the set goals

Figure 4. Reward for doing the exercises every day for a week

SDT consists of different types of motivation, and the external motivational aspects were relevant for the design of the application. The concept of autonomy – the freedom of choices a user has – was applied through GST and particularly by implementing self-set goals. The concept of competence – the person's belief in the extent to which they can conduct a task – was applied through personalization of services, in this case through discussion with the physiotherapist in which they together set exercises, repetitions and goals in a way that made the healthcare user feel competent and capable. This was also covered in the concept of self-efficacy – the belief in one's own ability to conduct a task – which is part of SCT.

To cover the relatedness part of SDT and the social aspect of SCT, social comparison theory was applied. Due to the tight timeframe of the project, it was a challenge to apply social comparison theory fully since a number of different users are needed for comparisons. However, potential users

of the application were asked during the interviews about what types of comparisons they would like to be engaged in. The phrasing was altered in order to avoid the word comparison due to its challenging nature (i.e., people often reject comparisons). In the application, comparisons were implemented in relation to the physiotherapist's advice and the healthcare user's adherence to it. Such comparisons between the healthcare users could be related to performing the exercises suggested by the physiotherapist in a consistent manner, or in relation to the completion of the medical questionnaires that had been made available by the physiotherapist.

The HBM was applied in relation to the visualizations of the scans. For example, perceived severity and susceptibility could be influenced through the 3D scan visualizations. The idea behind this was that people are unaware of how their body moves, but by looking at the visualizations they could be able to better understand harmful movement patterns that on a long-term basis could lead to problems and chronic pain.

Stage theories such as PAPM and TTM were found to be difficult to apply. The TTM is time-based and at least one month to one year of interaction with a functional prototype of the application is needed in order to be able to see any changes in the user's behavior. However, we used the PAPM to understand the users' initial intentions and readiness to change.

B. Applying personalization approaches

Personalization approaches can provide support at different levels. Personalization approaches at the technological application level [57] can help in understanding what can be personalized, for instance, in the training program. On the other hand, personalization through technological applications [60] can provide insights to consider, for instance, how the physiotherapist can adjust the training program based on the feedback. Finally, the role of technological applications in personalization [64] can help in considering the role of technology, whether it can provide data-display support for the physiotherapist only, or whether it is intended more as a tool for collaborative decision-making between the user and physiotherapist for personalization. Personalization approaches in the developed application were connected to the use of the technological application in order to understand the user needs, but also to the use of the technological application in order to personalize the physiotherapy services for these needs.

VI. Explorative user study

As a part of the iterative development, a set of user tests were conducted. One of these tests focused on motivational features and personalization. This was an explorative study in which the aim was to gather feedback from possible users about how motivational features could be integrated into the application in a meaningful way.

A. Method

There were seven participants in the test, five men and two women, in an age range between 33 and 52 years. All participants had university educations and held a Master of Science degree or higher. The materials used were a digital mock-up prototype designed in Figma [65] and a scenario description (Figure 5).

> **Scenario**: You have done the scan and discussed your result with your physiotherapist. Imagine that a hip problem has been detected (or another problem that you want to choose). You have received a training program from the physiotherapist to improve the hip problem and to prevent hip pain. In the web app, you can see what exercises to do and how often, as well as the number of repetitions. You can also see the results of the scans.

Figure 5. The scenario presented to the participants

Semi-structured interviews were conducted based on an interview guide. The guide was formulated to collect information about motivational features and personalization, for example, types of features that would motivate the participants to use the system and follow the exercise plan, how they would like to receive feedback on their progress, if they would like to be able to see the progress of other users, or to which extent they wanted the system to be adapted to their preferences and needs or to support the individualized treatment in physiotherapy. Each interview lasted about an hour. The interviews were recorded, and the data collected was transcribed and thematically analyzed.

B. Results

Motivational profile. The participants' general profile, based on the interviews, was that they were extrinsically motivated to follow their physiotherapist's advice based on progress improvement and pain reduction. All the participants were aware and engaged with the matters related to physical posture, physical activity, and pain issues due to bad posture or sedentary life. All participants had at some point been instructed to act to reduce the risk of posture issues. However, they had been acting on this on different levels, e.g., they were exercising but often had to skip it because daily life got in the way. Finally, all described issues with maintenance, e.g., when the pain stopped, they started to neglect their physiotherapist's advice.

Being able to see progress. One of the most motivating factors was being able to see progress. The participants described the possibility of being able to see improvement as the most motivating feature, for example by comparing their past scan data with the results from the latest scan or being able to see progress with respect to goals or in terms of reduced pain. Being able to track the progress was described as one of the most important features, because the lack of progress could be demotivating. This showed that the participants were extrinsically motivated to adhere to the physiotherapist's instructions. However, if the outcome was negative or stable, there was risk of getting the feeling of doing something wrong, which could lead to reduced motivation to continue exercising. The lack of pain (external motivator) could also lead to forgetting to do exercises prescribed by the physiotherapist.

Goal setting and feedback from the physiotherapist. Goal setting and feedback from the physiotherapist was described as vital. The participants thought that frequent

personalized interaction with and feedback from the physiotherapist would increase their motivation to continue to do exercises, answer questionnaires and report pain. In general, goal setting was perceived as positive among the participants. However, they were hesitant towards setting their own goals. They perceived the physiotherapists as experts and were expecting them to set the goals. Even though the users had full confidence in the physiotherapist when it came to planning/rehabilitation, there was a desire to do the planning and set the goals together with the physiotherapist.

Reminders. Adherence to the exercises could be externally motivated by discomfort induced by pain. The participants also pointed out that it is easy to forget to do exercises when the pain is no longer present. The possibility to get reminders was described as important by the participants, regardless of the existence of pain. However, they pointed out that the reminders should be adjustable and optional.

Sharing health-related information. To be able to connect with others, there is a need to share information and be able to see other people's information. In practice, this raised a lot of questions in relation to data handling, security, and reasons for collecting the data. Data handling and sharing were described as important aspects for using this kind of feature.

Sharing progress with other users. Sharing progress with other users was a feature that some participants liked, and others strongly disliked. For some, it might be too personal to share health-related aspects, but for others it is a way of sharing experiences and motivating each other. To apply relatedness, we used social comparison theory. We asked the users to report how it would influence their motivation to see other people's data on their persistence in following the physiotherapist's advice (doing the exercises regularly) and in terms of filling in personalized health-related questionnaires. Most of the participants (6 out of 7) thought that we asked them to compare their health progress, but it was clarified that we were asking only about their persistence in sticking to the training program or filling in the health-related questionnaires. Their reply was generally that they were uninterested in knowing about how persistent other users were in following their training programs or filling in their health-related questionnaires. However, the participants pointed out that gamification features in the application could make it more interesting to relate to other people's data if, for example, the data was used for contributing to a group target or used in competing about being the most persistent user. Table 3 shows some of the comments the participants shared about their persistence in sticking to the training program or filling in the health-related questionnaires.

Personalization as a motivating factor. It was important for the users to understand how the technology used the information they provided in the health-related questionnaires. The participants pointed out the importance of a clear connection between questionnaires and the feedback that was given by the physiotherapist. They stated that if they could not understand this connection, they would hesitate to answer health-related questions because the value of answering the questions would not be clear. Being able to report pain and to get personalized feedback specifically based on pain level were described as important aspects. This was understandable, because one of the primary goals for a healthcare user is to get rid of the pain. The importance of personalized feedback can also be seen by the need to have direct contact with the physiotherapist.

For the participants, personalization was connected primarily to being treated as individuals, rather than to interaction with the technological application. In this case, the technology generated additional data points through user reporting that could be used in personalization. The participants expected that the generated data would not only support the physiotherapist in prescribing the most optimal exercises or treatment for them, but also help the physiotherapist to track their progress. In the study, some of the participants expected that the technology would enable advanced forms of personalized feedback from the physiotherapist in terms of care progress and potential improvement in condition. Other participants expected that the technology would generate data in a way that could trigger a personalized intervention based on input from the user. That is, if the user were to report an increased level of pain, the physiotherapist could use the data and contact the user with a personalized intervention.

TABLE III. COMMENTS FROM THE PARTICIPANTS ABOUT SOCIAL COMPARISON

Comments about comparing exercise persistence
A. "If I could see how much I contributed to the group, in a gamified group goal" B. "So as to get the feeling that you are in this together" C. "If we collected points together, I would be more interested than if competing. If other people were persistent, then I would be more persistent" D. "I would be more motivated by competing against the others in the group and try to beat them" E. "Competition is sometimes good but not here, if you make it more like collaboration" F. "It matters more to me if I am doing it than if other people are doing it"
Comments about comparing questionnaire completion persistence
G. "If I was the only one who didn't fill them in, it would have motivated me to fill them in"
Other insights
H. One participant would have liked to be compared only to a standard value or to a value close to a standard based on a statistical average. I. One participant compared their scan results with those of a colleague to understand how their bodies were crooked. This was perceived by the researchers as a comparison that promoted awareness. However, the participant thought that this was a novelty effect and could not see any value in continuing to compare future data.

VII. IMPLICATIONS AND USEFULNESS OF THE APPLIED THEORIES

A good understanding of the most common behavior change and motivation theories is an advantage when

designing applications within this area. It increases the understanding of users' attitudes to performing exercises, their behavior, and their motivation. For example, we expected users to be reluctant when it came to comparison since theory supports that many people will refuse to engage in comparisons for sociocultural reasons [41][42]. Therefore, alternative ways to apply comparison were considered. The design presented in this work was strongly based on theories of motivation, behavior, and personalization. However, the design process of an artifact is complex since it seeks to solve "wicked problems" in the real world [66]. According to Buchanan, a "wicked problem" is a complex problem that has many dependencies, contradictory or incomplete information and changing conditions. Buchanan furthermore links this to the need to understand the problem in its context in order to be able to develop solutions that are valuable from the user's point of view. Therefore, a combination of theories was applied in the development of the application.

A. Goal-setting theory

Goal-setting theory (GST) is a commonly applied theory for health-related behavior change [67]. In the health-related area, the goal setting should be designed with care as it may harm the healthcare users, e.g., if the goal is too difficult to reach and the healthcare users strive to reach it. Locke and Latham [67] define a goal as "*the object or aim of an action, for example, to attain a specific standard of proficiency, usually within a specified time limit*" [48]. Four principles specified in the theory of goal setting – ability, commitment, feedback and situation resources [68] – can be taken into account when setting performance goals.

Individuals need the *ability* to achieve a specific, challenging performance goal: "People cannot attain goals if they do not know how to do so" [68]. Studies have shown that difficult and specific performance goals can be detrimental to performance when people have not acquired the abilities or skills for a particular task [69]. Performance goals should not be set if the necessary ability is lacking [46]. To meet the ability of the health care user, support in setting the goals could be provided. The need for this was also shown in the conducted interviews, where the participants descried a desire to have goals set either by the physiotherapist or in cooperation with a physiotherapist. Goals should also be realistic and appropriate for each participant's situation. Otherwise they may be too hard or too easy to achieve, and this can cause the user to become demotivated [45]. This was also something that was mentioned during the interviews. Besides being realistic, goals need to be concrete and measurable, and it could also be beneficial to have explanations of the goals. Motivation can be further supported by making it possible for the user to track progress towards the goals, and to see goals that have been reached.

Being *committed* to the goals is an important prerequisite for success. The individual him/herself needs to be involved in the goals. They must be seen as important if people are to be committed to achieving the goals. *"A goal that one is not committed to attain will not affect that person's actions"* [68]. The participants in the study described themselves as more committed to and more confident about goals that had been set together with a physiotherapist.

Feedback should be given regularly throughout the treatment or process to provide support in achieving the goals. If the healthcare users set goals that are difficult to achieve, they must be able to get feedback about their performance in relation to their goals [70]. In the absence of feedback, the healthcare users have no information that could support whether they should change their strategies to achieve the goals or whether they should just continue in the same way as previously [46]. Therefore, it is important to understand and receive feedback regarding goals in order to be able to track progress towards the goals or see if you are moving away from them. It should be noted [46] that it may not only be enough to set performance goals; other strategies such as self-monitoring may also be needed to make it easier to achieve the goal. Self-monitoring is a personality trait that involves the capacity to regulate and monitor behaviors and emotions in different situations. Here, an individual focuses on skills that lead to the achievement of the goals. The individual can complement this with their own checklists and/or take notes about their own development of the performance of the exercises. A user should also be able to understand the connection between the goals and the overall objective. This should be done by the physiotherapist, and it could be beneficial to have explanations of the goals in the app.

According to Latham [68], the *resources* that are necessary must be available (such as equipment needed for the exercises), as the lack of these may otherwise influence the individual's ability to achieve the goals.

Goal-setting theory is consistent with the results obtained from the conducted study (see Section 6) in terms of need for feedback. The participants pointed out that it is significant to gain feedback about progress, and that information about progress is one of the most important features to include in an application like this. This can be achieved, for example, by showing improvements in terms of comparing past and present performance, or in relation to the goals that have been set.

B. Theories that include social aspects

To apply the relatedness aspect of self-determination theory, social comparison theory was applied. Social comparison theory suggests that people evaluate their abilities and that they have a willingness to improve. When people feel insecure about their abilities, they usually compare themselves with others. If an individual has several people to compare themselves with, it is likely that he/she will choose someone with abilities at a similar level (e.g., in terms of fitness) for comparison [35]. The more different one individual is from others, the less he or she tends to compare him/herself with them. Other studies, on the other hand, show that some people compare themselves with people they are different from [71].

Sharing health-related data is related to the relatedness aspect of SDT and our will to interact with others. The participants in this study pointed out that they were reluctant to share health data since they perceived their situation and

their health data as unique. They therefore did not wish to have any comparisons, and they were also uninterested in seeing other people's health-related data. However, they were positive to sharing data in relation to the adherence to their physiotherapist's advice and with respect to questionnaire completion, particularly if the comparison was applied in the shape of gamification [1]. One thing that is important to note for the comparison to be meaningful is that the people compared should perceive each other as similar [35]. One solution to this is to categorize users into different groups based on their condition and/or exercises they have to do. This would enable the healthcare users to see the group they belong to and recognize that there are several other people in the same cluster. Being in contact with people who are struggling with similar issues could provide psychological support for negative feelings such as feelings of deviance or isolation [38].

In the study, some of the participants were positive towards sharing advice to help others or asking for advice from those who had managed to follow the physiotherapist's advice better than they had. Healthcare users who are unable to keep up with new exercise routines could thus benefit from having the possibility to ask for advice from people who have managed to engage in new routines [44]. It is recommended that this be designed in a way so as to not trigger competition if that is unwanted. However, not everyone in the study disliked competition related to adherence to the physiotherapist's advice, and several also thought that gamified competition could support motivation [43].

C. External motivation and rewards

Extrinsic or external motivation is behavior driven by external rewards. Extrinsic motivation is a concept that applies every time an activity is performed in order to achieve a desirable outcome. Extrinsic motivation thus contrasts with intrinsic motivation, which refers to doing an activity simply for the enjoyment of the activity itself, rather than its contributing value [72]. SDT suggests that extrinsic motivation can vary greatly in the extent to which it is autonomous (related to the feeling of having control over our own selves and our actions). This can be exemplified by comparing two students doing their homework [72]. One does the homework only so as to not be "punished with sanctions" by his parents, and the other student does the homework because he personally thinks it is good for his future career. Neither is doing it because they find it interesting to learn. Both cases are extrinsically motivated, but the latter example has a sense of personal feeling and approval while the former example is only about external control. Both represent conscious behavior, but the two examples of extrinsic motivation differ in their relative autonomy.

Feedback that is based on data from other users could also serve as external motivation. Progress may be shown based on reported data from another group of users. For example, it can be shown how successful the proposed exercises have been in terms of rapid progress.

The participants in the study pointed out that they were motivated to follow the physiotherapist's advice based on situation and existence of pain, which could be described as external aspects for motivation. Even if exercises are forgotten, their value was understood. Therefore, the participants expressed the need for reminders. Reminders are related to the HBM and its cues to action. Reminders in combination with motivational messages can support the users in remembering to do the exercises, especially when health improves and pain has vanished. Within an application, motivational messages could be based on the user's actions, for example referring to a situation in which the healthcare user had performed the exercises. However, it is important that reminders are adjustable, that they are based on the users' needs, and that the user can deactivate them if desired since they can be overwhelming.

Another feature that might be motivating for some users is to add gamification. This feature is based on external motivation in terms of different kinds of rewards, such as being rewarded for performance in comparison with other users. Comparison, in this case, can be used for compliance in performing exercises and in answering health-related questionnaires. This can be done regardless of progress and without users sharing sensitive information about their health. Another comparable measurement is the streak (the number of days in a row the user did the exercises). This shows the user's compliance on a daily basis. Compliance is also a usable motivational aspect when there has been no progress since it will still be possible to give rewards [73].

VIII. Conclusions and Future Work

Due to the nature of the application, motivational aspects related to goal setting and social motivational theories were the most relevant aspects to apply in the development of the application. Goal setting and being able to follow progress were important features to include. It was also shown that the goals for the user's exercises needed to be realistic and set together with the physiotherapist. This was explained in terms of that the physiotherapists were experts in the physiotherapy domain and could therefore estimate true progress. However, it was important that the goals were meaningful and motivating for the user as otherwise compliance and performance could be affected [48].

With respect to being able to compare performance or progress with other users, our results were in line with the research conducted in the psychological field regarding the rejection of comparison [41][42]. However, if the comparison was disguised as a gamification element, the participants thought that people would be more willing to compare with others for competing, for feeling a part of a group or for contributing to a team. Due to the rejection of comparison in this study, it was impossible to get detailed user specifications about the design of social comparison features. For example, if they would like to compare specific individuals, compare random users of the application, or compare with statistics created by all the users. More research is needed to understand how we can make the users feel comfortable talking about comparisons they engage in.

The need for personalization was mainly related to receiving personalized feedback from the physiotherapist in a way that takes into consideration the user's condition. Users described that the frequent interaction with the physiotherapist and the individualized exercise plan based on input from the users was an important aspect for sharing health-related data with the system. The users in our study were willing to provide a variety of personal information as long as it was used in a meaningful way that supported their progress. Other studies have also shown the importance of social interaction and of being seen by the physiotherapist. For some users, this social aspect might be the most important motivational feature [73].

Finally, one motivational feature that was not initially discussed with the participants but which came up during the interviews was awareness of body posture and that the visualization of the body could be a motivating feature. This could provide the user with feedback about their existing posture and goals showing what to strive for [73].

To summarize, this study conveys insights about applying motivational theories and provides suggestions for developing motivational features in applications that support performing exercises based on recommendations from a physiotherapist. We have not systematically investigated the use of different motivational theories and are not suggesting which motivational theories can be most successfully applied in this context. In an exploratory manner, and for this particular application, practical combinations of different theories were applied. Future work needs to be conducted, both in terms of applying other motivational theories and in terms of evaluating the applied motivational features.

ACKNOWLEDGMENTS

The authors would like to thank EIT Digital, which funded the project that this work was conducted within. We would also like to thank the participants, who provided us with valuable information during the interviews.

REFERENCES

[1] M. Sjölinder, A. Avatare Nöu, V. Mylonopoulou, and O. Korhonen, "Motivational features in an application for presenting dysfunctional movement patterns and for providing support in conducting exercises," CENTRIC 2019, The Twelfth International Conference on Advances in Human-oriented and Personalized Mechanisms, Technologies, and Services, pp. 37–42, 2019.

[2] 1 EUMSUC (2008-2013) Musculoskeletal Health in Europe Report v5.0. EUMSUC, Executive Agency for Health and Consumers.

[3] K. Areskoug-Josefsson and A. C. Andersson, "The co-constructive processes in physiotherapy," Cogent Medicine, vol. 4, no. 1, 2017, doi: 10.1080/2331205X.2017.1290308.

[4] M. Batalden et al., "Coproduction of healthcare service," BMJ Qual vol. 25, no 7, pp.509–517, 2016, doi: 10.1136/bmjqs-2015-004315.

[5] G. Postolache, R. Oliveira, and O. Postolache, "Designing digital tools for physiotherapy," Interactivity, Game Creation, Design, Learning, and Innovation, pp. 74–88. Springer, Cham. 2016

[6] T. J. Hoogeboom, J. J. Dronkers, E. H. Hulzebos, and N. L. Meeteren, "Merits of exercise therapy before and after major surgery," Curr Opin Anaesthesiol, vol. 27 no. 2, pp. 161–166, 2014, doi: 10.1097/ACO.0000000000000062.

[7] I. M. Rosenstock, "Historical origins of the health belief model," Health Education Monographs, vol. 2 no. 4, pp. 328–335, 1974.

[8] N. K. Janz and M. H. Becker, "The Health Belief Model: A Decade Later," Health Education Quarterly, vol. 11, no. 1, pp. 1–47, 1984.

[9] V. J. Strecher and I. M. Rosenstock, "The health belief model," Health behavior and Health Education: Theory, Research and Practice (2nd ed.), 1997.

[10] A. Bandura, "Perceived self-efficacy in the exercise of personal agency," Journal of Applied Sport Psychology, vol. no. 2, pp. 128–163, 1990.

[11] S. U. Ayubi and B. Parmanto, "PersonA: Persuasive social network for physical activity," Annual International Conference of the IEEE Engineering in Medicine and Biology Society, pp. 2153–2157, 2012.

[12] G. Giunti, V. Mylonopoulou, and O. R. Romero, "More stamina, a gamified mHealth solution for persons with multiple sclerosis: research through design," JMIR mHealth and uHealth, vol. 6, no. 3, 2018.

[13] J. A. Naslund et al., "Health behavior models for informing digital technology interventions for individuals with mental illness," Psychiatric Rehabilitation Journal, vol. 40, no. 3, pp. 325–335, 2017.

[14] R. Wahyuni, "Explaining acceptance of e-health services: An extension of TAM and health belief model approach," 5th International Conference on Cyber and IT Service Management, IEEE press, pp. 1–7, 2017.

[15] A. S. Ahadzadeh, S. P. Sharif, F. S. Ong, and K. W. Khong, "Integrating health belief model and technology acceptance model: An investigation of health-related internet use," Journal of Medical Internet Research, vol. 17, no. 2, 2015.

[16] R. M. Ryan and E. L. Deci, "Self-determination theory: Basic psychological needs in motivation, development, and wellness," Guilford Publications, 2017.

[17] E. L. Deci and R. M. Ryan, "Self-determination theory: A macrotheory of human motivation, development, and health," Canadian Psychology/Psychologie canadienne, vol. 49, no. 3, pp. 182–185, 2008.

[18] M. Becker, "The health belief model and personal health behavior," Slack, vol. 2, no. 4, pp. 324–508, 1974.

[19] S. Asimakopoulos, G. Asimakopoulos, and F. Spillers, "Motivation and User Engagement in Fitness Tracking: Heuristics for Mobile Healthcare Wearables," Informatics, vol. 4, no. 1, p. 5, 2017.

[20] C. Kerner and V. A. Goodyear, "The motivational impact of wearable healthy lifestyle technologies: a self-determination perspective on Fitbits with adolescents," American Journal of Health Education, vol. 48, no. 5 pp. 287–297, 2017, doi: 10.1080/19325037.2017.1343161

[21] A. Jansen, M. Van Mechelen, and K. Slegers, "Personas and behavioral theories: A case study using self-determination theory to construct overweight personas," Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems, pp. 2127–2136, 2017.

[22] K. Glanz, B. K. Rimer, and K. Viswanath, "Health behavior and health education: Theory, research, and practice," (4th ed.), Jossey-Bass, 2008.

[23] J. O. Prochaska and W. F. Velicer, "The Transtheoretical Change Model of Health Behavior," American Journal of Health Promotion, vol. 12, no. 1, pp. 38–48, 1997.

[24] E. Haas, "Applying the Precaution Adoption Process Model to the Acceptance of Mine Safety and Health Technologies," Occupational Health Science, vol. 2, pp. 43–66, 2018.

[25] J. Chapin, Adolescents and Cyber Bullying: The Precaution Adoption Process Model Education and Information Technologies volume 21, pp. 719–728, 2016.

[26] N. D. Weinstein, "The precaution adoption process," Health psychology, vol. 7, no. 4, p. 355, 1988.

[27] N. D. Weinstein and P. M. Sandman, "The precaution adoption process model and its application," Emerging Theories in Health Promotion Practice and Research, Jossey-Bass, pp. 16–39, 2002.

[28] J. J. Lin, L. Mamykina, S. Lindtner, G. Delajoux, and H. B. Strub, "Fish'n'Steps: Encouraging physical activity with an interactive

computer game," International Conference on Ubiquitous Computing, Springer, pp. 261–278, 2006.

[29] S. R. Paige, J. M. Alber, M. L. Stellefson, and J. L. Krieger, "Missing the mark for patient engagement: mHealth literacy strategies and behavior change processes in smoking cessation apps," Patient Education and Counseling, vol. 101, no. 5, pp. 951–955, 2018.

[30] A. Bandura, "Health promotion from the perspective of social cognitive theory," Psychology and Health, vol. 13, no. 4, pp. 623–649, 1998.

[31] A. Bandura and R. H. Walters, "Social learning theory," Englewood Cliffs, NJ: Prentice-Hall, vol. 1, 1977.

[32] M. Duncan et al., "Effectiveness of a web- and mobile phone-based intervention to promote physical activity and healthy eating in middle-aged males: randomized controlled trial of the ManUp study," Journal of Medical Internet Research, vol. 16, no. 6, e136, 2014.

[33] M. Iio, "Beneficial Features of a mHealth Asthma App for Children and Caregivers: Qualitative Study," JMIR mHealth and uHealth, vol. 8, no. 8, e18506, 2020.

[34] B. K. White et al., "Theory-based design and development of a socially connected, gamified mobile app for men about breastfeeding (Milk Man)," JMIR mHealth and uHealth, vol. 4, no. 2, e81, 2016.

[35] L. Festinger, "A Theory of Social Comparison Processes," Human Relations, vol. 7, no. 2, pp. 117–140, 1954, doi:10.1177/001872675400700202.

[36] R. L. Collins, "For better or worse: The impact of upward social comparison on self-evaluations," Psychological Bulletin, vol. 119, no. 1, pp. 51–69, 1996, doi:10.1037/0033-2909.119.1.51.

[37] J. Suls, R. Martin, and L. Wheeler, "Social comparison: Why, with whom, and with what effect?," Current Directions in Psychological Science, vol. 11, no. 5, pp. 159–163, 2002, doi:10.1111/1467-8721.00191.

[38] S. E. Taylor, J. V. Wood, and R. R. Lichtman, "It Could Be Worse: Selective Evaluation as a Response to Victimization," Journal of Social Issues, vol. 39, no. 2, pp. 19–40, 1983.

[39] J. V. Wood, S. E. Taylor, and R. R. Lichtman, "Social Comparison in Adjustment to Breast Cancer," Journal of Personality and Social Psychology, vol. 49, no. 5, pp. 1169–1183, 1985, doi:10.1037/0022-3514.49.5.1169.

[40] S. M. Garcia, A. Tor, and T. M. Schiff, "The Psychology of Competition: A Social Comparison Perspective," Perspectives on Psychological Science, vol. 8, no. 6, pp. 634–650, 2013, doi:10.1177/1745691613504114.

[41] K. J. Hemphill and D. R. Lehman, "Social Comparison and Their Affective Consequences: The Importance of Comparison Dimension and Individual Difference Variables," Journal of Social and Clinical Psychology, vol. 10, no. 4, pp. 372–395, 1991.

[42] F. X. Gibbons and B. P. Buunk, "Individual Differences in Social Comparison: Development of a Scale of Social Comparison Orientation," Journal of Personality and Social Psychology, vol. 76, no. 1, pp. 129–142, 1999, doi: 10.1037/0022-3514.76.1.129.

[43] S. A. Mummah, A. C. King, C. D. Gardner, and S. Sutton, "Iterative development of Vegethon: a theory-based mobile app intervention to increase vegetable consumption," International Journal of Behavioral Nutrition and Physical Activity, vol. 13, no. 1, p. 90, 2016.

[44] V. Mylonopoulou, "Design for Health Behavior Change Supportive Technology: Healthcare Professionals' Perspective," NordiCHI '18, pp. 82–92, 2018.

[45] S. Consolvo, P. Klasnja, D. W. Mcdonald, and J. A. Landay, "Goal-Setting Considerations for Persuasive Technologies that Encourage Physical Activity," Persuasive '09, Article 8, pp. 1-8, 2009.

[46] C. Swann et al., "Updating goal-setting theory in physical activity promotion: A critical conceptual review," Health Psychology Review, vol. 15, no. 1, pp. 34–50, 2020, doi:10.1080/17437199.2019.1706616.

[47] V. J. Strecher et al., "Goal setting as a strategy for health behavior change," Health Education Quarterly, vol. 22, no. 2, pp. 190–200, 1995.

[48] E. A. Locke and G. P. Latham, "Building a Practically Useful Theory of Goal Setting and Task Motivation: A 35-Year Odyssey," American Psychologist 57, pp. 705–717, 2002.

[49] R. Orji and K. Moffatt, "Persuasive technology for health and wellness: State-of-the-art and emerging trends," Health Informatics Journal, vol. 24, no. 1, pp. 66–91, 2018.

[50] D. L. Kappen and R. Orji, "Gamified and persuasive systems as behavior change agents for health and wellness," XRDS: Crossroads, The ACM Magazine for Students, vol. 24, no. 1, pp. 52–55, 2017.

[51] M. Zhou et al., "Evaluating machine learning-based automated personalized daily step goals delivered through a mobile phone app: Randomized controlled trial," JMIR mHealth and uHealth, vol. 6, no. 1, e28, 2018.

[52] Y. K. Bartlett, T. L. Webb, and M. S. Hawley, "Using persuasive technology to increase physical activity in people with chronic obstructive pulmonary disease by encouraging regular walking: A mixed-methods study exploring opinions and preferences," Journal of Medical Internet Research, vol. 19, no. 4, e124, 2017.

[53] G. Giunti, V. Mylonopoulou, and O. R. Romero, "More stamina, a gamified mHealth solution for persons with multiple sclerosis: Research through design," JMIR mHealth and uHealth, vol. 6, no. 3, e51, 2018.

[54] A. Tuzhilin, "Personalization: The state of the art and future directions," Business Computing, vol. 3, no. 3, pp. 3–43, 2009.

[55] A. Sunikka and J. Bragge, "Applying text-mining to personalization and customization research literature – Who, what and where?," Expert Systems with Applications, vol. 39, no. 11, pp. 10049–10058, 2012.

[56] J. Blom, "Personalization: a taxonomy," CHI'00 Extended Abstracts on Human Factors in Computing Systems, pp. 313–314, 2000.

[57] H. Fan and M. S. Poole, "What is personalization? Perspectives on the design and implementation of personalization in information systems," Journal of Organizational Computing and Electronic Commerce, vol. 16, no. 3–4, pp. 179–202, 2006.

[58] D. Wu, I. Im, M. Tremaine, K. Instone, and M. Turoff, "A framework for classifying personalization scheme used on e-commerce websites," 36th Annual Hawaii International Conference on System Sciences, Proceedings of the IEEE, 2003, doi: 10.1109/HICSS.2003.1174586.

[59] A. B. Kocaballi et al., "The personalization of conversational agents in health care: Systematic review," Journal of Medical Internet Research, vol. 21, no. 11, e15360, 2019.

[60] A. Shen and A. D. Ball, "Is personalization of services always a good thing? Exploring the role of technology-mediated personalization (TMP) in service relationships," Journal of Services Marketing, vol. 23 no. 2, pp. 79–91, 2009, doi:10.1108/08876040910946341.

[61] C. F. Surprenant and M. R. Solomon, "Predictability and personalization in the service encounter," Journal of Marketing, vol. 51, no. 2, pp. 86–96, 1987.

[62] K. P. Gwinner, M. J. Bitner, S. W. Brown, and A. Kumar., "Service customization through employee adaptiveness," *Journal of Service Research*, *8*(2), 131-148, 2005.

[63] B. Mittal and W. M. Lassar," The role of personalization in service encounters," Journal of Retailing, vol. 72, no. 1, pp. 95–109, 1996.

[64] O. Korhonen and M. Isomursu, "Identifying personalization in a care pathway: A single-case study of a Finnish healthcare service provider," 25th European Conference on Information Systems (ECIS) pp. 828–841, 2017.

[65] "Figma: the collaborative interface design tool," [Online]. Available from: https://www.figma.com/ 2021.06.29.

[66] R. Buchanan, "Wicked problems in design thinking," Design Issues, vol. 8, no. 2, pp. 5–21, 1992.

[67] G. Latham and E. A. Locke, "New developments in and directions for goal-setting research," European Psychologist, vol. 12, no. 4, pp. 290–300, 2007.

[68] E. A. Locke and G. P. Latham, "The development of goal setting theory: A half century retrospective," Motivation Science, vol. 5, no. 2, pp. 93–105, 2019, doi:10.1037/mot0000127.

[69] K. Williams, "Goal setting in sports," New Developments in Goal Setting and Task Performance, New York, NY: Routledge, pp. 375–398, 2013.

[70] A. Bandura and E. A. Locke, "Negative self-efficacy and goal effects revisited," Journal of Applied Psychology, vol. 88, no. 1, pp. 87–99, 2003.

[71] S. Hesse-Biber, P. Leavy, C. E. Quinn, and J. Zoino, "The mass marketing of disordered eating and Eating Disorders: The social psychology of women, thinness and culture." Women's Studies International Forum, 29(2), 208–224, 2006. https://doi.org/10.1016/j.wsif.2006.03.007

[72] R. Ryan and E. Deci, "Intrinsic and Extrinsic Motivations: Classic Definitions and New Directions," Contemporary Educational Psychology, vol. 25, no. 1, pp. 54–67, 2000, doi:10.1006/ceps.1999.1020.

[73] M. Sjölinder et al., "Perspectives on Design of Sensor Based Exergames Targeted Towards Older Adults," In: Zhou J., Salvendy G. (eds) Human Aspects of IT for the Aged Population, Applications in Health, Assistance, and Entertainment, ITAP 2018. Lecture Notes in Computer Science, vol 10927, Springer, Cham, pp. 395–414, 2018, doi:10.1007/978-3-319-92037-5_29.

Analysis of Short-term and Long-term Effects on Mental State of Suggestions Given by an Agent using Impasse Estimation

1st Yoshimasa Ohmoto
Department of Informatics
Graduate School of Integrated Science and Technology
Shizuoka University
Shizuoka, Japan
ohmoto-y@inf.shizuoka.ac.jp

2nd Hanako Sonobe
Department of Intelligence Science and Technology
Graduate School of Informatics
Kyoto University
Kyoto, Japan
sonobe@ii.ist.i.kyoto-u.ac.jp

3rd Toyoaki Nishida
Department of Informatics
Faculty of Informatics
The University of Fukuchiyama
Kyoto, Japan
toyoakinishida@gmail.com

***Abstract*—For an agent to teach a person a problem-solving attitude by giving him advice that does not directly contribute to solving the problem, a strategy that considers changes in the person's long-term attitude must be designed. This study aimed to investigate how the mental state of participants performing a task is affected during short-term and relatively long-term periods when they are advised either based on their conditions or mechanically at regular intervals. We focused on metacognitive suggestions during insight problem-solving as an example of advice that would be effective even if given by the agent. By these means, the effect on the human mental state over a relatively long period of time when the agent gives advice is examined. We conducted an experiment using two types of suggestion agents and observed that participants were likely to accept metacognitive suggestions provided by an agent when the suggestions were given based on an inner-state estimation of the participant. An analysis of mental state changes based on physiological indices suggested that the use of metacognitive suggestions by agents based on participants' conditions affected the mental state in problem-solving activities in the short and long term. It is also suggested that if the advice is not given depending on the situation, the effect of the advice in mitigating the impasse reduces as the task progresses. These findings will contribute towards the implementation of a tutoring agent.**

***Index Terms*—Human-agent interaction; metacognitive suggestion; insight problem solving.**

I. Introduction

This article is an extended version of the authors' paper "Difference in Attitudes toward Suggestions Given by an Agent using Impasse Estimation" [1], presented at the Twelfth International Conference on Advances in Human-oriented and Personalized Mechanisms, Technologies, and Services (CENTRIC2019). In this paper, based on the temporal changes in the participants' physiological indices measured during the experiment, we analyzed in detail the changes in mental state throughout the experiment, the changes in mental state for each task with different properties, and the changes in mental state due to the intervention by the agent. In this way, we investigated not only the effect on the mental state immediately after the agent's intervention, but also the long-term effect on it due to the repetition of the interventions of the agant. In addition, we described details of the specific interventions of the agents and the tasks performed in the experiment, which were omitted in the proceedings of CENTRIC2019 due to space limitations.

In learning and teaching situations, when learners are working on problem-solving, those who are knowledgeable about the task are often encouraged to develop an attitude toward learning. This is a way of thinking about problem-solving itself by encouraging them to broaden their horizons and learn via trial and error rather than by giving them direct advice to help solve the problem. This teaching strategy does not contribute directly to the solution of the problem. Therefore, if you do not consider the condition of the person you are communicating with, you may not be able to convey your intentions correctly or you may not be able to be considered your opinions. For example, if you repeatedly give advice about something that the listener does not perceive to be a problem, they may ignore the advice. Such problems become more pronounced when systems such as agents provide advice. One of the reasons for this is that agents' ability to grasp the situation seems to be relatively low from a human standpoint.

In order to avoid such issues, the agent needs to understand the human state and give advice. However, even a human often fails to estimate the mental state of the person they are communicating with. When you have a trustful relationship with the communication partner, this is not necessarily a fatal problem. However, in a human-agent interaction, the human often needs to infer the agent's behavior model based on a small number of interactions. Therefore, it is expected that a small number of failures in interaction will cause errors in the behavioral model of the agent constructed by humans. For example, one approach to getting people to accept an agent's advice is to show that the agent has expertise by having the agent consistently provide the appropriate advice [2]. The human often accepts the advice of the agent when the agent provides appropriate advice depending on the task situation.

However, if the agent fails early in the interaction because of misunderstandings, the human may stop accepting further advice. In addition, it is often difficult to determine whether or not the advice is appropriate when the advice does not directly lead to the correct answer for the task being performed. In such instances, effective advice cannot be given without considering the situation and intention of the person performing the task. In order for an agent to teach a person something like a problem-solving attitude by giving him advice that does not directly contribute to solving the problem, it is necessary to consider an advice-dispensing strategy that takes changes in the long-term attitude of the person into account.

Metacognitive suggestions are useful for problem-solving, although they do not lead to the direct outcome of the task being performed. Several previous studies had attempted to improve task performance using pre-training to induce metacognition. Patrick et al. [3] reported the impact of general metacognitive training on performance. Metacognitive suggestions may convey knowledge of how participants solve problems and can facilitate changes in their way of thinking during insight problem-solving (e.g., [4]). In this study, we consider metacognitive suggestions during insight problem-solving as an example of advice that would be effective even if the agent gave it. The effect of an agent's metacognitive suggestions on the human mental state over a relatively long period of time is then examined.

We tried to encourage the acceptance of metacognitive suggestions from agents by controlling the contents and timing of suggestion presentation as per the state of the participant. In many previous studies (e.g., [5] [6]), the agent advises a participant when there is a pause in the conversation. In this study, in order to provide convincing advice to humans, we focused not on the content of the problem, but on the state of thinking about the problem and the awareness of problems. The advantage of this method is that it does not recognize the content of the task and can provide advice based on how difficult the person is feeling the task at an appropriate time.

On the other hand, it takes a relatively long period of time for the learners to develop an attitude toward learning and toward problem-solving. In this study, participants were asked to repeatedly perform a similar task that became progressively more difficult. The purpose of this study was to investigate how the mental state of the participants in performing a task is affected when the participants are given advice either depending on their conditions or mechanically at regular intervals. If it is important to be able to customize the timing of providing advice as per the condition of the participant in order for the agent's advice to be understood with long-term effect, the method proposed in this paper may be useful when developing a tutoring agent.

The present paper is organized as follows. The Suggestion system using impasse estimation section contains an explanation of a system developed to give metacognitive suggestions based on the estimated state of the person performing the insight problem-solving task. The Experiment section describes the results of an experiment to evaluate the system implemented on the agent. In the Discussion section, the achievements of this research and some future works are described. The conclusions are presented in the Conclusion section.

II. Suggestion system using impasse estimation

Insight problem-solving contains four steps: impasse, incubation, illumination, and validation [7]. We focus on the impasse step in which people repeatedly searches inappropriate problem space that does not include a solution. In the impasse step, advice from other perspectives is useful for constraint relaxation and a switch of problem space. Metacognitive suggestion is one method of providing acceptable advice for the constraint relaxation [3] [8] [9] [10]. The metacognitive suggestion is confirmed to be effective even if it is presented at random timing. This is because the insight problem-solving task is prone to fall into the impasse state, and therefore there is a certain probability of being in the impasse state when presented at random.

In order to confirm the appropriate timing of advice in an insight problem-solving task, we conducted a preliminary experiment in which an experimenter determined the content and timing of the agent's metacognitive suggestions using the Wizard of Oz (WoZ) and presented them to the person performing the task. The task in the preliminary experiment was an "escape room game" in which players were often at a stalemate because they were required to think from a different perspective to win this game. In this game task, the participant must escape from a virtual room using various game objects. After this preliminary experiment, some participants reported that they were "given proper advice", so we thought that the advice of the agent by WoZ operation was accepted. When we observed the behavior of the participants and advice of the experimenter in the preliminary experiment, the experimenter provided suggestions when the participant seemed to be at a stalemate. We regarded this state of stalemate as an impasse. We consider that the state of stalemate is one of the appropriate clues to provide metacognitive suggestions. In accordance with this concept, we developed a system to provide advice by estimating whether the interaction partner is in an impasse state while working on an insight problem-solving task.

A. Estimation of strategies to perform the insight problem solving task

In order to find typical strategies to perform the insight problem-solving task, we observed the behavior of the participants in the preliminary experiment. As a result, it was expected that the participants switched two strategies: depth-first search and breadth-first search. In the state of depth-first search, participants focused on a particular object and looked for ways to use it successfully. In the state of breadth-first search, participants saw the overall situation of the task to search whether there were any missing or untried methods. Since it is conceivable that a stalemate may occur while executing each strategy, human inner states in insight problem-solving can be classified into 4 states (table in Figure 1). In

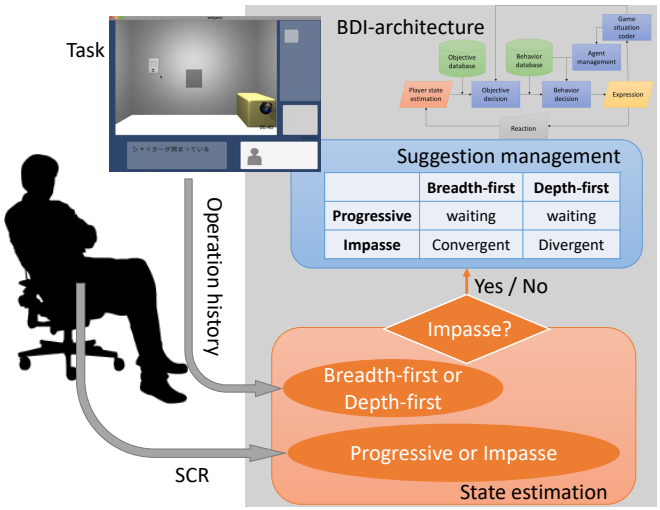


Fig. 1. Outline of the system architecture.

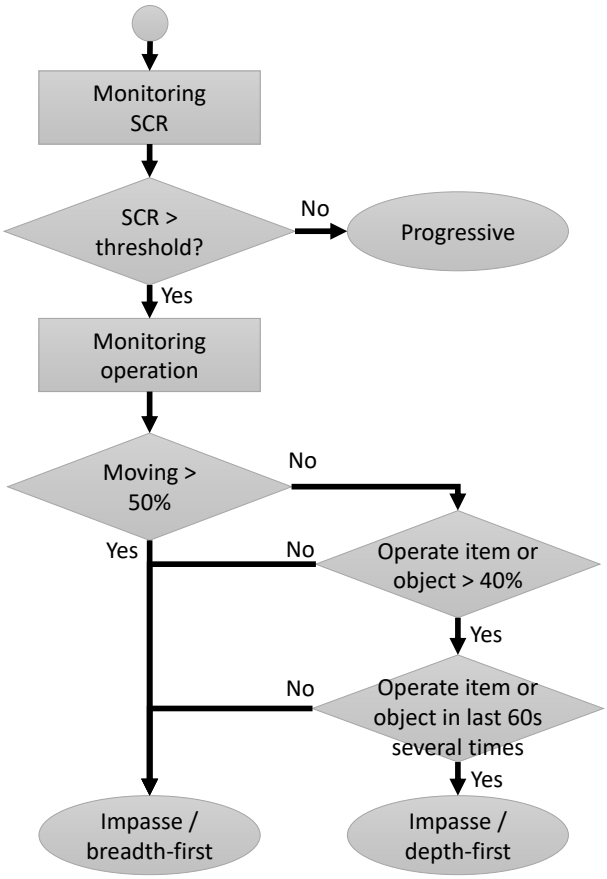


Fig. 2. Outline flowchart of the thinking-mode estimation.

the advice by the experimenter, there were many suggestions that urged the participant to look for other ways to solve the task in the depth-first search, and there were many suggestions that encouraged the participant to look back on his/her own behavior and to focus on the specific object in the breadth-first search.

It is difficult to infer the inner state of thinking from the participant's behavior, specifically the inner state of thinking whether the participant is at a stalemate. To estimate this inner state, we analyzed physiological indices obtained during the preliminary experiment. In our previous work, we reported to estimate the feeling of difficulty of the task by using physiological indices [11]. As a result, it was frequently observed that Skin Conductance Response (SCR) was often activated, when the unfamiliar object was discovered during the task and when the situation in the task was changed. In addition, even when the situation did not change, the SCR was often responsive when with repeated trial and error such as looking for hints or checking previous information. Therefore, we regard the state as a non-impasse state (the participant is not at a stalemate) when the responses of SCR are frequently observed, and we regard the state to have shifted to the impasse state (the participant is at a stalemate) when the response of SCR is not observed for a certain time.

We also measured the electrocardiogram. However, we have not been able to obtain a useful feature for estimating task impasse from the electrocardiogram. Therefore, no electrocardiogram data was input to the system. We used electrocardiographic data to assess participant's mental states to tasks.

B. The outline of the suggestion system using impasse estimation

Figure 1 shows an outline of the agent design. This agent basically decides own behavior based on the Belief-Desire-Intention (BDI) architecture. This agent estimates two kinds of user states: a thinking mode (depth-first or breadth-first) and a state of the stalemate (impasse or progressive). The user's overall states are categorized into one of four combinations: depth-first/progressive, breadth-first/progressive, depth-first/impasse, and breadth-first/impasse. The agent provides a metacognitive suggestion when the estimated user's state includes "impasse." A convergent suggestion is provided in the state of breadth-first/impasse. A divergent suggestion is provided in the state of depth-first/impasse.

The user's physiological index and behavior are measured to estimate the user's state. The state of the stalemate is estimated using the measured SCR. The agent estimates the user's state as impasse when the SCR does not respond during a defined time window. The time window and the threshold to estimate the state of the stalemate are decided based on the measured data for two minutes from the start of the task. To estimate the thinking mode, the operation history log of the user is used. When the user repeatedly operates a game object (such as, a key, a scissors, a piece of paper, a door, a dial plate, a drawer of a desk, a closet, a button and a safe) in high frequency, the agent estimates the thinking mode to be depth-first search. The outline flowchart of the thinking-mode estimation is shown in Figure 2.

Ten convergent and ten divergent suggestions were pre-

pared. The suggestions were not dependent on a particular task because they were metacognitive suggestions. One of the suggestions is selected randomly when the agent provides an advice. In general situations, it is necessary to give advice considering the context of the task though it is a metacognitive suggestion. In this study, we focused on the effect of controlling the timing of the metacognitive suggestions provided based on the state of the user. Therefore, the agent advised only considering whether the context was divergent or convergent in our experiment. The following is a list of metacognitive suggestions we prepared.

Divergent suggestions

- Why don't you look at something a little different?
- Don't stick to the way you've been doing things, think about different ways.
- Why don't you consider some other ways to proceed?
- Think of a way that's different from the way you failed.
- Is there anything else you can do?
- Why don't you think about something different from what you have been looking at?
- Please try to look at the situation from a different point of view.
- Let's think about what else you can do.
- Is there anything else you haven't done yet?
- Try to get rid of your assumptions.

Convergent suggestions

- Let's think about what's important in what you have seen.
- Let's try what comes to mind.
- Try to sort out what you have been doing.
- Think about what you need to do to escape.
- Have you missed anything so far?
- Let's think about what's been the inspiration so far.
- What element do you think is involved in the escape?
- Let's think back to what you have done so far.
- If you think you can do something, try it.
- Why don't you narrow down what you're focusing on?

III. Experiment

When we try to intervene in the behavior or decision-making of the other person by providing advice, especially during interaction with a less socially related interaction partner, it is important to provide appropriate advice based on the estimation of the partner's inner state. We considered the metacognitive suggestion in the insight problem-solving task as an example of the useful advice that the agent can provide. Then we proposed the suggestion system based on the impasse state estimation of the partner in order to accept the suggestion by the agent. We used two types of suggestion agents in this experiment. One was a state-considering agent that estimated the user's state before providing a metacognitive suggestion (sc-group). Another was a fixed-interval agent that provided a metacognitive suggestion in three-minute intervals (fi-group).

A. Task

Participants played an "escape from the room" game. The objective of this game is to escape from a closed space such as a room by utilizing game objects and items that are placed in that space. In most escape games, the player cannot escape from a room in a simple way, such as unlocking the door. The player escapes by searching for keys that are hard to find, by manipulating game objects in specific steps, and/or by using items in ways that are different from their common uses. In this game, players were often at a stalemate because they are required to think from a different angle to escape from a virtual room using various game objects.

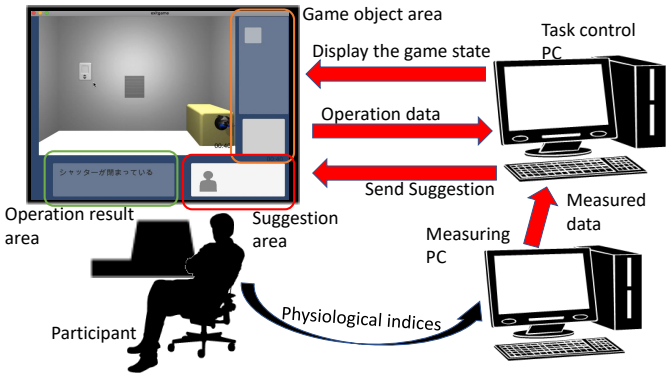


Fig. 3. The experimental setting.

The player can see images representing the four directions of the room, as well as partial enlargements of the images. If the player can explore the room and is able to move to the other side of the wall by opening a door or making a hole in the wall, the escape is successful. There are several non-movable game objects in the room (desk, chair, window, safe, etc.) and movable items (hammer, key, notepad, etc.). Descriptions of the game objects or items are displayed in the description display area in the game screen. The player can use the keyboard and mouse to change his/her viewing direction, zoom in on objects of interest, and use items.

The participants are asked to escape from three rooms. As the number of game objects and items and the steps to escape increase, the difficulty of escaping gradually increases. The order of the rooms that the participant escaped from was fixed. There was a 15-minute time limit to escape from each room. The suggestion agent explained the procedure for escaping from the room when the participant exceeded this time limit. After escaping from a room, the participant was allowed to rest. Participants were able to continue the game by pressing the start button when they wanted to resume.

B. Exprimental setting

Figure 3 shows the experimental setting. Each participant sat in front of a 27-inch monitor that displayed the game. A video camera was placed behind the participant to record his/her behavior and the game playing screen. The participant's voice was recorded using microphones. Polymate was used to measure SCR and the electrocardiogram (heart rate variability). SCR was measured with electrodes attached to the first and third fingers of the participant's non-dominant hand.

The electrocardiogram was measured by connecting electrodes with paste to the participant's left side, the center of the chest, and both ears for ground and reference. The experimenter sat out of view of the participant and observed the participant's behavior. The suggestions by the agent were provided using audio and text. The participants performed the task using a mouse.

C. Procedure

First, each participant was briefly instructed on the experimental procedure. Electrodes for measuring SCR and the electrocardiogram values were then attached to the participant's left hand and chest. After the installation, each participant played a practice game to confirm the operating method and basic flow of the game. The experimenter instructed the participant on the basic operation method. In addition, the participant was given an overview of the agent providing metacognitive suggestions. After receiving questions from the participant and confirming his/her understanding, the participant started the "escape from the room" experiment. After the experiment, the participant answered NASA Task Load Index (NASA-TLX) to measure the mental workload.

Forty-two undergraduate students, 27 males and 15 females, participated in the experiment. The average age was 20.8 years with a standard deviation of 1.9 years. We eliminated 13 participants because they did not need suggestions to escape from one of the rooms. Therefore, we used data of 29 participants (sc-group: 14 participants, fi-group: 15 participants).

D. Results

We analyzed the frequency of metacognitive suggestions, operation history log, mental workload, and physiological indices. For the frequency of metacognitive suggestions, we analyzed how many times each of the participants in each group provided metacognitive suggestions to encourage divergence and convergence in each room. From the operation history log, it was analyzed whether the state transition was carried out within a fixed time. The analysis range was 10 seconds after the suggestion. In the analyses of the physiological indices, we used heart rate variability (this is converted to cardiac sympathetic index (CSI) and cardiac vagal index (CVI)), and SCR. In the analysis of mental workload of the task, we used the Japanese version of the NASA-TLX, which represents the physical and psychological load of the task.

1) The frequency of metacognitive suggestions : We analyzed whether there was a difference in the frequency of metacognitive suggestions provided in the sc-group compared with the fi-group. There were two types of the metacognitive suggestions (divergent and convergent), so we performed a 2 (group: state-considering or fixed-interval) x 3 (room: first, second or third) analysis of variance (ANOVA) separately. Since each participant spent different amounts of time in each room, we compared the number of suggestions per minute. Logit transformed values were used in ANOVA to test for differences. The results are shown in Figure 4, Figure 5, Table I, Table II, Table III, and Table IV. In the tables, "SS" means the sum-of-squares, "df" means the degrees of freedom, "MS" means the mean squares, "F" means the F ratio, and "p" means the p-values.

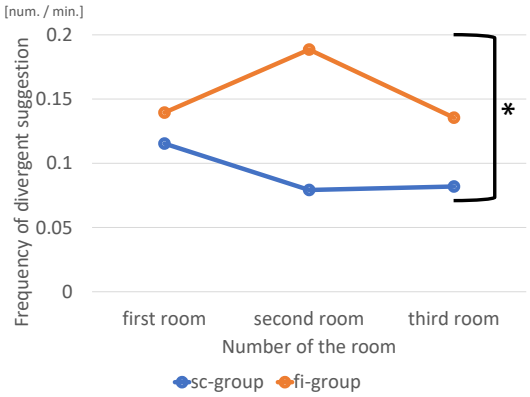


Fig. 4. The frequency of divergent metacognitive suggestions per minute.

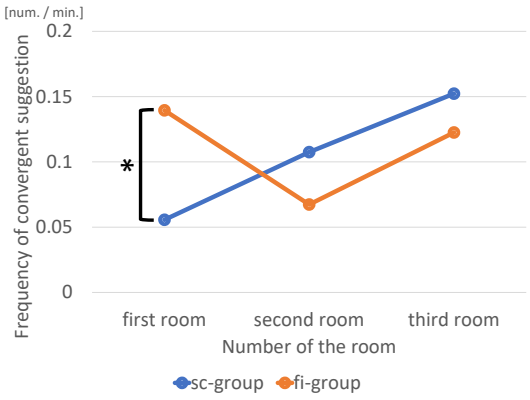


Fig. 5. The frequency of convergent metacognitive suggestions per minute.

In the divergent suggestions, there were significant differences between groups (sc-group $<$ fi-group) and between rooms (first and second $>$ third). The interaction was also significant. When tested for simple main effects, there were significant differences between groups in second room and third room (sc-group $<$ fi-group). There was also a significant difference between the rooms in the sc-group (first $>$ second and third). This result indicates that, in the sc-group, a relatively large amount of divergent suggestions was provided in the first room where the task execution method is unclear for the participants, and that trial and error is encouraged. In addition, in second room and third room where people seem to be used to the task, suggestions were reduced.

In the convergence suggestions, there was no significant difference between groups, but there were significant differences between rooms (first and second $<$ third). The interaction was also significant. A simple main effect test showed a significant difference between groups in first room (sc-group $<$ fi-group). It was also found that there were significant differences between the rooms in the sc-group (first $<$ second and third). This result shows that the convergent suggestions in

TABLE I
RESULT OF THE ANOVA ON THE FREQUENCY OF DIVERGENT METACOGNITIVE SUGGESTIONS.

source	SS	df	MS	F	p
A: group	6.50	1	6.50	14.83	<0.001 ****
error[S(A)]	11.84	27	0.44		
B: room	1.81	2	0.90	5.13	0.0091 **
AB	1.17	2	0.58	3.32	0.044 *
error[BS(A)]	9.49	54	0.18		

+ $p < .10$; * $p < .05$; ** $p < .01$; *** $p < .005$; **** $p < .001$

TABLE II
THE SIMPLE MAIN EFFECT OF THE ANOVA ON THE FREQUENCY OF DIVERGENT METACOGNITIVE SUGGESTIONS.

effect	SS	df	MS	F	p
A(first)	0.38	1	0.38	1.43	0.235
A(second)	4.32	1	4.32	16.39	<0.001 ****
A(third)	2.98	1	2.98	11.30	0.0012 ***
error		81	0.26		
B(state-considering)	1.90	2	0.95	5.42	0.0072 **
B(fixed-interval)	1.07	2	0.53	0.04	0.056 +
error		54	0.18		

+ $p < .10$; * $p < .05$; ** $p < .01$; *** $p < .005$; **** $p < .001$

TABLE III
RESULT OF THE ANOVA ON THE FREQUENCY OF CONVERGENT METACOGNITIVE SUGGESTIONS.

source	SS	df	MS	F	p
A: group	0.77	1	0.77	2.03	0.165
error[S(A)]	10.21	27	0.38		
B: room	1.15	2	0.57	4.10	0.022 *
AB	3.17	2	1.58	11.30	<0.001 ****
error[BS(A)]	7.57	54	0.14		

+ $p < .10$; * $p < .05$; ** $p < .01$; *** $p < .005$; **** $p < .001$

TABLE IV
THE SIMPLE MAIN EFFECT OF THE ANOVA ON THE FREQUENCY OF CONVERGENT METACOGNITIVE SUGGESTIONS.

effect	SS	df	MS	F	p
A(first)	3.83	1	3.83	17.42	<0.001 ****
A(second)	0.015	1	0.015	0.067	0.796
A(third)	0.099	1	0.099	0.45	0.50
error		81	0.26		
B(state-considering)	3.19	2	1.60	11.39	<0.001 ****
B(fixed-interval)	1.12	2	0.56	4.01	0.024 *
error		54	0.14		

+ $p < .10$; * $p < .05$; ** $p < .01$; *** $p < .005$; **** $p < .001$

the sc-group was reduced in first room, which was a relatively simple.

Overall, the control of metacognitive suggestions was reasonable to some extent.

2) Operation history log: After the metacognitive suggestion was provided, we analyzed from participants' operation history log to determine whether they were acting in line with the suggestion. From the operation history log, we checked whether the transition to another state occurred within 10 second after the suggestion was given. In divergent suggestions, if a state transition was made, it was considered that the suggestion was accepted. In convergent suggestions, if no state transition was made, the suggestion was accepted. The result is shown in Figure 6 and Figure 7. The chi-squared test was applied to determine whether there was a difference between the groups in the acceptance rate of divergent suggestions and the acceptance rate of convergent suggestions.

We compared the acceptance rate of all suggestions between groups. As a result, the acceptance rate of all suggestions in the sc-group was significantly higher than that in the fi-group (p = 0.0013). We compared the acceptance rates of divergent suggestions and convergent suggestions between groups. Although there was no significant difference in divergent suggestions (p = 0.01), the acceptance rate of convergent suggestions in the sc-group was significantly higher than that in the fi-group (p = 0.0061). We compared the acceptance rates of divergent suggestions and convergent suggestions in each group between rooms. In third room, the acceptance rates of both divergent suggestions and convergent suggestions in the sc-group were significantly higher than those in the fi-group (divergent: first room p = 0.72, second room p = 0.16, third room p = 0.005, convergent: first room p = 0.44, second room p = 0.36, third room p = 0.014). In addition, in the sc-group, the acceptance rates in third room were higher than those in first room, and the acceptance rates seems to be gradually increasing. It is not clear whether this is because the difficulty of the room is increasing or because the reliability of the agent's suggestions is increasing. In any case, the results showed that the participants were likely to accept the metacognitive suggestions provided by the agent when the suggestions were given based on the inner state estimation of the participant.

3) Mental workload: We measured mental workload using NASA-TLX. This is major method to measure the mental workload. Figure 8 shows the results. With the exception of "performance," the sc-group reported an overall lower mental workload than the fi-group. We performed Welch's t-test on the total score between the two groups and there is no significant difference (t(27)=-1.42, p=0.17). We also performed Welch's t-test on each individual score between the sc-group and the fi-group. There was a significant difference regarding the data of "temporal demand" (sc-group $<$ fi-group, t(27)=-2.18, p=0.038). The results suggest that advice based on human internal state estimation reduces some of the human mental workload. At the same time, it shows that the overall effect is not significant.

4) Analysis of changes in mental state based on physiological indices: To investigate whether a change that was not apparent from the human's behavior occurred in their mental state, we analyzed physiological indices. The physiological indices used in this study are heart rate variability (this is converted to CSI and CVI), and SCR.

CSI is one of the indices of sympathetic nerve activity. The sympathetic nervous system's primary function is to stimulate the body's fight-or-flight response, in terms of perceptible reactions such as tension and excitement. The CVI is one of

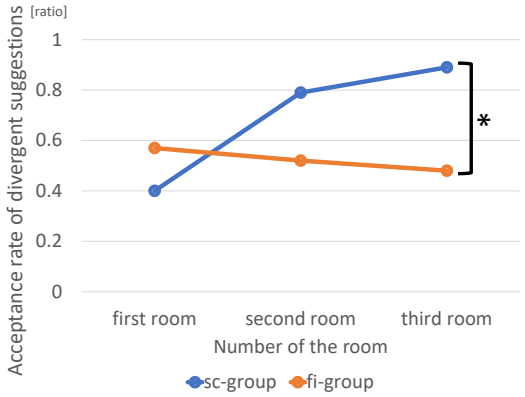


Fig. 6. Acceptance rates of divergent metacognitive suggestions.

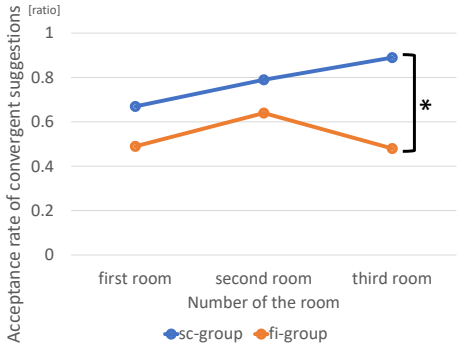


Fig. 7. Acceptance rates of convergent metacognitive suggestions.

the indices of parasympathetic nerve activity. The parasympathetic system is responsible for stimulating the ”rest-and-digest” activities that occur when the body is at rest and relaxed. We used the geometric Lorenz plot method [12] to calculate the CVI and CSI.

SCR is a kind of electro-dermal activity that includes skin potential activity and skin conductance activity. People sweat during exercise but their palms have only a few sweat glands for body temperature adjustment. Therefore, by measuring the electrical resistance on the palms, it is possible to check for the presence or absence of emotional perspiration [13]. Given that the underlying mechanisms of SCR and electrocardiograms are different, we assumed that they could be used to distinguish between different responses from different sources of stress. Sweating is controlled by the sympathetic nervous system [14] and can be induced by emotional stimuli, intellectual strain, or painful cutaneous stimulation. The underlying mechanisms of SCR are related more to anticipation, expectation, and attention concentration [15]. We thus expected that the SCR could be used to tell when someone is dealing with an unexpected or thrilling situation.

a) Analysis of changes in mental state in each room: In order to analyze the changes in the mental state of participants who tackled a difficult task while receiving advice from an agent, we analyzed the state in which they started working on the task and the state before they reached a solution in each

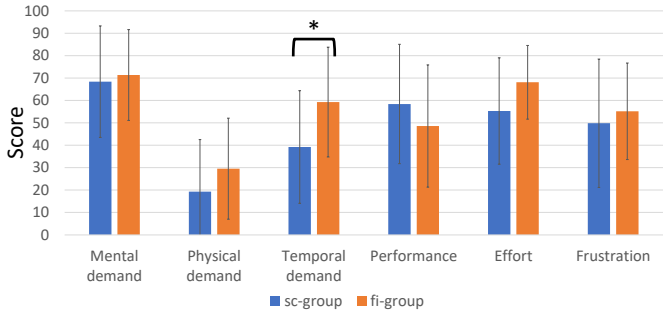


Fig. 8. Results of mental workload measurements.

room. For this purpose, the time required for escaping from each room was divided into three for each participant and the early and last data were extracted. For this data, CSI, CVI, and SCR were calculated. The results are presented in Table V. Paired t-test was performed on these data for each group.

In first room, the sc-group showed a significant increase in CSI as problem-solving progressed (CSI: $t(13)=-2.56$, $p=0.024$), but there was no significant difference in CVI and SCR. In the fi-group, there were no significant differences in CSI and CVI, but there was a marginally significant increase in SCR (SCR: $t(14)=-2.10$, $p=0.054$).

In second room, the sc-group showed a marginally significant increase in CSI as problem-solving progressed (CSI: $t(13)=-2.03$, $p=0.063$) but there were no significant differences in CVI and SCR. In the fi-group, there were no significant differences in CSI and CVI but there was a significant increase in SCR (SCR: $t(14)=-2.38$, $p=0.032$).

In third room, the sc-group showed a marginally significant increase in CSI as problem-solving progressed (CSI: $t(13)=-1.94$, $p=0.074$), but there were no significant differences in CVI and SCR. In the fi-group, there were no significant differences in CSI, CVI, and SCR.

Overall, the sc-group was adaptively advised via metacognitive suggestions and the sympathetic nervous system was more active in the late phase of problem-solving than in the early phase of problem-solving, suggesting that they had been working diligently on the problem-solving until escape. In other words, it can be suggested that by providing advice based on the condition of the participants, they were able to increase their involvement in the problem-solving efforts. In the fi-group, effects on SCR are observed in first room and second room. SCR tends to be lower when a person is absorbed in repeating tasks and higher when they are engaged in various trials and errors [16]. As the difference seems to be caused by the low SCR when they started working on the problem-solving, it is presumed that the fi-group might have fallen into an impasse and repeated the same action in that time. We think that the reason here were no differences in any of the physiological indices when the participants were in third room is that the problem-solving became more complicated, so the problem space became wider, and impasse in the form of repeating the same action was less likely to occur.

TABLE V
THE AVERAGES OF CSI, CVI AND SCR IN EACH ROOM.

room	group	CSI		CVI		SCR	
		early part	last part	early part	last part	early part	last part
first	sc	1.18	1.29	3.22	3.26	15.03	15.20
	fi	1.18	1.26	3.27	3.32	14.79	15.30
second	sc	1.29	1.38	3.29	3.30	14.78	14.97
	fi	1.42	1.36	3.38	3.36	14.00	14.96
third	sc	1.43	1.54	3.33	3.37	14.78	14.97
	fi	1.40	1.52	3.39	3.40	14.99	15.21

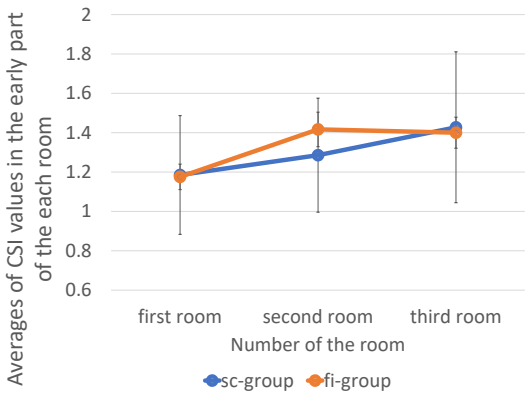


Fig. 9. The averages of CSI in the early part of the problem-solving in each room.

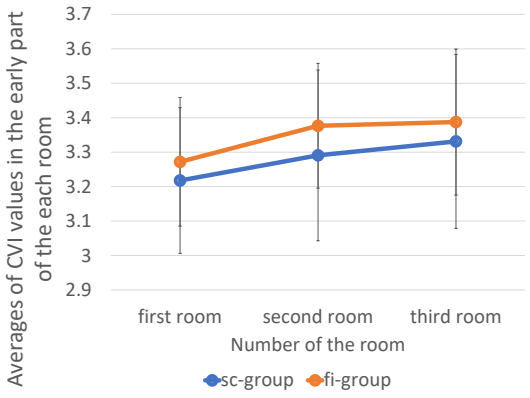


Fig. 10. The averages of CVI in the early part of the problem-solving in each room.

b) Analysis of changes in mental state in three consecutive rooms: In the early part of the problem-solving, differences between sc-group and fi-group seem to appear. Therefore, we analyzed the effect of the three rooms. A one-way analysis of variance was performed on the CSI, CVI, and SCR data in each group. The averages of each data in the early part of the problem-solving in each room are listed in Figure 9, Figure 10, and Figure 11.

The results indicate that there were no significant differences in CSI, CVI, and SCR in the sc-group, but there was a significant difference in CSI in the fi-group ($F(2, 14)=3.40$, $p=0.048$). There was a significant difference between first

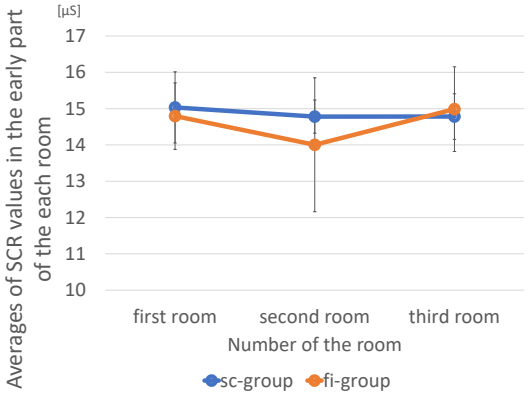


Fig. 11. The averages of SCR in the early part of the problem-solving in each room.

TABLE VI
MULTIPLE COMPARISONS IN THE MAIN EFFECT OF CSI BY RYAN'S METHOD.

pair	r	nominal level	t	p	sig.
second - first	3	0.33	2.33	0.027	s.
third - first	2	0.66	2.17	0.038	n.s.
second - third	2	0.66	0.16	0.874	s.

room and second room as well as between first room and third room (Table VI). The participants in the fi-group were more engaged in the problem-solving from the beginning in second room compared to first room, while participants from the sc-group demonstrated a gradual increase as they moved through the rooms. These differences may reflect the different attitudes toward problem-solving that were learned during the problem-solving trials in the previous room(s).

c) Analysis of changes in mental state after advice was given: The CSI, CVI, and SCR were calculated and compared for 10 seconds before and after giving the advice in each room to investigate the effect on the mental state of the participants immediately after the advice was given. Averages of values before and after advice are presented in Table VII. Paired t-tests were performed on these data for each group.

In first room, the sc-group showed a significant increase in SCR immediately after advice was given (SCR: $t(34)=-4.51$, $p<0.0001$) but there were no significant differences in CSI and CVI. In the fi-group too, there was a significant increase in SCR immediately after advice was given (SCR: $t(45)=-4.41$,

TABLE VII
THE AVERAGES OF CSI, CVI AND SCR BEFORE AND AFTER SUGGESTIONS IN EACH ROOM.

		CSI		CVI		SCR	
room	group	before	after	before	after	before	after
first	sc	1.21	1.20	3.25	3.26	13.26	16.56
	fi	1.23	1.25	3.31	3.29	14.11	19.14
second	sc	1.44	1.42	3.32	3.32	13.59	16.88
	fi	1.41	1.33	3.36	3.36	15.00	18.43
third	sc	1.48	1.44	3.34	3.34	13.44	16.41
	fi	1.35	1.38	3.36	3.37	15.02	16.73

$p<0.0001$) but again there were no significant differences in CSI and CVI.

In second room, the sc-group demonstrated a significant increase in SCR immediately after advice was given (SCR: $t(39)=-3.64$, $p=0.00039$) but there were no significant differences in CSI and CVI. In the fi-group, there was a significant increase in SCR (SCR: $t(37)=-3.10$, $p=0.0019$) but no significant differences in CSI and CVI.

In third room, the sc-group demonstrated a significant increase in SCR immediately after advice was given (SCR: $t(57)=-4.30$, $p<0.0001$), but there were no significant differences in CSI and CVI. In the fi-group, there was a marginally significant increase in SCR (SCR: $t(55)=-1.88$, $p=0.065$) but no significant differences in CSI and CVI.

Observing the overall trend, the sc-group appears to demonstrate that there is a relatively consistent and strong influence of advice on the participants, whereas the fi-group appears to be less influenced by advice as the task progresses.

IV. DISCUSSION

We hypothesized that participants would be likely to accept the advice based on the estimation of the inner state of the human even when the agent provided advice. In this research, we focused on metacognitive suggestions in an insight problem-solving task, which is one of the examples of useful advice that the agent can provide. We investigated the effects of metacognitive suggestions that controlled the timing of presentation based on human inner state. We implemented an agent that estimated two kinds of user states: a thinking mode (depth-first or breadth-first) and a state of stalemate (impasse or progressive). The agent categorized the participant's overall state as one of four combinations: depth-first/progressive, breadth-first/progressive, depth-first/impasse, and breadth-first/impasse. The agent provided a metacognitive suggestion with the goal of getting humans out of the impasse state.

We conducted an experiment using two suggestion agents. One was a state-considering agent that estimated the user's state before providing a metacognitive suggestion. The other was a fixed-interval agent that provided a metacognitive suggestion at three-minute intervals. Based on results from the analysis of the operation history log, the acceptance rate of suggestions in the sc-group was significantly higher than that in the fi-group. In other words, the attitude to the metacognitive suggestions given by the agent was different between the participants in fi-group and those in sc-group. The participants in sc-group believed that the content of the suggestions given by the agent should always be considered. By contrast, participants in the fi-group typically thought that the agent's suggestions presented general knowledge and tended to accept useful ones regardless of the task status. The results of the mental workload suggest that participants in the fi-group might interpret the agent's suggestions as a kind of facilitation of the task execution rather than human assistance.

An analysis of mental-state changes based on physiological indices suggested that the use of metacognitive suggestions by agents based on participants' conditions affected the mental state in problem-solving activities in the short as well as long term. Overall, changes in mental state were mainly reflected in CSI and SCR. While working on a single problem-solving task, it was suggested that when the agent's advice was given based on the participant's condition, the participants worked more diligently on the problem-solving task. In multiple tasks in succession, especially in situations where participants were beginning to tackle a new problem, they tended to be less likely to fall into the impasse of trying the same solution over and over again when they received advice that was provided based on their condition. As a short-term effect, a comparison of physiological indices before and after advice was provided suggested that both groups responded to such advice by trying new trial-and-error activities. However, it was also suggested that if the advice was not given depending on the situation, the effect of the advice in mitigating the impasse would be reduced as the task progressed.

During one room trial, the effect of advice based on the participant's condition appeared in the CSI. By contrast, the direct effect of the advice appeared in the SCR, while there was no change in the CSI. In addition, the direct effect of the advice was basically unaffected by the timing of the advice. The direct effect of advice on participants is providing a specific way of problem-solving. Therefore, it is suggested that the change in the current way of problem-solving by receiving the advice itself affects the mental state of the participants. It also suggests that mitigating the susceptibility to falling into impasses and changing attitudes toward the problem-solving requires appropriate advice based on the participant's condition. This is consistent with the different interpretations of the sc-group and fi-group advice described above. In other words, advice on actual task performance is considered to be acceptable to the person by providing appropriate advice on

the status of the task, regardless of the state of the person. However, if we expect to improve the attitude toward the task through advice, such as to make the person have a broader perspective and be less likely to fall into an impasse, it is important to understand the person's condition at the time of giving the advice. The advice provided by the agent in this study was not directly helpful in solving the problem, but was a metacognitive suggestion that indirectly suggested how to solve the problem. Nonetheless, it is interesting to note that the immediate effect of the advice differed from the overall effect of the task execution. This is a finding that will contribute to the implementation of a tutoring agent. It suggests that an agent supporting active learning, which needs to maintain a positive attitude toward learning tasks, should provide advice and intervention based on an understanding of the human's current condition.

V. Conclusions

The aim of this study is to investigate how the mental state of participants performing a task is affected in the short term and for relatively long-term periods when the participants are given advice based on their conditions as opposed to mechanically at regular intervals. We implemented an agent that estimated two kinds of user states: a thinking mode (depth-first or breadth-first) and a state of stalemate (impasse or progressive). Based on experiments using two types of suggestion agents, we suggest that participants are more likely to accept metacognitive suggestions provided by agents when the suggestions are provided based on an inner-state estimation of the participant. With respect to the participant's mental state, an analysis of mental-state changes based on physiological indices suggests that the use of metacognitive suggestions by agents according to participants' conditions affects the mental state in problem-solving activities in the short and long term. As a short-term effect, a comparison of physiological indices before and after advice suggests that both groups responded to advice by trying new trial-and-error activities. It is also suggested that if the advice is not given depending on the situation, the effect of the advice in mitigating the impasse reduces as the task progresses. These findings will contribute towards the implementation of a tutoring agent.

Acknowledgment

This research is supported by Grant-in-Aid for Young Scientists (B) (KAKENHI No. 16K21113), and Grant-in-Aid for Scientific Research on Innovative Areas (KAKENHI No. 26118002) from the Ministry of Education, Culture, Sports, Science and Technology of Japan.

References

[1] Y. Ohmoto, H. Sonove, and T. Nishida, "Difference in attitudes toward suggestions given by an agent using impasse estimation," in *The Twelfth International Conference on Advances in Human-oriented and Personalized Mechanisms, Technologies, and Services*. IARIA, 2019, pp. 11–16.

[2] R. G. Hass, "Effects of source characteristics on cognitive responses in persuasion," *Cognitive Responses in Persuasion*, pp. 141–172, 1981.

[3] J. Patrick, A. Ahmed, V. Smy, H. Seeby, and K. Sambrooks, "A cognitive procedure for representation change in verbal insight problems," *Journal of Experimental Psychology: Learning, Memory, and Cognition*, vol. 41, no. 3, p. 746, 2015.

[4] Y. Hayashi, "Social facilitation effects by pedagogical conversational agent: Lexical network analysis in an online explanation task," *International Educational Data Mining Society*, 2015.

[5] S. T. Iqbal and B. P. Bailey, "Leveraging characteristics of task structure to predict the cost of interruption," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 2006, pp. 741–750.

[6] K. Isbister, H. Nakanishi, T. Ishida, and C. Nass, "Helper agent: Designing an assistant for human-human interaction in a virtual meeting space," in *Proceedings of the SIGCHI conference on Human Factors in Computing Systems*. ACM, 2000, pp. 57–64.

[7] S. Ohlsson, "Information-processing explanations of insight and related phenomena," *Advances in the Psychology of Thinking*, vol. 1, pp. 1–44, 1992.

[8] T. Okada and H. A. Simon, "Collaborative discovery in a scientific domain," *Cognitive Science*, vol. 21, no. 2, pp. 109–146, 1997.

[9] H. Shirouzu, N. Miyake, and H. Masukawa, "Cognitively active externalization for situated reflection," *Cognitive Science*, vol. 26, no. 4, pp. 469–501, 2002.

[10] Y. Hayashi, "Togetherness: Multiple pedagogical conversational agents as companions in collaborative learning," in *International Conference on Intelligent Tutoring Systems*. Springer, 2014, pp. 114–123.

[11] Y. Ohmoto, T. Matsuda, and T. Nishida, "Experimentally analyzing relationships between larner's status in the skill acquisition process and physiological indices," *International Journal on Advances in Life Sciences*, vol. 9, no. 3 and 4, pp. 127–136, 2017.

[12] M. Toichi, T. Sugiura, T. Murai, and A. Sengoku, "A new method of assessing cardiac autonomic function and its comparison with spectral analysis and coefficient of variation of r–r interval," *Journal of the Autonomic Nervous System*, vol. 62, no. 1, pp. 79–84, 1997.

[13] R. Edelberg, "Electrical activity of the skin: Its measurement and uses in psychophysiology," *Handbook of Psychophysiology*, vol. 12, p. 1011, 1972.

[14] E. F. Bartholomew, F. Martini, and W. B. Ober, *Essentials of Anatomy & Physiology*. Benjamin Cummings, 2007.

[15] K. Hugdahl, *Psychophysiology: The Mind-Body Perspective*. Harvard University Press, 1995.

[16] Y. Ohmoto, S. Takeda, and T. Nishida, "Effect of visual feedback caused by changing mental states of the avatar based on the operator's mental states using physiological indices," in *International Conference on Intelligent Virtual Agents*. Springer, 2017, pp. 315–324.

Playing Halma with Swarm Intelligence

Isabel Kuehner

Baden-Wuerttemberg Cooperative State University
Mannheim, Germany
Email: isabel.kuehner@isik-media.de

Adrian Stock

Baden-Wuerttemberg Cooperative State University
Mannheim, Germany
Email: adrian.stock@isik-media.de

***Abstract*—Swarm Intelligence algorithms are inspired by animals living together in swarms. Those algorithms are applicable to solve optimization problems like the Travelling Salesman Problem. Furthermore, they can be extended for playing games, e.g., board games. This paper proposes a novel approach for playing the board game Halma by combining Swarm Intelligence algorithms. It focuses on the implementation of a Swarm Intelligence player for the Halma game by combining two state-of-the-art algorithms, namely the Ant Colony Optimization, and the Bee Colony Optimization. In addition, we propose a modular Model View Controller software architecture for implementing the game and its players. Moreover, this paper evaluates the performance of the Swarm Intelligence agent for the single player and two player cooperative version of Halma. The algorithm presented in this paper is successful in learning the dynamics of the game and provides a stable basis for further research in this area.**

***Keywords*—Swarm Intelligence; Traveling Salesman Problem; Ant Colony Optimization; Bee Colony Optimization; Halma.**

I. Introduction

Animals like bees or ants live together in huge swarms. Although the number of members in the swarm is large, they are able to coordinate and divide tasks, e.g., splitting up the food searching process. The tasks are optimized within the swarm. Therefore, it is possible to derive algorithms, called Swarm Intelligence (SI) algorithms, for optimization problems, e.g., the Traveling Salesman Problem (TSP). Our previous work, which compared and applied different SI algorithms to the TSP problem [1], was published at "The Twelfth International Conference on Information, Process, and Knowledge Management eKNOW 2020". As SI algorithms are well suited for optimization problems like the TSP, several applications arise. SI approaches can be used for robotic swarms to explore unknown environments, e.g., in space. Another application for those algorithms are video or board games. This paper extends a previously published paper [1] by implementing and testing SI for a more complex application, namely the board game Halma.

Halma is a traditional board game which can be played using two different types of boards. Either on a square board or, as in our case, on a star-shaped board. Halma on a star-shaped board is also called "Sternhalma" in Germany or "Chinese Checkers" in the rest of the world, although it is not a variant of Checkers [2]. This paper focuses on "Sternhalma", but we will use the name Halma to refer to it. The main similarity between the TSP and the board game Halma is the structure of the problem. The Halma board, shown in Figure 1, is divided into nodes and edges, which is also the basis of the TSP. For a TSP, nodes represent cities which are visited by a salesman. He uses edges to travel from city to city. The aim is to find the optimal solution, so the salesman visits all nodes exactly once and reaches his starting point in the end. For a game of Halma, a player tries to find an optimal solution to get from its starting position to the goal position by using the edges. Due to the similarities and successful tests of SI algorithms for the TSP, it is promising to use SI approaches to construct a player for the Halma game.

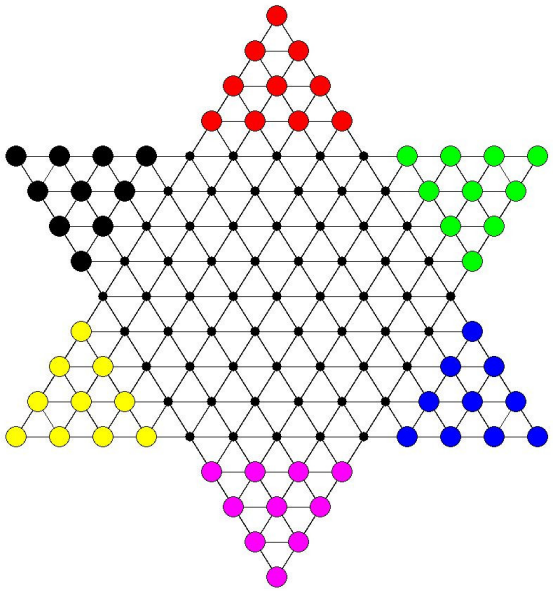


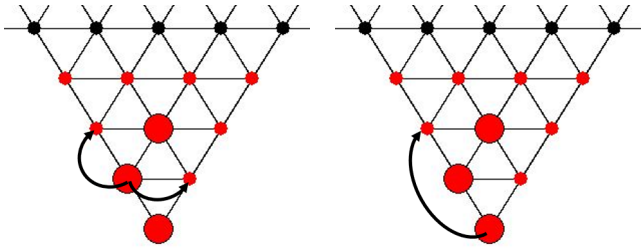
Figure 1. A Halma ("Sternhalma" or "Chinese Checkers") board for six players

The rules of the game are simple. Yet, the game is still challenging and interesting [3]. Its complexity is discussed in Section IV. Halma is played by one to six players having ten game characters each. A board with six players is shown in Figure 1. The board is star-shaped and consists of nodes and edges. Characters are allowed to stand on nodes and move from node to node using the edges. Goal of the game is to bring all game characters to the opposite side of the board.

For example, blue needs to place all its characters on the starting position of black and black needs to place all its game characters on the starting position of blue. To reach this goal, the player is allowed to move one game character at a time. Each character has two possible types of moves which are illustrated in Figure 2. It can either

- make a step move or
- make a jump move.

If it decides to make a step move (Figure 2(a)), the character can be moved in any direction to the next node, requiring the destination node to be unoccupied. If the player decides to jump, a neighboring node needs to be occupied and the node on a direct line behind this node needs to be unoccupied. It is irrelevant if the neighboring node is occupied by an own or an opponent's piece. This move is shown in Figure 2(b). If possible, the player is allowed to do multiple jumps in a row with the same character. He is able to decide how many jumps he wants to do as long as the move is valid.



(a) A character is allowed to make one step in any direction

(b) A character is allowed to jump over another character

Figure 2. The two different kinds of valid moves in Halma

The game ends if one player places all its game characters on the opposite side of its starting position. As Halma is a competitive game, it is a valid strategy to prevent other players from winning.

In general, games are interesting problems in the fields of Artificial Intelligence (AI). The complexity of games is often high and there are several possible strategies. Solving them needs a large amount of time and computational resources. Different games provide different challenges for AI. Most of all, board games are well suited to do research in the fields of AI, as they have simple rules which lead to a simple implementation. Additionally, experiments can be conducted on hardware with less computational power [4].

An AI player can play a game to either win or to gain experience out of it [4]. Halma, like most board games, is a competitive game. It ends when one of the players wins. Nevertheless, for the two player case, the shortest possible solution to win the game can only be reached, if the two players cooperate with each other. Therefore, this paper presents a strategy for a single player, forming a swarm of ten characters, as well as a strategy for two cooperating players.

Additionally, we propose a Model View Controller (MVC) architecture to realize the game. It is designed to easily exchange the AI player with different algorithms.

For accomplishing this, the paper makes use of the results gained in [1]. First, related work is presented. Second, two SI algorithms, the Ant Colony Optimization (ACO) and the Bee Colony Optimization (BCO), are presented. The complexity of the game is discussed in Section IV. The combination of the two SI algorithms to form a Halma player is presented in Section V. Section VI focuses on the experimental setup, our software architecture, and the results gained for a single player game and a two player game. The last section (Section VII) concludes the work and presents future work.

II. Related Work

Both, [5] and [6], give a good overview of the different SI approaches and their analogue in nature. This paper focuses on two out of several SI algorithms, the ACO [7] and the BCO [8]. Recent researches utilize those algorithms for a wide range of applications. Approaches based on the ACO are used, among others, for a routing protocol for Wireless Sensor Networks (WSN) [9], to load balance peer-to-peer networks [10] or a fuzzy logic controller [11]. Furthermore, the ACO has been applied in swarm robotics, e.g., for Unmanned Aerial Vehicles (UAVs) [12], or path planning on mobile robots in [13] and [14]. Variants of the BCO have been employed, e.g., for a swarm of autonomous drones [15] or path planning [16].

An extensive analysis of the game Halma can be found in [17] studying the six-piece game and in [2] focusing on the ten-piece game.

To the authors knowledge solving the Halma game with AI, especially with SI, is not a widely researched area. In [18], the authors design a Halma player based on deep reinforcement learning. Roschke and Sturtevant [19] use an Upper Confidence Bounds (UCB) applied to trees (UCT) algorithm to solve the Halma game. Both papers focus on the two-player game reducing the star-shaped Halma board to the two player square board. Additionally, both approaches are applied to a Halma game with six game characters per player instead of ten, which is the number of pieces used throughout this paper.

There are also only a few approaches using SI algorithms to learn playing games. In [20], the authors use an SI approach for a board game called "Terra Mystica". Kapi et al. [21] consider SI as a method to solve path planning in video games. In [22], the ACO has been used to for a video game called "Lemmings". Daylamani-Zad et al. [23] discuss a variant of BCO and its applicability to strategy games.

Thus, we can see that this paper tackles two ill-researched areas and presents a novel approach in playing Halma with SI based methods.

III. SI Algorithms

In the following, two state-of-the-art SI algorithms are presented. In this paper, the application for SI algorithms is the

board game Halma. The Halma game, as well as the TSP, is based on nodes connected by edges. Consequently, this chapter focuses on the definition of both SI algorithms for discrete problems. The introduction of the algorithms has been adapted from our previous publication [1]. First, the ACO is presented. Second, the BCO is summarized. Third, both algorithms are compared for a TSP application and we evaluate their use for the Halma board game.

A. Ant Colony Optimization (ACO)

When searching for food, ants leave pheromone trails on their path. Other ants can sense the pheromones and plan their path accordingly. This behavior is adapted for the ACO algorithm [24]. Ants mediate information over the environment and communicate therefore indirectly with the other members of the swarm. This form of communication is known as stigmergy communication [25]. Assume a simple example, where an ant can choose between two possible ways to get from the nest to the food source. One path is shorter than the other as visualized in Figure 3.

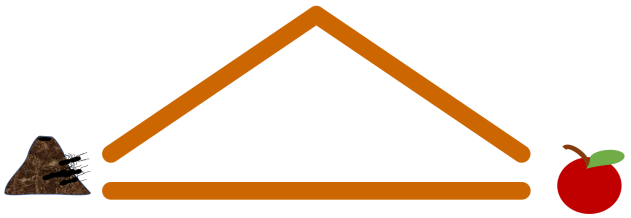


Figure 3. A swarm of ants will choose the short path in favor of the long path to get to the food source over time.

In the beginning, each individual ant chooses its way randomly, i.e., both paths have the same probability to be chosen. The members of the swarm that have chosen the short path will reach the food source earlier than the ants that chose the long path. When they arrive at the food source, they take a piece of food and return to the nest. Now, they have to make the decision once more, which way to chose. As ants leave pheromones on the path, while they move, they sense those pheromones on the short path. If the ants that took the long path have not arrived at the food source yet, the ants sense less pheromones on the long path. Ants choose the path where they can sense more pheromones. Therefore, they will choose the short way to get back to the nest.

To avoid a convergence of the swarm towards local minima, the pheromones on the paths evaporate partly [26]. Nevertheless, the pheromone value on the shorter path is higher than on the longer one. When the ants go to the food source and back to the nest multiple times, the pheromone value on the short path will grow over time. As a result, all ants decide to take the short way in the end [24].

The ACO algorithm simulates this food searching process. In the following, the algorithm is explained for a TSP application. For this application, the path is represented by a sequence of nodes, which are connected by edges. Table I summarizes the symbols used in the equations throughout this section.

The procedure of the ACO is divided into four phases:

1) path planning depending on the pheromone values on the path,
2) pheromone update on each ant's path,
3) pheromone update of the global-best path,
4) pheromone evaporation on all edges.

All phases are iterated multiple times, which is visualized in Figure 4. The next paragraphs focus on the explanation of each phase.

1) Path Planning: Path planning of each individual ant is based on the State Transition Rule

$$s = \begin{cases} \arg\max_{u \in J_{k(r)}} \{[\tau(r,u)] \cdot [\eta(r,u)]^{\beta}\}, \\ \quad \text{if } q \leq q_0 \text{ (exploitation)} \\ S, \quad \text{otherwise (biased exploration)} \end{cases}, \quad (1)$$

where r is the current node of the ant k, s is the next node, and q is calculated randomly [27]. Equation (1) defines the weighting between exploration and exploitation. If q is smaller than or equal to q_0, the ant chooses exploitation. Otherwise it chooses exploration. When choosing exploitation, path planning is based on the value of pheromones on the edge $\tau(r,u)$ and the distance $\eta(r,u)$ between the current node r and a possible next node u. The parameter β regulates the balance between distance and pheromone value. The maximum is chosen from calculating $[\tau(r,u)] \cdot [\eta(r,u)]^{\beta}$ for all nodes that have not been visited yet (for all $u \in J_{k(r)}$) [27].

TABLE I. SYMBOLS USED IN THE FORMULAS EXPLAINED IN SECTION III-A [1]

Symbol Used	Meaning
s	next node
r	current node
u	next possible node
k	ant
$J_{k(r)}$	all nodes that have not been visited yet by ant k
$\tau(r,u)$	pheromone value of an edge between r and u
$\eta(r,u)$	inverse of distance between r and u
β	parameter to manipulate the proportion between distance and pheromone value ($\beta > 0$)
q	random number between [0...1]
q_0	proportion between exploration and exploitation ($0 \leq q_0 \leq 1$)
S	random variable connected to the random-proportional rule
$p_k(r,s)$	probability to choose node s as next node
ρ	pheromone decay parameter for local update ($0 < \rho < 1$)
τ_0	initial pheromone value
α	pheromone decay parameter for global update ($0 < \alpha < 1$)
δ	evaporation parameter ($0 < \delta < 1$)

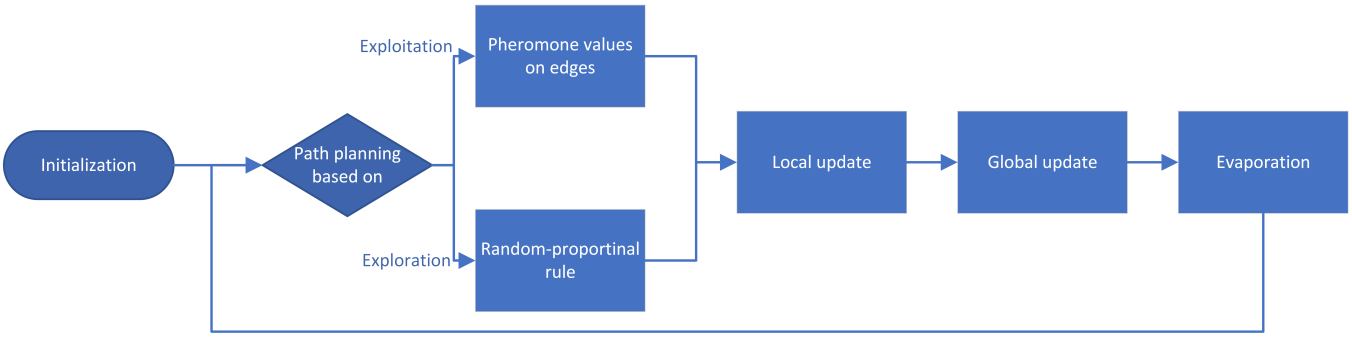


Figure 4. Steps of the ACO algorithm [1]

For biased exploration, the node selection is made with the random-proportional rule

$$S = p_k(r,s) = \begin{cases} \frac{\tau(r,s)\cdot\eta(r,s)^\beta}{\sum_{u\in J_{k(r)}} \tau(r,u)\cdot\eta(r,u)^\beta} & \text{if } s \in J_{k(r)} \\ 0, & \text{otherwise} \end{cases}, \quad (2)$$

where S represents the result of this random-proportional rule [27]. Equation (2) calculates a probability for each node. It is based on the length of an edge and its pheromone value. The shorter the edge and the higher the pheromone value, the more likely this node is chosen. As both, the pheromone value and the length of an edge are considered, $[\tau(r,u)] \cdot [\eta(r,u)]^\beta$ is calculated as well [27]. This results in a weighted value and the exploration is referred to as biased exploration [7]. The term is divided by the sum of all $[\tau(r,u)] \cdot [\eta(r,u)]^\beta$ to calculate the probability.

Each ant leaves a pheromone trail on its path. The trail is updated after each ant has finished its tour and has returned to the initial node. This pheromone update is called the local update.

2) Local Update Rule: Real ants leave pheromones on their trail. The pheromone values increase with the quality and the quantity of the food at the food source [24]. In analogy to real ants, the pheromone values on a path constructed by the artificial ants are updated with

$$\tau(r,s) = \tau(r,s) + \rho \cdot \Delta\tau(r,s)\,, \quad (3)$$

where $0 \leq \rho \leq 1$ [27]. The local update rule is influenced by $\Delta\tau(r,s)$. Depending on the implementation, you can choose different approaches to set $\Delta\tau(r,s)$ [27]. There are implementations which use, e.g., reinforcement learning to determine an appropriate $\Delta\tau(r,s)$ [27]. For sake of simplicity, we use a constant value

$$\Delta\tau(r,s) = \tau_0\,, \quad (4)$$

where τ_0 corresponds to the initial pheromone value.

3) Global Update Rule: After the local update has been performed and all ants have returned to the nest, the global update is conducted. All paths that have been found by the individual ants are compared to the global best path. For the TSP this is the path with the shortest length. For the Halma game this can, e.g., be a move that brings the game character closest to the goal positions or a move that uses many jumps. If one of the ants found a better path, the global best path is updated. No matter if it has been updated in the current iteration or not, the pheromone values on the edges are updated for the globally best path. Extra pheromones are added with

$$\tau(r,s) = \tau(r,s) + \alpha \cdot \Delta\tau(r,s)\,, \quad (5)$$

where α is a predefined pheromone decay parameter between 0 and 1 [27].

By adding pheromones to the edges in each iteration, the pheromone values on the edges increase over time. When a good solution was found in, e.g., iteration 10 and the swarm does not find a better solution during the following iterations, the pheromone value on this path is high. If an ant finds a better solution, it takes many iterations until the swarm will use this solution, as the pheromone value on the old best path is still high. Therefore, it is difficult to abandon old solutions. Consequently, we have to introduce evaporation to overcome this issue.

4) Evaporation: To avoid rapid convergence towards a non-optimal solution, pheromone values evaporate partly when they are updated. It offers the possibility to explore new areas [24]. The evaporation is regulated with a parameter δ and results in

$$\tau(r,s) = \delta \cdot \tau(r,s)\,. \quad (6)$$

We can also combine the evaporation with the local and global update rule [27]. The global update rule is now calculated with

$$\tau(r,s) = (1-\alpha) \cdot \tau(r,s) + \alpha \cdot \Delta\tau(r,s)\,, \quad (7)$$

whereas the local update is computed with

$$\tau(r,s) = (1-\rho) \cdot \tau(r,s) + \rho \cdot \Delta\tau(r,s)\,. \quad (8)$$

Combining all steps, we result in the ACO as summarized in Figure 4.

The second algorithm, our SI Halma player is based on, is the BCO, which is presented in the following section.

B. Bee Colony Optimization (BCO)

Another state-of-the-art SI algorithm is the BCO. It is derived from the foraging behavior of bees [28]. In contrast to ants, bees use a different kind of communication. Ants communicate indirectly over the environment by leaving pheromone trails. On the contrary, bees communicate directly with the other swarm members by means of dancing. Bees fan out the hive to search for food, return and dance to the other bees to communicate the location of a food source. The dance is a form of advertisement to convince the other bees to choose the food source they are advertising [28].

The BCO was inspired by this behavior to solve optimization problems, e.g., the TSP. The algorithm mainly consists of two steps,

1) the forward pass,
2) and the backward pass.

The algorithm is visualized more detailed in Figure 5. Like the ACO, the BCO uses multiple iterations to converge towards a solution. In contrast to the ACO, which constructs a path at once, the BCO is divided into several stages. During each stage, a bee conducts the forward, as well as the backward pass and builds a partial solution, i.e., a part of the path. The number of stages for the TSP application depends on the number of nodes m the path is appended by during each stage. The two steps, conducted during each stage, are explained in the following.

1) Forward Pass: The forward pass is the step of building a partial solution. Each bee fans out the hive and appends its path by its own partial solution. This represents exploration, as the partial solutions are calculated randomly. For the TSP the bees append their current path by m nodes they have not visited yet [8]. After appending the path and returning to the hive, the bees perform the backward pass.

2) Backward Pass: The backward pass is the phase, where the bees perform the waggle dance. Each bee has two options to either

- abandon its partial solution (exploitation) or
- dance and advertise its solution to the other bees (exploration) [28].

By abandoning its solution, the bee will exploit another bee's solution (one of the bees that dances) or the global best solution (the best solution that has been found so far). For the TSP application, the shorter the path length of another bee's partial solution, the more likely the bee will choose it. After choosing a partial solution the bee will add this partial solution to its path that has been constructed during the previous stages. It basically exchanges its partial solution constructed during the forward pass with another partial solution [8]. For the TSP, cities are visited only once. Consequently, it is crucial to check whether the chosen partial solution does contain cities that have already been added to the path. If that is the case, for our implementation, the honeybee returns to its own partial solution constructed during the forward pass.

A solution for the problem has been found if the forward and the backward pass have been finished for all stages. Subsequently, the global best solution is updated [28] and one iteration has been finished.

The following section focuses on the comparison of the two algorithms and their application for the TSP as well as for the Halma game.

C. Comparison of the Algorithms

In order to decide which of the algorithms to use for the Halma board game, we first tested them for the TSP. The TSP was simulated by placing ten nodes, representing the cities, randomly on a grid. The edges, connecting all nodes, have different length. The algorithms are supposed to find a path which connects all nodes while traveling a minimum distance. Each algorithm performed 200 iterations per test and 1000 tests have been conducted. The results shown in Figure 6 give an idea of the performance of the algorithms. As the TSP is only an exemplary application to compare the algorithms, the implementations are not optimized with respect to time efficiency and performance. Additionally, this paper focuses on a board game application and the goal is to play against a human player, so time efficiency can be neglected. For the evaluation of performance of each algorithm, the interested reader is referred to [29]. Figure 6 visualizes the average path length for each iteration. The ACO is visualized in blue, whereas the BCO is shown in red. The parameter configurations used for the experiments visualized in Figure 6 are summarized in Table II.

TABLE II. PARAMETERS USED FOR EXPERIMENTS [1]

ACO		BCO	
parameter	**value**	**parameter**	**value**
iterations	200	iterations	200
population	100	population	100
β	0.7	m	3
q_0	0.8		
ρ	0.7		
τ_0	10.0		
α	0.9		

The ACO converges towards the optimal solution, which is shown on the bottom in black, whereas the BCO converges quickly but towards a non-optimal path. Consequently, the ACO has a better balance between exploration and exploitation than the BCO. For the BCO, exploitation predominates over exploration. From Table II we can see that the BCO needs less parameters than the ACO that need to be tuned. Nevertheless, the balance between exploration and exploitation is much better for the ACO.

The ACO is well suited for the TSP problem, as the swarm contributes from all solutions of all other ants and does not only consider the global best solution found at some point in the past. Distributing pheromones on the edges makes an easy planning possible. If we decide to us the ACO on its own for a Halma player, the algorithm outputs us 10 different paths

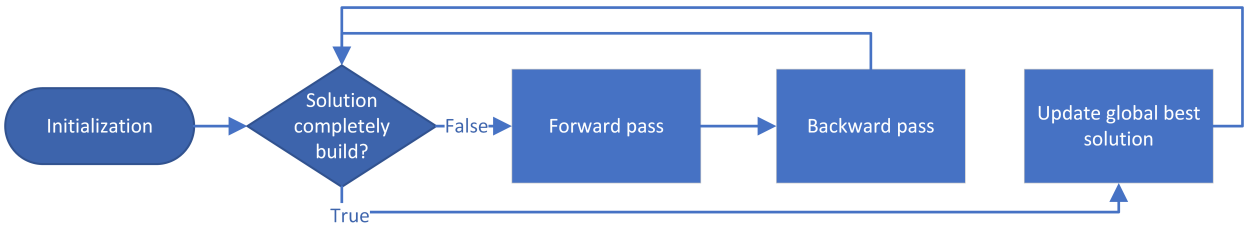


Figure 5. Steps of the BCO algorithm [1]

for 10 different characters. Then, we have to decide which game character executes its path, as the rules only allow one move of one game character at a time. With its abandoning and advertising scheme, the BCO is well suited to find the best solution within a swarm and to decide for one game character to move. Therefore, the ACO and the BCO are combined for the board game Halma. All game characters first plan a path with the ACO algorithm and then the BCO is used to decide which character will make the move. The implementation and combination of the two algorithms is further explained in Section V.

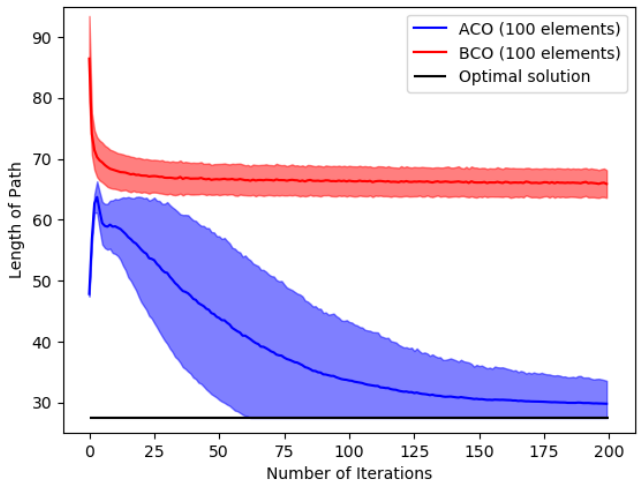


Figure 6. Average length of path by solving the TSP with the ACO (blue), the BCO (red) and the optimal solution (black) [1]

IV. Complexity of Halma

Even though its rules are simple, the board game Halma offers a high complexity. In [17], Sturtevant mentions that there are $1.73 \cdot 10^{24}$ states referring to a game with six game characters per player. The number of states is even higher for the ten-piece version of the game.

Due to the different possible moves, i.e., steps and jumps, and a varying number of players per game, and therefore also a varying number of game characters per game, there are usually a lot of different options per turn a player can choose from. Furthermore, players are allowed to perform multiple jumps in one turn with the same figure, if possible. Successive jumps do not necessarily have to follow the same direction. This contributes to even more different turns one can perform. As players are allowed to jump over the characters of other players, more players in a game allow a higher number of possible moves. In general, it is advantageous to jump as often as possible as this allows extra moves per turn. To reach this, the player needs to build ladders which transport characters quickly from one side to the other [2].

When playing Halma with only one player, the shortest solutions are often palindromic. This means, that the second half of moves is symmetric to the first half. According to [2] the shortest solution possible for one player has 27 moves and no shorter solution is possible.

For a game of two players the shortest game consists of 30 moves, that is 15 moves per player [2]. However, this only works if both players cooperate. In this scenario, two ladders are built. The first one is built by both players whereas the second one is only built by the player that is going to lose.

The following chapter considers this information and presents an implementation for a SI player for Halma using the algorithms introduced in Section III.

V. Combination of SI Algorithms for Halma

To use Swarm Intelligence for playing a game of Halma, the algorithms mentioned in Section III are combined. Figure 7 gives an overview of the implementation of an SI player for Halma. Table III includes the symbols used in equations throughout this chapter that have not been already introduced in Table I.

The ACO algorithm is well suited for the local path planning of each character. A single character is able to plan a path based on the pheromone values on the edges between the nodes. It decides whether to choose exploration or exploitation. The path is planned accordingly. In Figure 7 this process is marked in orange. The implementation of the BCO is illustrated in green. Originally, the forward pass of the BCO is used to make a local decision for each member of the swarm. In our algorithm, this decision is based on the ACO, so the forward pass is neglected, whereas the backward pass plays a major role. It is used to make the decision which character is going to make a move. The characters can either abandon their choice or advertise their solution to the others. This decision making process is further discussed in Subsection V-C.

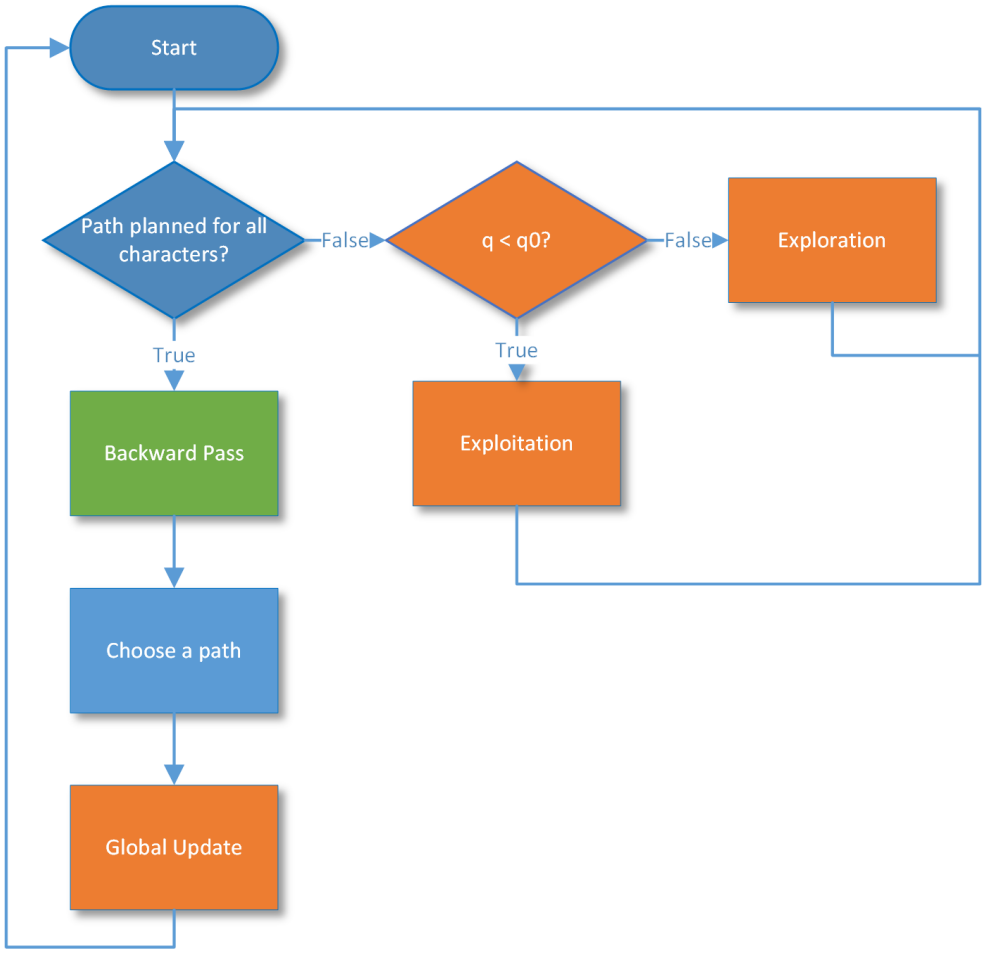


Figure 7. Process of choosing a character and path to make a move

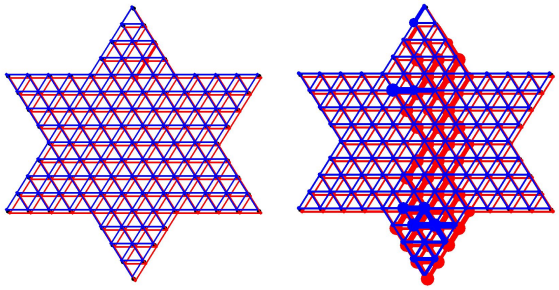
TABLE III. SYMBOLS USED IN THE FORMULAS EXPLAINED IN SECTION V THAT ARE NOT CONTAINED IN TABLE I.

Symbol Used	Meaning
$\tau 0_{goals}$	initial pheromone value for the two edges connected to the first row
j	jumping factor to reward moves with many jumps
g	in goal factor to increase the pheromones on edges that lead to the goal
e	evaporation rate
w_d	weight of distance to the first row
w_l	weight of the length of the path
w_p	weight of the pheromone values of the path
f_g	fitness value for characters that already reached the goal area
b	influence of best game played so far on the initial pheromone distribution
c	pheromone update value added during cooperative games
t_o	threshold to update the pheromone value during cooperative games

In the beginning, the pheromone values are distributed on all edges. This initialization is described in Subsection V-A. Subsection V-B discusses how the ACO can be applied to Halma. The global backward pass and the update rule, including evaporation, are further explained in Subsection V-C and V-D, respectively. Subsection V-E introduces some limitations for the pieces. The last subsection of this section focuses on the two-player cooperation.

A. Initial Distribution of Pheromones

The Halma board is divided into nodes and edges. The game characters need to move from node to node by using the edges. All edges are bidirectional and it is possible to use the same edge to move from node A to node B and from B to A. Therefore, it is necessary to distinguish which direction of the edge the agent moves along in order to judge if it is approaching the goal. As a consequence, it is necessary to double all edges for the implementation of the ACO. As a result, there are two edges connecting two nodes A and B, one to move from A to B and one to move from B to A. Each of the doubled edges has its own pheromone value. This value differs if the edge between A and B or the edge between B and A is examined. As each edge is doubled to have both directions, in Figure 8, one direction is marked in blue and the other in red.



(a) Pheromone distribution if all edges have the same starting pheromone value

(b) Pheromone distribution if the best solution found in a previous game is considered

Figure 8. Different pheromone distribution modes

Tests have shown that distributing the pheromones equally on all edges with

$$\tau(r,s) = \tau_0\,, \tag{9}$$

as visualized in Figure 8(a), the swarm needs a lot of time to reach the goal. Therefore, we introduced a second pheromone initialization mode. The pheromone values are initialized depending on their distance to the goal with

$$\tau(r,s) = \frac{1}{d} \cdot \tau_{0_{goals}} + \tau_0\,, \tag{10}$$

where d is the distance to the last goal position (the first row in Figure 9(a)), and $\tau_{0_{goals}}$ is a parameter for the initial pheromone value for the two edges connected to the first row. Edges connecting goal nodes have a higher initial pheromone value than edges connecting the start nodes. As a result, the swarm knows the direction where to go from the beginning. Additionally, we gave the swarm the ability to remember pheromone values of previous games. The pheromone distribution of the best game with the lowest number of moves played so far is stored. Its influence on the initial distribution of the pheromones can be weighted with the parameter b. Figure 8(b) illustrates the pheromone distribution for considering the best solution found so far. It can be exploited by the characters in the beginning of the new game.

After distributing the pheromones on all edges, the characters start planning their path by using the ACO algorithm.

B. Implementation of the ACO for Halma

Each game character needs to choose between exploration or exploitation. The parameter q_0 defines the probability to choose either of them. In the following exploration and exploitation are examined separately.

1) Exploitation: The character chooses exploitation if the randomly generated variable $q < q_0$. Exploitation means that the path is generated by using the trail with the highest pheromone value. Therefore, a character calculates all its valid moves. For each neighboring node of the character's position, all valid step and jump moves are calculated. To choose one of those paths, each path is evaluated with

$$p = l \cdot \sum_{e}^{k} \tau(r,s), \tag{11}$$

where l is the length of the path, $\tau(r,s)$ the pheromone value of an edge, and k the number of edges in the path. In contrast to the TSP application, for Halma we prefer longer paths over short moves. Long paths mean that we found several characters to jump over. Making several jump moves at once is crucial to find a efficient way to approach the goal. We therefore choose the path with the highest p. The chosen path is then considered for the global decision, which character is going to move.

2) Exploration: If the character chooses exploration, the general procedure is similar to exploitation. In contrast to the standard ACO the random-proportional rule was replaced, so the edges are chosen completely random without any bias. The edges of the character's path are chosen randomly from all neighboring edges that enable a valid move. The exploration procedure also implements both, jump moves and step moves.

In contrast to exploitation, exploration only produces one path and no decision is needed to choose between potential paths.

3) Balance Between Exploration and Exploitation: The balance between exploration and exploitation is important to generate new solutions as well as to exploit good solutions. For exploration, choosing edges randomly leads to new paths to visit nodes and edges which have not been part of a path so far. With exploration it is possible to find paths to the goal which lead to fewer draws than the already found solutions. Those paths found by exploration can be used by other game characters throughout exploitation. They follow the paths that have already been chosen by other characters. The more characters follow a path the higher is the pheromone value on these paths. This results in the exploitation of good explored solutions.

With the ACO each character plans the path it would take if it is going to move. As only one character is allowed to move at a time, the global backward pass of the BCO is used to make this decision.

C. Global Backward Pass

For performing the backward pass, each character calculates a fitness for its path. This fitness is calculated with

$$f = w_d \cdot \frac{1}{d} + w_l \cdot l + w_p \cdot p\,. \tag{12}$$

To calculate the fitness, three components are taken into account:

- the distance d between the end node of the path and the first row (farthest goal position)
- the length of the path (l)
- the sum of all pheromone values on all edges of the path (p)

Each component has its own weighting factor w_d, w_l, w_p respectively. Those weighting factors are given as parameters. If a character is already at one of the goal positions the equation for calculating the fitness is changed to

$$f = w_d \cdot \frac{1}{f_g} + w_l \cdot l_p + w_p \cdot p\,. \tag{13}$$

This avoids draws inside the goal, although there are still characters at the starting position or in the middle of the field. Here, the distance d was replaced by a factor f_g that is associated with a parameter.

After calculating the fitness of each character, the sum of all fitness values is calculated. Then, the decision is made if a piece abandons its path or if it presents its solution to the others. The decision is based on a probability $\tau(r, s)$ for a character i which is calculated with the roulette wheel rule according to

$$\tau(r, s) = \frac{f_i}{\sum\limits_{c \in C} f_c}. \tag{14}$$

The fitness f_i of the character i is divided by the sum of all fitness values of all characters C.

The probability $\tau(r, s)$ defines if the character abandons or advertises its solution. Is $\tau(r, s) > \frac{1}{|C|}$, the character advertises its path. Otherwise, it abandons it.

Each character that decides to abandon its solution, needs to choose one of the solutions advertised by another character. Therefore, a similar equation to Equation (14) is used, but only the fitness values of characters are summed up that advertise their path.

One character is chosen from all characters that advertise their solution by an Equation similar to (14) and makes a move.

D. Update Rule

After a character has been chosen to make a move, the edges of its path are updated. In contrast to the original ACO algorithm, the update rule is only used on a global level and the local update rule has been neglected. We use only the best result of the paths proposed by our ten characters. The paths of the other nine characters may be worse and would negatively influence the pheromone distribution. Therefore, when testing the local update rule, their results have been worse and we decided to neglect it. The pheromone update is illustrated with

$$\tau(r, s) = \tau(r, s) + \alpha \cdot l \cdot \tau_0 \cdot \begin{cases} 2 \cdot j, & \text{jump moves and } l > 2 \\ 1 \cdot j, & \text{jump moves} \\ 1, & \text{else} \end{cases}. \tag{15}$$

A pheromone value, depending on the properties of the path, is added to the pheromone value of an edge $\tau(r, s)$. If the character makes a move without jumping, a pheromone value of $\alpha \cdot l \cdot \tau_0$ is added, where l is the length of the path. If the character's move includes one or more jumps, a jumping factor j, given as a parameter, is taken into account. If the character jumps more than once, this factor is multiplied by two. This has the effect that a long path raises the pheromone value of the edges significantly and other characters might consider this path as well in the next iteration.

If a character reaches one of the goal positions, the pheromone values on all edges the character has been visiting throughout the game are updated. This increases the pheromone values on edges which lead to the goal. Other characters are able to plan their path accordingly. Consequently, after one character reaches a goal position, the probability for the other characters increases to get to the goal faster. The edges are updated with

$$\tau(r, s) = \tau(r, s) + \rho \cdot \tau_0 \cdot g\,. \tag{16}$$

The initial pheromone value τ_0 is multiplied by an in goal factor g and by ρ which are both given as a parameter. The result is then added to the edge's pheromone value.

To avoid the algorithm to get stuck into a local minimum the pheromone values on all edges evaporate partly. The pheromones only evaporate after a predefined number of moves (e), to avoid an evaporation of the pheromones on long paths too quickly. The updated pheromone value of an edge is calculated with

$$\tau(r, s) = (1 - \alpha) \cdot \tau(r, s)\,. \tag{17}$$

E. Limitations for the Characters

The characters plan their path autonomously. After doing first tests a few problems were detected that needed to be restricted. First of all, the characters are supposed to stay in the goal area once they arrive there. Figure 9(a) illustrates the goal positions of one player in red. If a character reaches one of the red points, it is in the goal region.

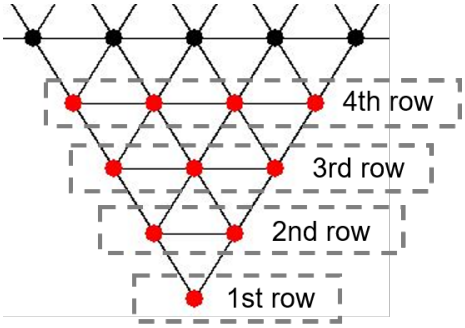
To avoid that the characters do small moves in front and inside the goal, they need to fill the goal from the end. This is shown in Figure 9(b). If a character reaches the first row, it is not allowed to move anymore. The second row needs then to be filled next. The character on the left of the two nodes is not able to move. For the characters that are not allowed to move, no local path is calculated and they are not part of the decision making process. If the character in the Figure 9(b) moves to the node where the arrow is pointing to, it is not allowed to move in future moves. The list of nodes that will be filled next is updated so the third row is filled during the next draws.

F. Cooperative Game

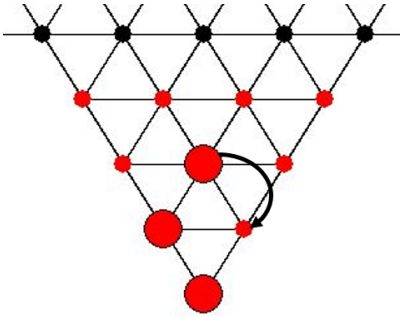
As mentioned in [2], an optimal solution for the two-player case can only be reached if the two players cooperate. Consequently, we implemented an option to start a cooperative two-player game. After each move, the two players exchange their pheromone distribution. As both players start at the opposite of each other, the pheromone distributions are then inverted and compared to the own distribution. The pheromone values are then updated according to

$$\tau(r,s) = \tau(r,s) + \begin{cases} c, & \text{if} \tau(s,r)_o > t_o \\ 0, & \text{otherwise} \end{cases}, \qquad (18)$$

where $\tau(s,r)_o$ is the pheromone value of the other agent matching $\tau(r,s)$ and c and t_o are given as a parameter.



(a) Visualization of the different rows in the goal positions



(b) The goal is filled from the end

Figure 9. Goal positions of one player

The implementation proposed throughout this section has been tested for playing Halma games with SI. The experimental results are presented in the following chapter.

VI. Experiments

This section focuses on the experiments conducted for the SI player. First, the experimental setup is defined and the architecture of our implementation is proposed. Second, the experimental results for single as well as two-player games are presented.

A. Experimental Design

In order to optimize the algorithm designed for Halma we implemented a framework for Halma. As depicted in Fig. 10 the application is based on a Model View Controller (MVC).

Due to the modularization of the different components, we can connect both, an agent based on the presented SI algorithm but also an agent that uses any other algorithm. This has the following advantages:

- It allows future research with algorithms of different archetypes. We could, e.g., include agents based on reinforcement learning by solely exchanging the agent.
- A learning algorithm can play against a human player and learn from their strategy.
- It allows the initialization of the presented SI algorithm with different hyperparameters. This can lead to one algorithm playing ”defensive”, i.e., building up a ladder and the other algorithm playing ”offensive”, i.e., making use of the ladder to reach the other side more quickly. This could greatly benefit the optimization of our solutions.

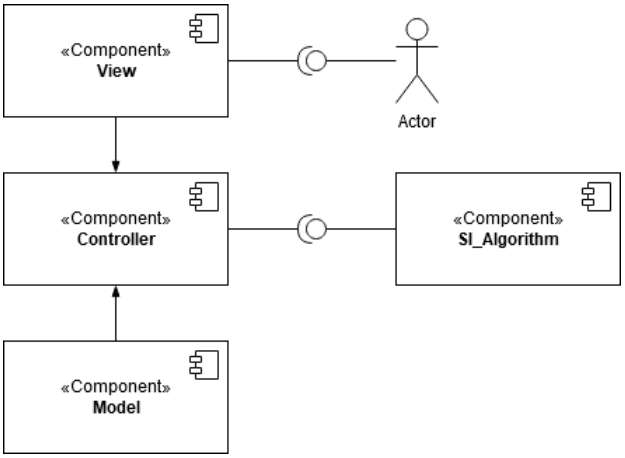


Figure 10. Component Diagram of the Halma framework

To test the SI player for a Halma game, we conducted experiments for the single player and two-player game. For both setups, the parameter configurations are the same and summarized in Table IV.

TABLE IV. Parameters used for playing Halma with SI

Parameter	Value
τ_0	5
$\tau_{0_{goals}}$	10
q_0	0.8
β	2.7
α	0.1
ρ	0.5
j	5000
g	10
e	11
w_d	2
w_l	50
w_p	5
f_g	50
b	2.5
c	5
t_o	20

We want to force the players to favor jump moves over step moves because they contribute to less draws. Therefore, the parameter value for j and w_l are high. Additionally, the large value for f_g reduces the number of draws within the goal region. If a game has not been finished after 500 moves, the game has been stopped and restarted.

Goal of the experiment with one agent is the comparison of number of moves a single SI player needs to win the game, in comparison to the best possible solution as discussed in Section IV. Additionally, we want to evaluate the effect of the initial distribution of pheromones on the number of moves needed to finish a game. Moreover, we conducted two-player experiments to evaluate the cooperation between two agents.

B. Experimental Results

In the following, we present the results gained for single, as well as two-player experiments. All configurations and the evaluation of number of moves are summarized in Table V and Figure 11.

1) Single player experiments: Figure 12 shows the number of moves of three test scenarios for 1000 games each. The more bold a point, the more often this number of draws has been achieved. The blue points on the bottom represent the number of moves without using the best solution found so far in a previous game (Id 1). Nevertheless, it uses an initial distribution of the pheromones introducing the direction where the swarm has to go. The orange points in the model represent the number of moves when using the best solution found so far with 68 moves (Id 2). The game with 68 moves has been the game with the smallest number of draws found when doing experiments. It has been achieved in a previous game. The pheromone distribution at the end of this game has been used to initialize the pheromones. The green points on the top represent the total number of moves when starting without the best solution found so far and updating it throughout the game, i.e., the first game of the 1000 played games was the best solution in the beginning and was updated during the experiment (Id 3). Figure 12 shows that using the best solution that was found by the swarm in the past has an effect on the number of draws. Using the best solution found so far reduces the median of the number of moves. Without using best result of the past, the median is 133. In comparison, when starting without a best solution and updating it throughout the experiment, the median is 120. When initializing the pheromone values considering the best solution with 68 moves, the median of total moves is 113. Furthermore, from Table V, the mean, as well as the standard deviation is lowest when using the best result found so far with 68 moves.

As a consequence, the initial distribution of pheromones has a large effect on the number of moves needed by the swarm. Without a hint in which direction to move, the swarm needs more moves to win the game.

Furthermore, we tested the SI player for different balance values between exploration and exploitation shown in Figure 13. In the aforementioned experiments, the parameter q_0 balancing between exploration and exploitation is 0.8. For the following test, we choose an initial pheromone distribution exploiting the best solution of 68 moves. The experiment has been conducted for 1000 games each by varying the q_0 parameter. Three values for q_0 have been tested, namely 0.6 (blue on the bottom), 0.8 (orange in the middle), and 1.0 (green on the top).

As 68 moves was the best result that has been reached so far with the algorithm, exploiting this solution more, results in a lower total number of moves than when increasing exploration. The SI player can exploit the pheromone values of the 68-moves solution, because they lead to a good result in the past. As seen in Figure 8(b), when using the best solution for pheromone initialization, the pheromones are mostly distributed on the area directly connecting the start and the goal. More exploring will therefore lead to worse results, as no new faster ways besides the direct paths to the goal can be found. Consequently, choosing $q_0 = 1.0$ leads to better results than $q_0 = 0.6$. As stated in [2], the 68 draws solution is still far from optimal. In order to improve this, we need exploration and therefore, we chose for the following experiments $q_0 = 0.8$.

2) Two-player Experiments: As mentioned in [2], the perfect game with two players consists of less moves per player than single player games. In order to validate if that is also the case for our SI Halma player, we present the results of several experiments in the following.

First, we directly compare a single player game with a two-player game using the same parameters. For this experiment, the players do not cooperate while playing. The results are presented in Figure 14. 1000 tests have been conducted for the single player scenario on the bottom (blue, Id 2), whereas the two-player scenario on the top (orange, Id 5) has been repeated 500 times. Comparing the median of the number of moves for both experiments, the two-player scenario outperforms the single player setup with 104 compared to 113 moves.

The following experiment focuses on the two-player setup. Four scenarios with 500 games each have been compared and are shown in Figure 15. If the best solution found so far has been used, the 68 moves solution already used for previous experiments has been chosen. The results for using the best solution and having a cooperation between the two players (Id 4) is shown in blue on the bottom. Using the best solution, but not having a cooperation is shown in orange (second from bottom, Id 5). For the other two results, no best solution has been used. For the green results (second from top, Id 6), the players have cooperated, wheres for the tests resulting in the red dots (top, Id 7), no cooperation has been introduced.

Both using the best result and the cooperation affect the median of the number of moves as visualized in Figure 11. The median of the number of moves when using both is 102, whereas it is 121.5 when not using both. Introducing a cooperation when not using the best solution improves the median from 121.5 to 114. When using the best solution found so far, the cooperation slightly improves the median of moves from 104 to 102. As visible in Figure 15, the standard deviation when not using the best solution is much higher than when relying on it. This is also proven by the standard deviations shown in Table V. Despite the fact, when using the best solution, the median of moves for cooperation is lower, the standard deviation is higher. Although cooperation can

TABLE V. RESULTS FOR THE NUMBER OF MOVES FOR DIFFERENT EXPERIMENTAL SCENARIOS.

ID	Players	Tests	Best Solution	Cooperation	Median	Mean	Standard Deviation
1	1	1000	x	-	133	153.46	63.27
2	1	1000	✓	-	113	116.82	33.42
3	1	1000	construct	-	120	132.81	39.2
4	2	500	✓	✓	102	118.35	58.45
5	2	500	✓	x	104	121.75	54.96
6	2	500	x	✓	114	142	73.32
7	2	500	x	x	121.5	152.91	81.4

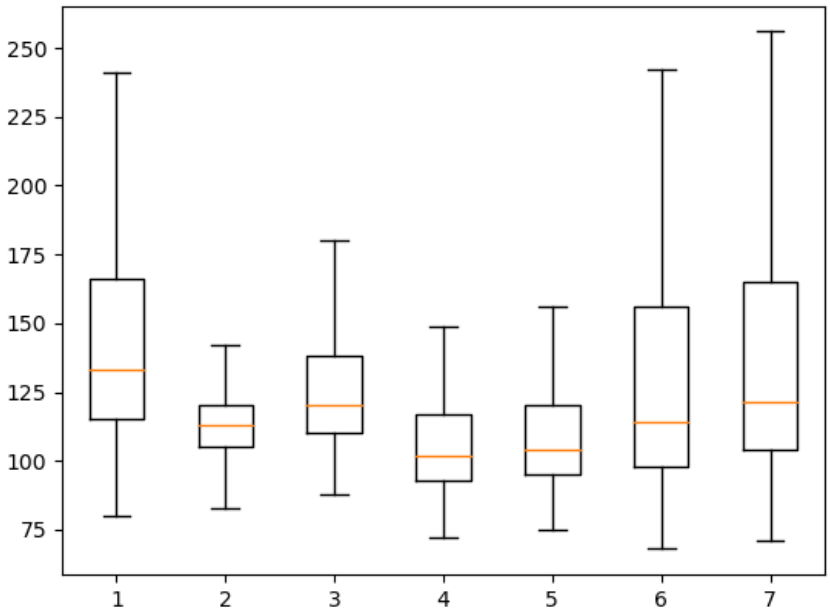


Figure 11. Distribution of draws until a player wins the game for the scenarios in Table V identified by the ID.

improve the number of moves, the solutions are far from the optimal solution of 15 moves per player specified in [2].

In general we can see from Table V and Figure 11 that we were able to improve the performance of the algorithm by remembering the best solution found so far. In the single, as well as the two-player scenarios, the number of total draws is less when using the best solution found so far than when the initial pheromone distribution is only based on the distance to the goal. Additionally with this approach, it was possible to reduce the standard deviation significantly. We can see that our algorithm is able to solve the game and our modifications including a ”memory” increased its performance. Although, the results for the single as well as the two-player case are far from optimal, further modifications and optimization can help to reach the optimal solutions.

VII. CONCLUSION AND FUTURE WORK

This paper presented an application for using a combination of SI algorithms. The ACO algorithm and the BCO algorithm have been combined to realize a nonhuman player for the board game Halma. Experiments have shown that the initial distribution of pheromones has a big influence on the performance of the SI player.

Nevertheless, the number of moves resulting from the experiments is still high in comparison to the optimal solution. Future work will therefore focus on decreasing the number of moves needed to win a game with one and multiple players. Furthermore, it is possible to compare the SI player to a human player. Most humans will not find the optimal solution when playing Halma. Therefore, it is interesting to do experiments by comparing human players and SI players.

In future work, we want to make use of this architecture to find out about the performance of other algorithms searching

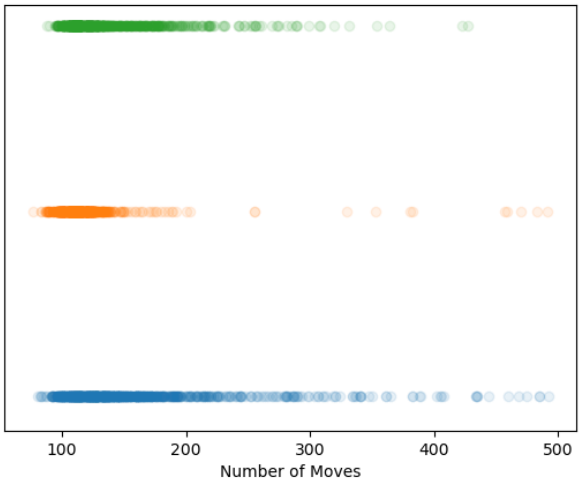


Figure 12. Number of moves for 1000 games. Blue (bottom) without using the best solution found in other games (Id 1). Orange (middle) using the best solution found so far with 68 moves (Id 2). Green (top) starting without a best solution and updating it whenever a better result has been found (Id 3).

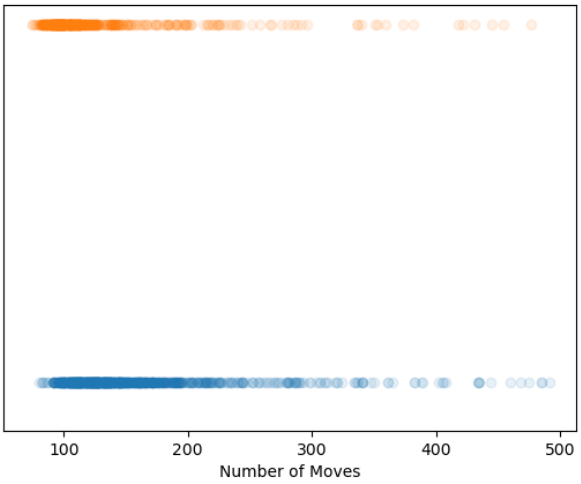


Figure 14. Number of moves for one player using the best solution found so far with 68 moves. 1000 games for a single player game (blue bottom, Id 2). 500 tests for two players without cooperation (orange top, Id 5).

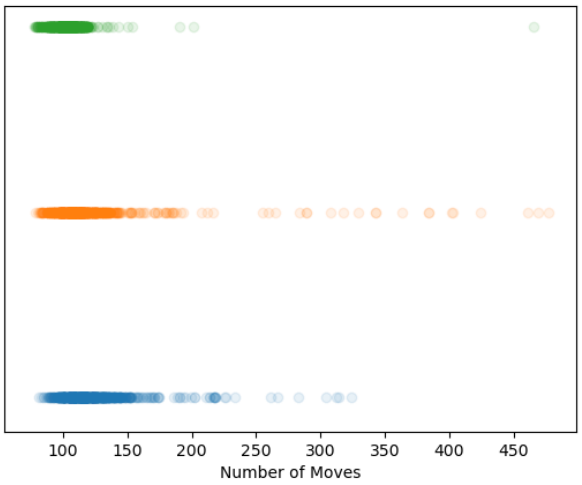


Figure 13. Number of moves for 1000 games. Blue (bottom) with $q_0 = 0.6$. Orange (middle) with $q_0 = 0.8$. Green (top) with $q_0 = 1.0$.

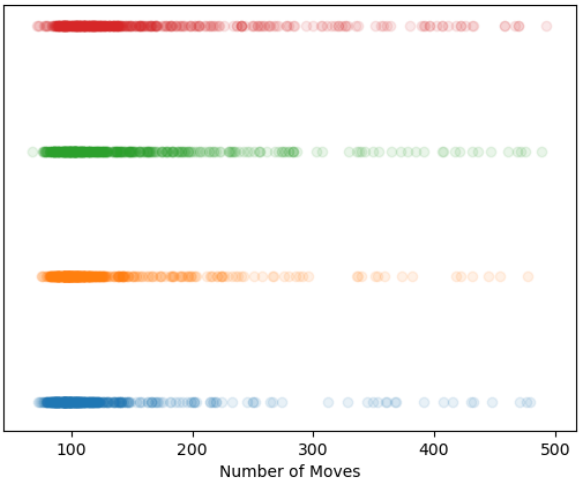


Figure 15. 500 Tests per experiment. Blue (bottom, Id 4) using the best solution (68 moves) and cooperation of the agents. Orange (second from bottom, Id 5) using the best solution (68 moves) and no cooperation. Green (second from top, Id 6) not using the best solution, but having a cooperation of the agents. Red (top, Id 7) no using the best solution and not cooperating.

for the shortest game possible. Furthermore, the architecture allows the initialization of the presented SI algorithm with different hyperparameters. This can lead to one algorithm building up a ladder and another exploiting the ladder to reach the other side more quickly. This could greatly benefit the optimization of our solutions.

This paper focuses only on one board game. SI can also be used for other board games with multiple characters. Another possible application would be Chess. In contrast to Halma, where all characters are equal, in Chess the members of the swarm have different roles. The decisions of the swarm need to consider the inequality of its members. SI can not only be used for board games, but for every game were multiple characters play together and need to make decisions. In video games, the human player often needs to play against other players and characters. If the game involves armies of opposing players, they can also act like a swarm. They need to find solutions themselves by making decisions which consider the solution of every single member of the swarm. Video games are more complex than board games and the effort for implementing an SI algorithm for a video game is higher than for a board game, but it enables a swarm-like behavior of the opposing player.

Not only simulated swarms in games can be optimized,

but the approaches can also be expanded for path planning in multi-agent systems and robotic swarms.

Acknowledgement

This paper has been written during a cooperative study program at the Baden-Wuerttemberg Cooperative State University Mannheim and the German Aerospace Center (DLR) Oberpfaffenhofen at the Institute of Communications and Navigation.

References

[1] I. Kuehner, "Swarm intelligence for solving a traveling salesman problem," in *eKNOW2020, The Twelfth International Conference on Information, Process, and Knowledge Management*. IARIA, 2020, pp. 20–27.

[2] G. I. Bell, "The shortest game of chinese checkers and related problems," *Integers*, vol. 9, no. 1, pp. 17–39, 2009.

[3] A. Elithorn and A. Telford, "Game and problem structure in relation to the study of human and artificial intelligence," *Nature*, vol. 227, no. 5264, p. 1205, 1970.

[4] G. N. Yannakakis and J. Togelius, *Artificial intelligence and games*. Springer, 2018, vol. 2.

[5] C. Blum and D. Merkle, *Swarm intelligence: introduction and applications*. Springer Science & Business Media, 2008.

[6] J. Kennedy, R. C. Eberhart, and Y. Shi, *Swarm intelligence*, ser. The Morgan Kaufmann series in evolutionary computation. San Francisco [u.a.] Morgan Kaufmann, 2009.

[7] M. Dorigo, G. D. Caro, and L. M. Gambardella, "Ant algorithms for discrete optimization," *Artificial life*, vol. 5, no. 2, pp. 137–172, 1999.

[8] D. Teodorović, "Bee colony optimization (bco)," in *Innovations in Swarm Intelligence*, C. P. Lim, L. C. Jain, and S. Dehuri, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2009, pp. 39–60. [Online]. Available: https://doi.org/10.1007/978-3-642-04225-6_3

[9] P. Maheshwari, A. K. Sharma, and K. Verma, "Energy efficient cluster based routing protocol for wsn using butterfly optimization algorithm and ant colony optimization," *Ad Hoc Networks*, vol. 110, p. 102317, 2021.

[10] S. Asghari and N. J. Navimipour, "Resource discovery in the peer to peer networks using an inverted ant colony optimization algorithm," *Peer-to-Peer Networking and Applications*, vol. 12, no. 1, pp. 129–142, 2019.

[11] O. Castillo, E. Lizárraga, J. Soria, P. Melin, and F. Valdez, "New approach using ant colony optimization with ant set partition for fuzzy control design applied to the ball and beam system," *Information Sciences*, vol. 294, pp. 203–215, 2015.

[12] Z. Ziyang, Z. Ping, X. Yixuan, and J. Yuxuan, "Distributed intelligent self-organized mission planning of multi-uav for dynamic targets cooperative search-attack," *Chinese Journal of Aeronautics*, vol. 32, no. 12, pp. 2706–2716, 2019.

[13] Y. Liu, J. Ma, S. Zang, and Y. Min, "Dynamic path planning of mobile robot based on improved ant colony optimization algorithm," in *Proceedings of the 2019 8th International Conference on Networks, Communication and Computing*, 2019, pp. 248–252.

[14] R. Uriol and A. Moran, "Mobile robot path planning in complex environments using ant colony optimization algorithm," in *2017 3rd international conference on control, automation and robotics (ICCAR)*. IEEE, 2017, pp. 15–21.

[15] A. Viseras, T. Wiedemann, C. Manß, V. Karolj, D. Shutin, and J. M. Gomez, "Beehive inspired information gathering with a swarm of autonomous drones," *Sensors*, October 2019. [Online]. Available: https://elib.dlr.de/129660/

[16] Z. Li, Z. Zhang, H. Liu, and L. Yang, "A new path planning method based on concave polygon convex decomposition and artificial bee colony algorithm," *International Journal of Advanced Robotic Systems*, 2020. [Online]. Available: https://doi.org/10.1177/1729881419894787

[17] N. R. Sturtevant, "On strongly solving chinese checkers," in *Advances in Computer Games*. Springer, 2019, pp. 155–166.

[18] Z. Liu, M. Zhou, W. Cao, Q. Qu, H. W. F. Yeung, and V. Y. Y. Chung, "Towards understanding chinese checkers with heuristics, monte carlo tree search, and deep reinforcement learning," *arXiv preprint arXiv:1903.01747*, 2019.

[19] M. Roschke and N. R. Sturtevant, "Uct enhancements in chinese checkers using an endgame database," in *Workshop on Computer Games*. Springer, 2013, pp. 57–70.

[20] L. J. P. de Araújo, A. Grichshenko, R. L. Pinheiro, R. D. Saraiva, and S. Gimaeva, "Map generation and balance in the terra mystica board game using particle swarm and local search," in *International Conference on Swarm Intelligence*. Springer, 2020, pp. 163–175.

[21] A. Y. Kapi, M. S. Sunar, and M. N. Zamri, "A review on informed search algorithms for video games pathfinding," *International Journal*, vol. 9, no. 3, 2020.

[22] A. Gonzalez-Pardo, F. Palero, and D. Camacho, "An empirical study on collective intelligence algorithms for video games problem-solving," *Computing and Informatics*, vol. 34, no. 1, pp. 233–253, 2015.

[23] D. Daylamani-Zad, L. B. Graham, and I. T. Paraskevopoulos, "Swarm intelligence for autonomous cooperative agents in battles for real-time strategy games," in *2017 9th International Conference on Virtual Worlds and Games for Serious Applications (VS-Games)*. IEEE, 2017, pp. 39–46.

[24] C. Blum and X. Li, "Swarm intelligence in optimization," in *Swarm Intelligence: Introduction and Applications*, C. Blum and D. Merkle, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2008, pp. 43–85. [Online]. Available: https://doi.org/10.1007/978-3-540-74089-6_2

[25] V. Trianni, *Evolutionary swarm robotics: evolving self-organising behaviours in groups of autonomous robots*. Springer, 2008, vol. 108.

[26] M. Dorigo and T. Stützle, *Ant Colony Optimization*. Cambridge : MIT Press, 2004.

[27] M. Dorigo and L. M. Gambardella, "Ant colony system: a cooperative learning approach to the traveling salesman problem," *IEEE Transactions on evolutionary computation*, vol. 1, no. 1, pp. 53–66, 1997.

[28] K. Diwold, M. Beekman, and M. Middendorf, "Honeybee optimisation – an overview and a new bee inspired optimisation scheme," in *Handbook of Swarm Intelligence: Concepts, Principles and Applications*, B. K. Panigrahi, Y. Shi, and M.-H. Lim, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2011, pp. 295–327. [Online]. Available: https://doi.org/10.1007/978-3-642-17390-5_13

[29] J. Odili, M. N. M. Kahar, A. Noraziah, and S. F. Kamarulzaman, "A comparative evaluation of swarm intelligence techniques for solving combinatorial optimization problems," *International Journal of Advanced Robotic Systems*, vol. 14, no. 3, p. 1729881417705969, 2017.

Reliability and Performances of Power Electronic Converters in Wind Turbine Applications

Aimad Alili, Mamadou Baïlo Camara, Brayima Dakyo and Jacques Raharijaona
GREAH Laboratory, University of Le Havre Normandy
Le Havre, France
Email: Alili.aimad@univ-lehavre.fr,camaram@univlehavre.fr,brayima.dakyo@univ-lehavre.fr,
jacques.raharijaona@univ-lehavre.fr

***Abstract*-The reliability of wind turbines system (WTS) is becoming a key issue as the penetration rate of wind energy is continuing to grow in the last decades. The reliability of a wind turbine is the reliability of all the components and sub-systems that compose the entire system. In this paper, we present a study of the wind turbines structures, currently used components and technologies. A review of wind turbine maintenance data from multiple wind turbines firms installed in different countries and different climatic zones. The study aims to identify the most critical components of the different technologies used in WTS and the effect of the wind turbine structure on the global failure rate of the WTS. We focus on two of the most used configurations in WTS (Variable speed wind turbine with partial-scale converter and Variable speed wind turbine with a full-scale converter) and the power converters associated with these configurations. These converters represent one of the most fragile components according to the data of major reliability studies. The comparison between the reliability rate of the different WTS topologies show the importance of the choice of the configuration and power converter topologies to ensure the availability of WTS.**

Keywords-Power Electronic Converters; Wind turbine; reliability; trends; Failure rate.

I. INTRODUCTION

In order to reduce the dependence of their countries on fossil fuels and to increase the production of electricity by clean energy. Government investment in renewable energies is increasing. Therefore, the renewable energy production is growing worldwide, in 2020, the global production capacity has reached 2799 GW, it is about a third of total installed electricity capacity. Wind power is the second most renewable energies installed in the word with 733 GW, which represents 26% of global energy renewable electricity production [1][2].

The used technology in wind turbine applications has changed since the power capacity penetration has grown dramatically to reach, for example, 14% of all electric energy consumption in Europe and 41% of all electric energy consumption in Denmark [3]. The first configuration used in wind turbine applications was a fixed-speed Squirrel-Cage Induction Generators (SCIGs) directly connected to the grid.

Recently, as the power capacity of the wind turbines increases, regulating the frequency and the voltage in the grid becomes a very important issue. Manufactures are moving toward variable speed Permanent Magnet Synchronous Generator (PMSG) connected to the grid through a power converter. This configuration shows nice properties like high efficiency, small size, and low maintenance; hence, it is a nice choice for wind turbine applications.

The purpose of this paper is to give an overview of recent converters technologies used in WTS. On the other hand, as reliability is a major challenge in WTS, a comparative study about the reliability of the converters is presented.

In Section II, an overview of existing technology market developments of wind power generation. In Section III, the most used wind turbine configurations and currents promising power converters topologies for WTS are presented. In Section IV, the reliability of WTS components is analyzed. In Section V, as they constitute one of the major sources of failure, a study about the reliability of power converters used in WTS is presented. Finally, the conclusions are presented in Section VI.

II. WIND TURBINE SYSTEMS

The wind power installed capacity is growing significantly since 1999 to reach 93 GW installed only in 2020. Therefore, the cumulative installed wind power capacity increased exponentially from 6100 MW in 1996 to 733 GW in 2020. Estimation predicts that this number would reach 2015 GW in 2030. Approximately 10 countries have more than 83% of all cumulative installed wind power capacity in the world, including 5 countries in Europe (Germany, Spain, UK, France, Italy), 2 in the Asia-Pacific (China, India), 2 in North America (US, Canada) and 1 in Latin America (Brazil) [2]. This dominance is shown in Figure 1 and it is obvious that countries with high technology advancements have a higher growth rate and higher penetration of wind power electricity.

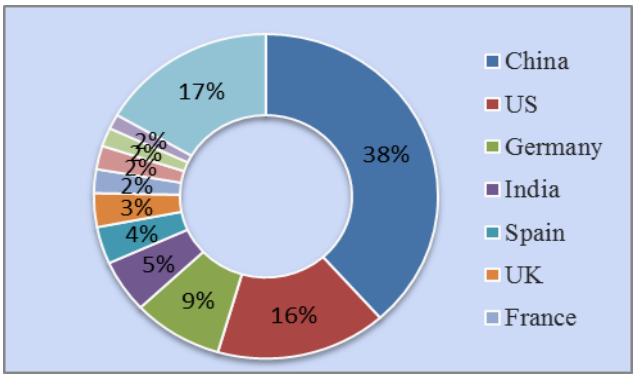


Figure 1. Renewable wind energy capacity in the word.

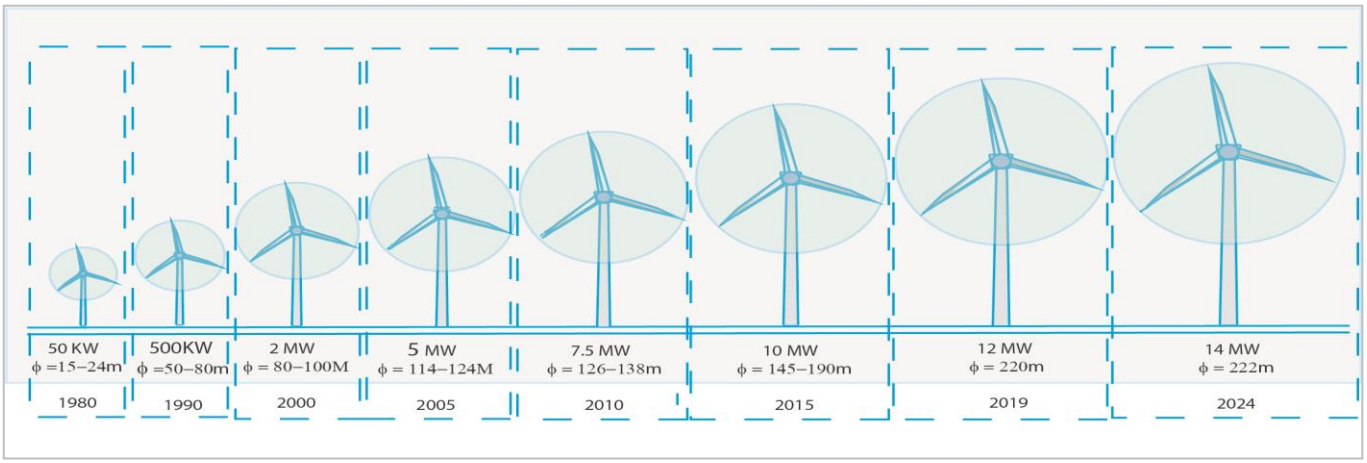


Figure 2. Evoluion of wind turbine size since 1980.

The large turbine presents a lot of advantages. They allow capturing a high power with low installation and maintenance costs compared to the small turbines. Hence, the size of the commercial wind turbine has greatly increased in the last decade, as presented in Figure 2. The largest wind turbine reported in 2021 is 12MW with a diameter of 220 m (General Electric Haliade-X 12 MW), and it will be tested to operate at 13MW. Siemens Gamesa has announced that they are developing 14 MW wind turbine with a rotor diameter of 222 m. It announced that the turbine will be available in 2024 [4].

Denmark based wind turbine company Vestas remained the world's largest wind turbine manufacturer and supplier in 2018 [5], due to its wide geographic diversification strategy and strong performance in the U.S. market.

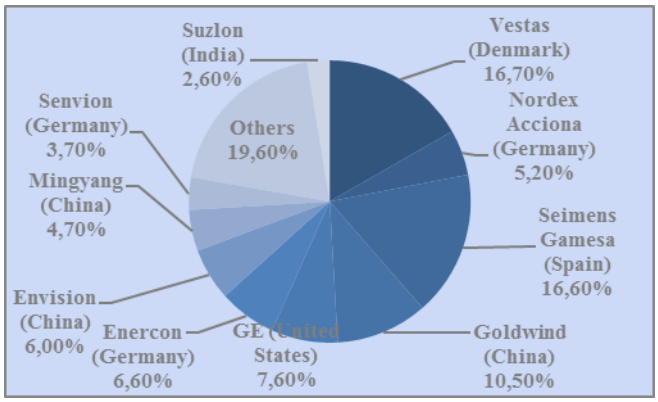


Figure 3. Top 10 wind turbine suppliers market share in 2018 [5].

Top 10 wind turbine manufacturers in the world are shown in Figure 3. The world's largest wind turbine companies account for over 75% of the global installed capacity every year, and their industrial dominance is expected to continue over the future.

III. WIND TURBINE CONCEPTS AND CONVERTERS TOPOLOGIES

A. *Wind turbine concepts*

Depending on the types of generator, power converters and speed control, most wind turbine structures can be classified into following four types:

- Type 1: Fixed-speed wind turbine systems;
- Type 2: Semi-variable speed wind turbine with variable rotor resistance;
- Type 3: Variable speed wind turbine with partial-scale converter;
- Type 4: Variable speed wind turbine with a full-scale converter.

All these wind turbine technologies have been used and commercialized in the last 30 years. Due to their efficiency, the two last configurations are the most dominant technologies in the market. In the following, these two wind turbine concepts are going to be exposed.

1) *Variable Speed Wind Turbine with a partial-scale converter*

Variable speed wind turbine with the partial-scale converter is generally associated with a doubly fed induction generator (DFIG), the typical configuration of this technology is shown in Figure 4. The induction generator is directly connected to the grid and the rotor is interfaced through a back-to-back power electronic converter. The converter system includes two AC/DC-based Voltage Source Converters (VSCs) connected by a DC-bus voltage. The power converter controls the rotor frequency and thus the rotor speed. Typically, the variable speed range is +30% around the nominal speed [7][8]. The main advantage is that only a part of the power production is fed through the power electronic converter. Hence, the nominal power of the power electronics converter system can be less than the nominal power of the wind turbine. In general, the nominal power of the converter is about 30% of the wind turbine power. The gearbox is essential in this type of configuration. Some commercial solutions using this technology are Repower 6M, 6.0 MW; Bard 5.0, 5 MW; Senvion 6.2m 126; General

electric GE.6-82.5 and Acconica AW-100/ 3000, 3 MW; Shanghai el W3600/122, 3.6MW; Nordex N80,1.5-2.5MW; Sinovel, 3MW [21].

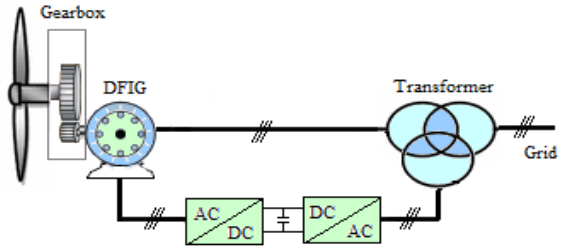


Figure 4. Variable-speed wind turbine with a partial-scale converter

2) Variable Speed with a Full-scale Converter

A variable speed wind turbine partial-scale converter is shown in Figure 5. In this configuration, the wind turbine uses a full-scale power converter between the generator and grid to enhance performance. Since all the generated power has to pass through the power converter, the power converter must be rated the same as generator capacity, which involves increasing the size, cost, and complexity of the system. However, the wind energy conversion efficiency is highest in this turbine compared to other types of turbines and the gearbox can be eliminated by using a high pole synchronous generator [13].

The PMSG, SCIG, and Wound Rotor Synchronous Generator (WRSG) have all been used in this type of configuration. However, due to the reduced losses, weight, and noise, the PMSGs are most commonly adopted and they are becoming the best seller technology in the wind energy market. Manufactures are commercialized several models of wind turbines based on PMSG full-scale technology: Goldwind GW140/3000; Enercon E126, 7.5 MW; Siemens Gamesa SG 8.0-167 DD, 8 MW; General electric Haliade 150, 6MW; Multibrid M5000, 5 MW; Adwen AD 5-135 and Vestas V-112, 3 MW [7][20].

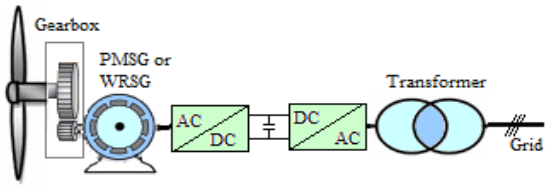


Figure 5. Variable speed wind turbine with a full-scale converter

B. Converters For Wind Turbine System

The power converter is one of the most important components of Wind Turbine System (WTS). The main objective of the power converters is to ensure the generator speed variation control in the turbine system. To accomplish this purpose, different topologies of converter have been proposed in the literature in the last decades. Recently, with the growing wind turbine penetration, these converters have to fulfil several technical requirements. The converter cost is an important factor, since it represents approximately 7%~8% of the global cost of the wind turbine system [9][10]. The cost of maintenance must also be as low as possible to reach less expensive and competitive energy compared with the others sources of energy, reliability is also an important element in the choice of the converters. The efficiency of the converters is also very important, especially in high power wind turbines where even, 1% efficiency improvement can save thousands of dollars over a period of a few years [20]. The output power quality of the converters is a primordial in the comparison between the different topologies. The output voltage should be as close as possible to the sinusoidal shape with low total harmonic distortion (THD) and small filter for a better converter [13][16][18]. The power converters can be classified as direct and indirect according to the different stages of the conversion. Overall, the indirect Back-to-Back (BTB) converter technology is the most used in the wind turbine applications [11].

1) Two-levels Voltage Source Converter (2L-VSC)

The two-level voltage source converters are the most widely used converters on the market. For its simple configuration this technology is mastered and well established in the field of wind energy conversion. It is considered a dominant topology used in around 90% of the wind turbines with power less than 0.75 MW. As illustrated in Figure 6, the Voltage Source Rectifier (VSR) and the Voltage Source Inverter (VSI) are back-to-back and are connected to a DC-bus capacitor. This DC-bus ensures the decoupling between the generator and the grid, therefore transient in the generator do not appear on the grid side. The VSR controls the torque and speed of the generator, while the VSI controls the voltage of the DC-link and the reactive power of the grid.

The VSR and the VSI are generally made with low-voltage transistors (LV-IGBT) arranged in a matrix. The switching frequency of VSR and VSI are fixed between 1 and 3 kHz to achieve low witching loss and high power density [6].

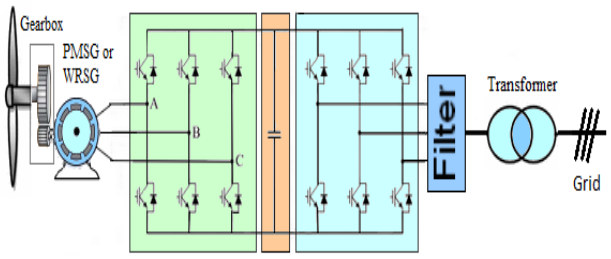


Figure 6. BTB based on the two-levels voltage source converter.

2) Parallel two-levels Voltage Source Converter (2L-VSC)

To achieve high current capacity, two or more VSC converters can be connected in parallel depending on the power required. As illustrated in Figure 7, two VSC modules are connected in parallel to reach a power of 1.5 MW corresponding to type 4 of the wind turbines. For type 3 of the wind turbines, connecting two modules in parallel can achieve a power of 5 MW. This configuration allows a wide margin for redundant operation. To improve the system efficiency in the case of underproduction, one or more converter modules can be put out of service. The redundancy of the converters allows the wind turbine to continue operating at reduced capacity in the case of a fault in the converters, after the faulty module is isolated. In the Gamesa G128, more than 6 power converters are connected in parallel to reach a nominal power of 4.5 MW [12]. However, the major disadvantage is that a large number of modules lead to the complexity control and congestion of the system.

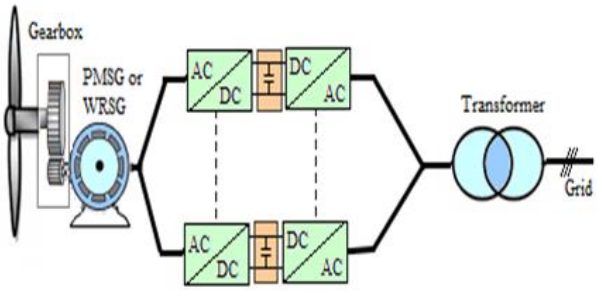


Figure 7. WTS with parallel connected BTB Two-levels VSCs

3) Three-levels Neutral-Point Clamped Converter (3L-NPC)

Another solution that has been widely studied in the literature for type 4 of the wind turbines is the three-levels Neutral Point Clamped converter (NPC). In this configuration, an arrangement of four power switches per leg, clamped with diodes to a midpoint of the dc-link. With this configuration, each power device has to block only half of the total converter voltage then the power of the converter can be doubled [14]. The output phase voltage of the converter contains three-levels leading to a reduced voltage variations dv/dt and electromagnetic interference compared to the 2L-VSC converters [13][17][21][22]. The main drawback of 3L-NPC is that the power switches do not have symmetric losses, forcing a derating of the devices. As shown in Figure 8, NPC converters enable medium voltage operation, and commercial wind turbines reached 6 MW rated power without connecting serial or parallel switching devices. These converters are installed and marketed with the "Multibrid M5000" wind turbine [7] [23].

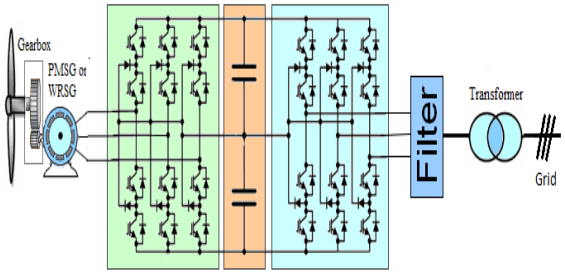


Figure 8. Three-levels Neutral-Point Clamped Converter (3L-NPC) .

4) Three-levels Active Neutral-Point Clamped Converter (3L-ANPC)

Active Neutral Point Clamp (ANPC) converters illustrated in Figure 9 have a structure almost identical to the NPC converters, the diodes are replaced by Insulated Gate Bipolar Transistor (IGBT) switches. Although more active switches are used, that allowing more redundancy to maintain the frequency and the same switching losses in all the IGBT switches [7][13][24][25]. In similar operations, BTB 3L-ANPC converters are capable of handling 32% higher power (up to 7.12 MW) and 57% higher switching frequency (1650 Hz) compared to 3L-NPC BTB converters. This configuration has been applied more recently in the field of MV drives and can also be used in the wind turbine system sector [19]. Vestas, one of the leading manufacturers, is currently studying this power converter topology [20].

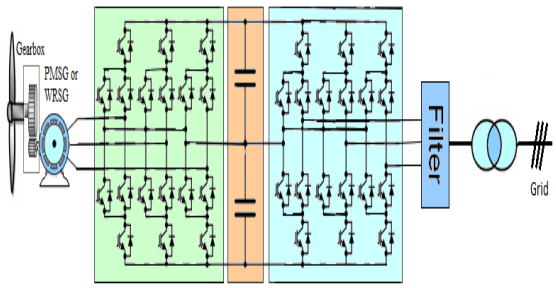


Figure 9. Three-levels Active Neutral-Point Clamped Converter (3L-ANPC).

5) Three-levels Flying Capacitor Converter (3L-FC)

The configuration of the Flying Capacitor converter (FC) is similar to the NPC converter, where the clamping diodes are replaced by the floating capacitors. The concept of FC was introduced in the early 1970, and was introduced into machine drives applications in 1990. The converter generates additional voltage levels while reducing voltage stress on the drive [15]. The power switches, setting an FC between two devices, are illustrated in Figure 10. Each pair of switches with an FC constitutes a power cell.

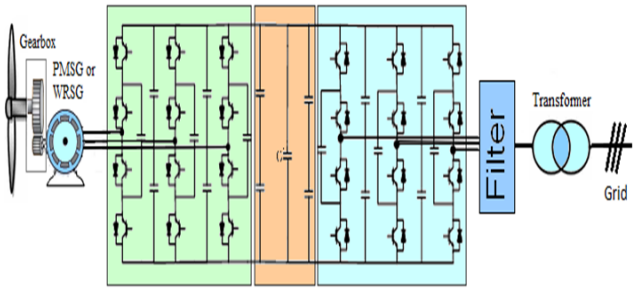


Figure 10. Three-levels Flying Capacitor Converter (3L-FC)

The most important difference with the NPC topology is that the FC has a modular structure and additional cells can be connected, increasing the number of voltage levels of the converter and the power rate.

The advantages of the flying capacitor multilevel converter are flexible switch mode, high protection ability to power devices, to control active power and reactive power conveniently [26][27]. The 3-levels configuration has found a practical application, but has not yet found a commercial success in wind turbines.

IV. WIND TURBINE SYSTEM RELIABILITY

Recently, with the orientation of wind energy manufacturers towards offshore wind turbines, the issue of the reliability of wind systems has become a major preoccupation due to the maintenance costs caused by limited accessibility of wind farms. This problem has been extensively studied in literature [28-33].To identify the major cause of failure in wind systems. Researchers have conducted surveys of the reliability of wind systems at various wind sites around the world to identify the most common faults in these systems:

- CIRCE

CIRCE is a research project on the reliability of wind systems conducted by the University of Zaragoza in Spain. SCADA data are collected over a period of three years on different wind farms, the totality of wind turbines studied is close to the 4300 onshore wind turbines of variable power between 0.3 and 3 MW. Two wind turbines configurations, geared and direct drive technologies are studied in the study. Data from the reliability analysis of the 23 wind farms included in this research are published in [34].

- CWEA

The CWEA study is carried out in China over a period of two years between 2010 and 2012. The data are published in [35] where the data analysis of 640 wind turbines with a power between 1.5 and 6 MW is presented. In this study, only the critical faults are considered, and the published data do not allow to differentiate the types of technologies used in these wind turbines.

- LWK

In [35] the LWK project data are presented. The data are collected over a period of 13 years from onshore wind sites in northern Germany. In total, the maintenance data of 643 wind turbines of power varying between 0.2 to 2 MW per wind turbine are exploited in this study. The reliability of wind power systems of both geared and direct drive concepts is studied and the failure rate of different components is exposed.

- Huardian project

The study includes maintenance data from 26 wind farms located in China. These sites are made up of 1313 wind turbi nes of different technology type and of unspecified power. The study is published in [36], failures are presented as a pe-rcentage, which makes it difficult to use the data.

- EPRI

Electric Power Research Institute (EPRI) based in the USA is at the initiative of this project. The data come from maintenance data from various wind farms in California. The number of wind turbines exploited in the study is small (290 wind turbines) of very low power, which varies between 0.04 to 0.6 MW. The technology of the wind turbines studied is very old, due to the fact that the study was carried out between the years 1986 and 1987. The study is published in [37].

TABLE I. WIND TURBINE SYSTEM COMPONENTS

Subsystem	Assembly	Subsystem	Assembly
Rotor system	Blade Hub Air brake Pitch system	Hydraulic system	Hydraulic system
		Yaw system	Yaw system
		Control system	Control system Sensors Data acquisition system
Drive train	Shafts and bearings Mechanical brake	Electrical system	Converter Transformer Electrical protection and switchgear
Gearbox	Gearbox		
Generator	Generator	Structure	Tower Foundations
Other	Other		

- ELforsk/Vindstat

The study was published in [38], it is based on the recovery of maintenance data from wind farms in Sweden. It comprises 723 onshore wind turbines with a capacity of 0.055 to 3 MW monitored between 1997 and 2004. The study provides the failure rate and downtime of the various wind turbine components for the period studied.

- Muppandal

This study is based on maintenance data from the Muppandal wind farm in southern India. The data are published in [39], where an analysis of the performances, failure and reliability of 15 wind turbines with a power of 225 kW are presented. The recovery of maintenance data is over a period of 4 years between 2000 and 2004.

- NEDO

The study is conducted by Japanese New Energy and Industrial Technology Development Organization (NEDO) and published in [40]. The study took place over a period of one year between 2004 and 2005. The number of wind turbines included in this study is 924 turbines. Only, faults with a downtime, greater than 72 h are considered as a failure in this study. This explains the very low failure rate of the various components of a wind power system, and makes any comparison with other reliability studies subjective.

- WMEP

WMEP (Wissenschaftliches Mess- und Evaluierungsprogramm) is a German project on the reliability of wind power systems. The data are published in [41], the study started in 1989 over a period of 17 years. The number of wind turbines contained in the study is 1500 wind turbines of different technologies and power varying between 0.03 MW and 1.8 MW. The study is rich in information, it allows us in particular to have the failure rate and the downtime of different wind turbine components over a long period of time.

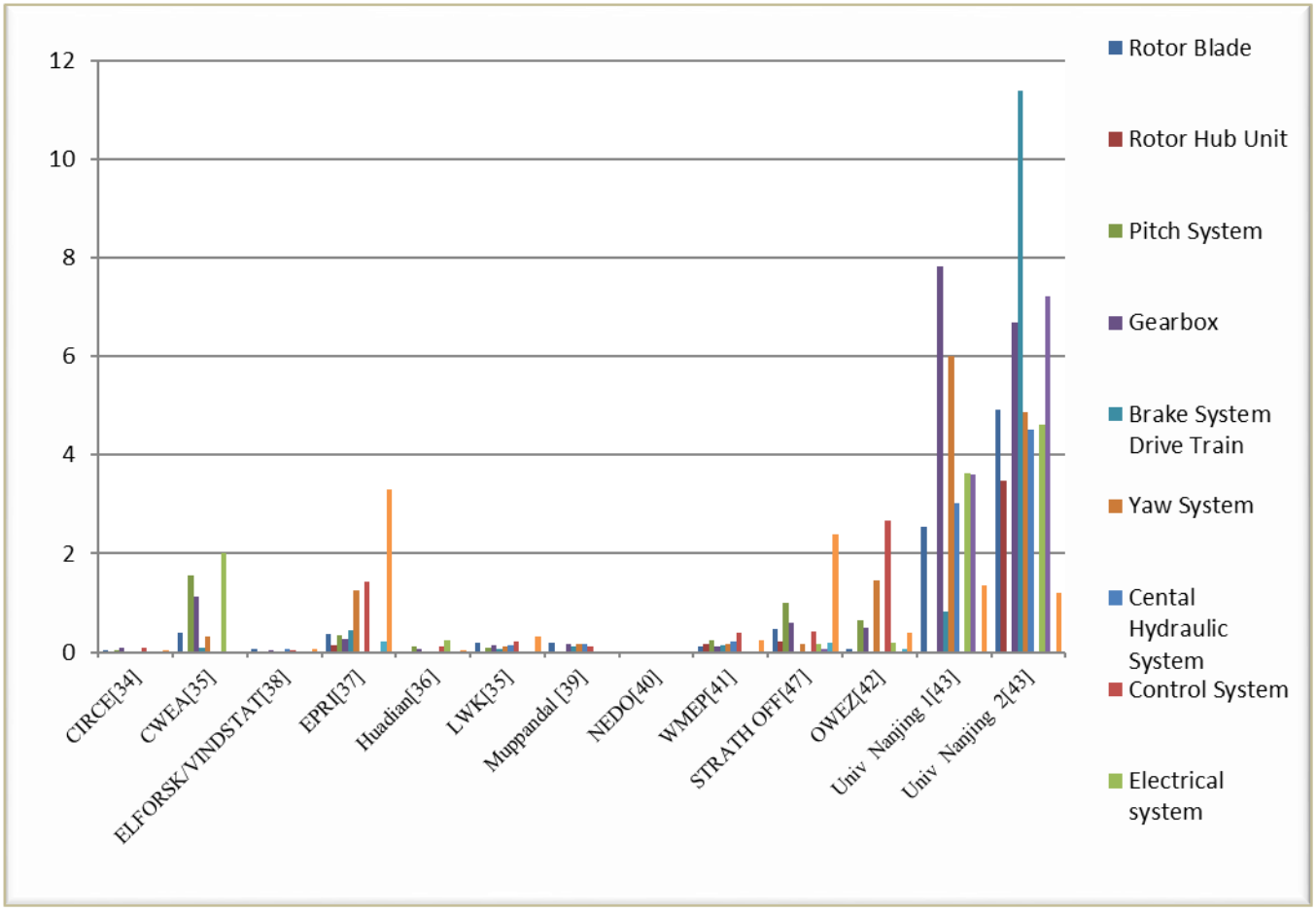


Figure 11. Wind turbine reliability study comparison [35-43].

- University of Strathclyde

In [47], maintenance data from 2220 turbines were studied to determine the failure rate in the various components of these systems. The wind turbines studied are modern types, with power ranging between 1.5 MW to 2 MW. They are divided into two groups, depending on the training configuration. The first group is made up of 1800 turbines based on the DFIG generators. The second group consists of 400 turbines equipped with the PMSG generators.

- OWEZ

Offshore Wind farm Egmond Aan Zee (OWEZ) is an offshore wind farm launched in 2007 in Netherlands. The site is made up of 36 Vestas V90 wind turbines, with a power of 3 MW per turbine. The maintenance data are published annually by NoordzeeWind and are analyzed in [42].

- University of Nanjing

In 2016, a study by the City University of Nanjing on two wind farms in China [43]: The first project, which contains 61 wind turbines of 1.5 MW, with data recovered over a period of 4 years between 2009 and 2013. The second project contains a few numbers of wind turbines, 46 wind turbines, but with a power greater than that of the first project, 2 MW by a wind turbine.

Figure 11 shows the results of the most relevant studies published in the literature and obtained from data recovery at different wind turbine sites around the world [34-43]. The distribution of wind turbine system components over the different subsystems is presented in Table I. The results show that the defect distribution rate varies between the different studies and this is mainly due to two reasons: the location of the wind turbines studied: the climatic regions can have considerable effects on the reliability of certain components. The second reason is the technology used in the different wind turbines, the reliability of the wind systems is also related to the manufacturers and the importance given to the reliability of the components during the development phase.

In this section, we focus on the result of some of majors studies. According to a study published by the University of Kassel, Germany in 2006 [44], based on the recovery of maintenance data for 13 years, the power converters are a leading cause of failure in a wind system as shown in Figure 12.

Another study [45], shows that the use of maintenance data LWK allowed researchers to identify the main causes of failures on this site. The conclusions given in [45] show that the defects in the converters represent a large part of these defects. They are ranked in 3rd position, just behind the faults in the electric system control and the mechanical defects in the rotor.

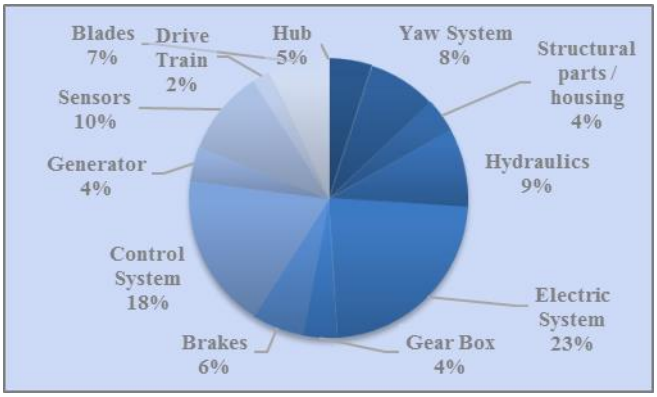


Figure 12. Share of main components of total number of failures [44].

More recently, in 2016, in the study [43], the researchers found different results: for the first project, the result shows that electrical systems (converters) account for 14% of the failures. The control of the wind system accounts for the largest share of these defects, with 35% of total defects recorded over this period. In the second project, the analysis of maintenance data over a period of two years shows that electrical systems (converters) account for 26% of failures, equal to failures rate found in the control system.

The wind turbines designed in 2000 are generally based on fixed or semi-variable speed technology, different from the technology generally used this last decade based on the synchronous machine with variable speed. It can be noted that the zone of the installation of these fields also plays an important role in the rates of defects of the components. Another point that can be made is that the reliability of wind turbine systems depends on the reliability of the used components and experience of the manufacturers. However, despite the difference in the failure rates between the different components of the wind system, which can be found in the different studies, the defects in the converters are considered as a major element in the shut-down of the service in the almost all of these studies.

V. CONVERTERS RELIABILITY IN WIND TURBINES APPLICATIONS

In the following, a deep analysis of the different reliability studies based on wind turbines around the world is proposed. The purpose of this analysis is to find a link between the different systems used in wind turbines and the failure rate in power converters. This study will allow us to identify the causes of failures in power converters and to propose the solutions and topologies to be used to improve the reliability of wind turbine systems. In [45] and [46], the results of the data recovered on the wind farms, affirm that, contrary to that is widespread in the literature, the failures in the systems with direct drive permanent magnet synchronous generator (PMSG) are more significant to those with indirect drive doubly-fed induction generator (DFIG). In their conclusions, the authors ask questions about the usefulness of the systems based on PMSG generators with the number of failures recorded, while it is supposed to improve the reliability of wind turbine systems. In the same study of the University of Columbia [46], the failure rates in the converters used in the case of the two systems (Geared/ Direct drive) are studied and compared. The failures at the converters are greater in the case where the system is based on direct drive technology. In another study presented in [47], maintenance data from 2220 turbines were studied to determine the failure rate in the various components of these systems. The wind turbines studied are modern types, with power ranging between 1.5 MW to 2 MW. They are divided into two groups, depending on the training configuration. The first group includes of 1800 turbines based on the DFIG generators. The second group consists of 400 turbines equipped with the PMSG generators. It should be noted that the converters used in the two configurations belong to the same manufacturer, which will allow us to analyze and compare the failures of these converters for each configuration. The comparison between the failure rates of the power converters of the two systems illustrated in Figure 13 shows that the converter in the direct drive system with PMSG generators presents an annual failure rate of 0.593, which is approximately four times more than the failure rates recorded in the system based on the DFIG generators.

TABLE II. COMPARISON OF BTB CONVERTERS TOPOLOGIES FOR HIGH POWER WIND TURBINES [7].

	2L-VSC	Parallel 2L-VSC	3L- NPC	3L- NPC-modified	3L-ANPC	3L-FC
Typical power	0.75 MW	5.0 MW	3.0-12.0 MW	3.0-12.0 MW	3.0-12.0 MW	3.0-12.0 MW
Number of converters	1	6	1	1	1	1
Number of switches	12	72	24	32	36	24
Switching devices	LV-IGBT	LV-IGBT	MV-IGBT/ICGT	MV-IGBT/ICGT	MV-IGBT/ICGT	MV-IGBT/ICGT
Diodes	0	0	12	16	0	0
Capacitors	0	0	0	0	0	6
Device voltage stress	V_{dc}	V_{dc}	$V_{dc}/2$	$V_{dc}/2$	$V_{dc}/2$	$V_{dc}/2$
Reliability of system	++	+++	+++	++++	+++	++
Redundancy	No	Yes. Module redundancy	No	Yes. Leg redundancy	No	No
Advantages	Simple and matured technology	Redundancy	Low harmonic matured technology	Low harmonic redundancy	Low harmonic Equal loss distribution	Low harmonic
Disadvantages	Limited power	Complex control	Unequal loss distribution	Unequal loss distribution	A large number of switches	Complex control
Technology status	Highly mature	Highly Mature	Well established	Research only	Research only	Research only
Power density	Moderate	Low	High	High	High	High

In [32], Koutoulakos presented a study of 643 WTs in Schleswig Holstein (LKW) Germany. The wind turbines are either fixed or variable speed configuration, and geared or with the direct drive concept. The study includes a data failure rates per turbine per year for different wind turbine sizes. We divided the WTs into two groups based on power (0.5 to 0.6 MW group and around 1 MW group). Reliability data of the various wind turbine components are separately analyzed to identify critical subassemblies of each topology. The comparison of geared and direct drive topologies shows that; the larger WTs had longer downtimes and higher cost. In Figure 14, the result shows that the electric failures in the direct driven wind turbine are more frequent for the two groups.

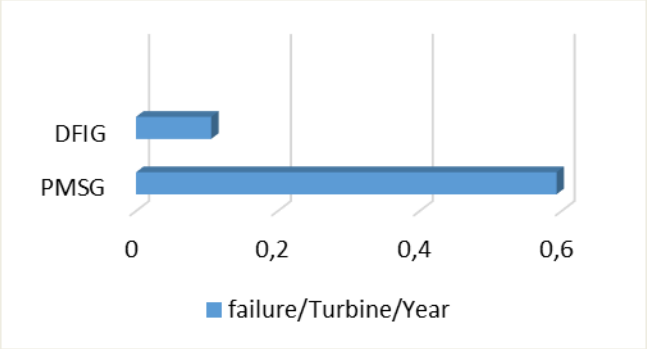


Figure 13. Annual failure rate [47].

To answer the questions on the number of significant failures recorded in PMSG generators systems compared to systems based on the DFIG generators, asked by the authors in, [45][46].This difference is mainly due to the increase in failures in power converters. Knowing that for the same power value of a turbine, the used converter in the PMSG systems must have a power three times higher than that of the converter used in DFIG systems. Since in the latter case, the power supplied via the converter represents only part of the overall power supplied by the turbine. Since the maximum power of a two-levels converter does not exceed 750 kW, to increase the powers in PMSG systems, manufacturers tend to put in parallel several two-level converters as presented in Figure 5. This solution enables to increase the number of switches and proportionally to the number of failures in the system. Figure 14 illustrates a comparison between the failures of converters recorded in three studies [32][46][47] with turbines of different powers. It is noted that the increase in the power of the turbine generates the increase in the difference between failure rates between the DFIG systems and the PMSG systems, due to the need of paralleling the two-level converters to achieve the desired power.

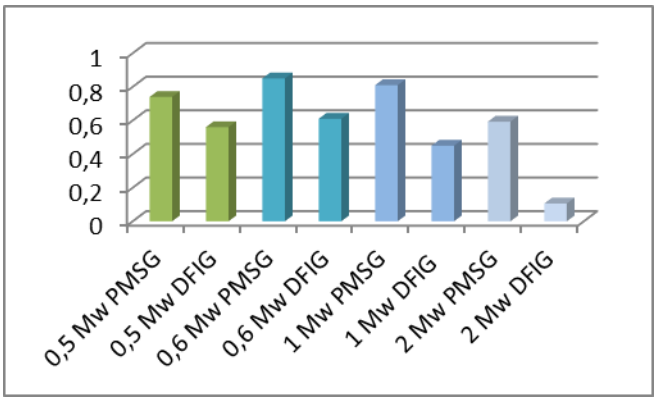


Figure 14. Converters failures rates based on the power of the turbine [32][46][47].

In the literature, the reliability of the power converters is linked only to the redundancy of the system, something which could be sufficient in the case of onshore wind turbines. Furthermore, in offshore wind turbines where for economy and production reasons, the reliability requirement is more important, any source of failure must be considered.

Therefore, in this study the analysis of several maintenance data from different wind farms around the world, allowed us to identify the importance and the need to take into account the number of switches used in the converters as a criterion for the reliability of energy conversation systems. Moreover, the choice of the most reliable converter topology is depending on the power of the application. In Table II, an example of reliability of converters for a 5 MW wind turbine, in this case the paralleling of 2-level converters does not offer higher reliability since the number of switches, which are fragile components becoming very important and decrease the reliability of the whole system.

Several attempts to develop methods to ensure continuity of service in converters have been developed in the literature. These methods can be classified into two categories, the one that uses hardware redundancy and the one that has no hardware redundancy. The first solution ensures continuity of operation in the nominal mode without power reduction while the second solution envisaged aims a degraded mode of the converter, which certainly continues to operate, but with lower power.

The hardware redundancy generally requires the use of other additional components, it was used for the 2-VSL converters in [51-55]. The most popular technique is based on the addition of a redundant arm can be activated when a fault is detected in one of the converter arms, this arm will be isolated using triacs or fuses. The same solution is used in [48-50] for NPC converters with the addition of an NPC redundant arm or FC redundant arm, as illustrated in Figure 15. Similar strategy is also used for ANPC converter as shown in [56-57].

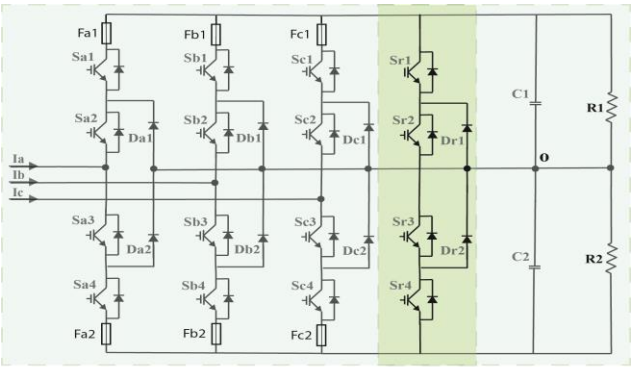


Figure 15. Redundant NPC Converter

On the other hand, in order to improve converters reliability and availability, various controls strategies to improve the system response in case of defects without hardware redundancy were studied. In [58-59], the authors proposed a control method using voltage vector redundancy for a Neutral Point Clamped (NPC) inverter. Similar approaches that do not require additional components are applied for other three-level topologies, such as the T-type converter [60], the active NPC converter [61-63].

VI. CONCLUSION

This paper presents the current technologies used in wind power systems and those may be a possible solution for WTS in the near future. A particular focus has been placed on power converters, as they are one of the most important components in the energy conversions in wind power systems. The other important point that we mention in this study is the reliability of wind power systems, a study of the various research and data available in the literature on the reliability of wind power systems is presented. Then an analysis and a comparison of the reliability between the different topologies used in wind power systems have been developed. This analysis allows us to observe:

-The data from different studies on the reliability of the component of wind turbine systems may be different and it is mainly due to: the different meteorological conditions of the wind turbines firms studied ; the different manufacturers and the importance that they give to the reliability in their designing processes; and the definition that researchers give to failure. For example, in the NEDO study, only breakdowns, which cause 72 hours of downtime are considered as failure, it explains the low rates of failures in this study.

-The choice of the converters is important in the global wind turbines systems. The two-level converters are most used but with the current trend towards systems with high power levels and high voltage levels, multilevel converters, especially three-levels NPC converters represent the most suitable system. For reliability of the power converter issues, the failures at the converters are proportional to the number of switches used in the energy conversion system.

- The majority of recent studies claim that, contrary to popular belief, failures in systems based on PMSG generators are greater than failures in systems based on the DFIG generators. In our study, we explain the reasons, which are mainly due to the significant increase in failures in the power converter systems. However, systems based on the PMSG generators remain the most reliable for onshore wind turbines since the downtime caused by a failure at the converters is significantly lower than that caused by a failure of the gearbox.

In the literature very few studies allow access to components maintenance data according to the structure of the studied wind turbines, most of this research separates the studied wind turbines only according to their power. Except these information are important for comparative analyzes and to deduce more reliable conclusions.

REFERENCES

[1] A. Alili, M. B. Camara, B. Dakyo, and J. Raharijaona, "Power Electronic Converters Review for Wind Turbine Applications: State of Art, Reliability and Trends," Paper presented at GREEN 2020 conference, IARIA, pp.12-18, Valencia, 21 – 25 Nov 2020.

[2] https://www.irena.org/publications/2021/March/Renewable Capacity-Statistics-2021, Nov. 2021.

[3] https://windeurope.org/wp-content/uploads/files/about-wind/statistics/WindEurope-Annual-Statistics-2018.pdf , Nov. 2021.

[4] https://www.siemensgamesa.com/en-int/products-and-services/offshore/wind-turbine-sg-14-222-dd , Nov. 2021.

[5] https://cleantechnica.com/2019/04/17/vestas-installed-1-out-of-5-wind-turbines-globally-in-2018 , Nov. 2021.

[6] D. Ikni, M. B. Camara, A. Payman, and B. Dakyo, "Dynamic control of wind energy conversion system," 2013 Eighth International Conference and Exhibition on Ecological Vehicles and Renewable Energies (EVER), Monte Carlo, 2013, pp. 1-5, doi: 10.1109/EVER.2013.6521633.

[7] V. Yaramasu et al., "High-power wind energy conversion systems: state-of-the-art and emerging technologies," in Proceedings of the IEEE, vol. 103, no. 5, pp. 740-788, May 2015, doi: 10.1109/JPROC.2014.2378692.

[8] X. Sun, D. Huang, and G. Wu, " The current state of offshore wind energy technology development," Int. J. Energy, vol. 41, no. 1, pp. 298–312, 2012.

[9] N. Flourentzou, V. Agelidis, and G. Demetriades, "VSC-based HVDC power transmission systems: An overview," IEEE Trans. Power Electron., vol. 24, no. 3, pp. 592–602, Mar. 2009.

[10] W. Kitagawa and T. Thiringer, "Inverter loss analysis and comparison for a 5 MW wind turbine system," 2017 19th European Conference on Power Electronics and Applications (EPE'17 ECCE Europe), Warsaw, 2017, pp. P.1-P.10.

[11] M. S. Camara, M. B. Camara, B. Dakyo, and H. Gualous, "Permanent Magnet Synchronous Generators for offshore wind energy system linked to grid-modeling and control strategies," 2014 16th International Power Electronics and Motion Control Conference and Exposition, Antalya, 2014, pp. 114-118, doi: 10.1109/EPEPEMC.2014.6980620.

[12] B. Andresen and J. Birk, "A high power density converter system for the Gamesa G10x 4,5 MW wind turbine," 2007 European Conference on Power Electronics and Applications, Aalborg, 2007, pp. 1-8.

[13] S. Kouro, J. Rodriguez, B. Wu, S. Bernet, and M. Perez, "Powering the Future of Industry: High-Power Adjustable

Speed Drive Topologies," in IEEE Industry Applications Magazine, vol. 18, no. 4, pp. 26-39, July-Aug. 2012.

[14] M. M. G. Lawan, J. Raharijaona, M. B. Camara and B. Dakyo, "Three level Neutral-Point-Clamped Inverter Control Strategy using SVPWM for Multi-Source System Applications," 2019 IEEE International Conference on Industrial Technology (ICIT), Melbourne, Australia, 2019, pp. 562-567, doi: 10.1109/ICIT.2019.8754968.

[15] W. Kitagawa and T. Thiringer, "Inverter loss analysis and comparison for a 5 MW wind turbine system," 2017 19th European Conference on Power Electronics and Applications (EPE'17 ECCE Europe), Warsaw, 2017, pp. P.1-P.10.

[16] J. Rodriguez et al., "Multilevel Converters: An Enabling Technology for High-Power Applications," in Proceedings of the IEEE, vol. 97, no. 11, pp. 1786-1817, Nov. 2009.

[17] Z. Zhang, Z. Li, M. P. Kazmierkowski, J. Rodríguez and R. Kennel, "Robust Predictive Control of Three-Level NPC Back-to-Back Power Converter PMSG Wind Turbine Systems With Revised Predictions," in IEEE Transactions on Power Electronics, vol. 33, no. 11, pp. 9588-9598, Nov. 2018.

[18] F. Blaabjerg and K. Ma, "Future on Power Electronics for Wind Turbine Systems," in IEEE Journal of Emerging and Selected Topics in Power Electronics, vol. 1, no. 3, pp. 139-152, Sept. 2013.

[19] M. Liserre, R. Cardenas, M. Molinas and J. Rodriguez, "Overview of Multi-MW Wind Turbines and Wind Parks," in IEEE Transactions on Industrial Electronics, vol. 58, no. 4, pp. 1081-1095, April 2011.

[20] F. Blaabjerg, K. Ma, and Y. Yang, "Power Electronics for Renewable Energy Systems - Status and Trends," CIPS 2014; 8th International Conference on Integrated Power Electronics Systems, Nuremberg, Germany, 2014, pp. 1-11.

[21] Z. Zhang, X. Cai, R. Kennel, and F. Wang, "Model predictive current control of three-level NPC back-to-back power converter PMSG wind turbine systems," 2016 IEEE 8th International Power Electronics and Motion Control Conference (IPEMC-ECCE Asia), Hefei, 2016, pp. 1462-1467.

[22] O. S. Senturk, L. Helle, S. Munk-Nielsen, P. Rodriguez, and R. Teodorescu, "Power Capability Investigation Based on Electrothermal Models of Press-Pack IGBT Three-Level NPC and ANPC VSCs for Multimegawatt Wind Turbines," in IEEE Transactions on Power Electronics, vol. 27, no. 7, pp. 3195-3206, July 2012.

[23] X. Jing, J. He, and N. A. O. Demerdash, "Application and losses analysis of ANPC converters in doubly-fed induction generator wind energy conversion system," 2013 International Electric Machines & Drives Conference, Chicago, IL, 2013, pp. 131-138.

[24] Y. Deng et al., "Improved Modulation Scheme for Loss Balancing of Three-Level Active NPC Converters," in IEEE Transactions on Power Electronics, vol. 32, no. 4, pp. 2521-2532, April. 2017.

[25] Q. Guan et al., "An Extremely High Efficient Three-Level Active Neutral-Point-Clamped Converter Comprising SiC and Si Hybrid Power Stages," in IEEE Transactions on Power Electronics, vol. 33, no. 10, pp. 8341-8352, Oct. 2018.

[26] A. Abdelhakim, P. Mattavelli, and G. Spiazzi, "Three-Phase Three-Level Flying Capacitors Split-Source Inverters: Analysis and Modulation," in IEEE Transactions on Industrial Electronics, vol. 64, no. 6, pp. 4571-4580, June 2017.

[27] A. B. Ponniran, K. Orikawa, and J. Itoh, "Minimum Flying Capacitor for N-Level Capacitor DC/DC Boost Converter," in IEEE Transactions on Industry Applications, vol. 52, no. 4, pp. 3255-3266, July-Aug. 2016, doi: 10.1109/TIA.2016.2555789.

[28] S. Pfaffel, S. Faulstich, and K. Rohrig, "Performance and Reliability of Wind Turbines: A Review," Energies. 2017; 10(11):1904. https://doi.org/10.3390/en10111904.

[29] X. Jin, W. Ju, Z. Zhang, L. Guo, and X. Yang, "safety analysis of large wind turbines," Renewable and Sustainable Energy Reviews, Elsevier, Volume 56, 2016, Pages 1293-1307, ISSN 1364-0321, doi: 10.1016/j.rser.2015.12.016.

[30] E. Artigoa, S. Martin-Martinez, A. Honrubia-Escribano, and E. Gômez-Làzaro, "Wind turbine reliability: A comprehensive review towards effective condition monitoring development," Applied Energy,Volume 228, 2018, Pages 1569-1583, ISSN 0306-2619, doi: 10.1016/j.apenergy.2018.07.037.

[31] J. M. P. Pérez, F. P. G. Màrquez, A. Tobias, and M. Papaelias, "Wind turbine reliability analysis," Renewable and Sustainable Energy Reviews,Volume 23,2013,Pages 463-472,ISSN 1364-0321, doi: 10.1016/j.rser.2013.03.018.

[32] E. Koutoulakos, "Wind turbine reliability characteristics and offshore availability assessment," [Master's thesis]. Delft University, Wind Energy Research Institute, 2010.

[33] S. Faulstich, B. Hahn, and P. J. Tavner, "Wind turbine downtime and its importance for offshore deployment," Wind Energy. 14. 10.1002/we.421.

[34] M. D. Roder, E. Gonzalez, and J. J. Melero, "failures—Tackling current problems in failure data analysis," J. Phys. Conf. Ser. 2016, 753, 072027. 27.

[35] P. J. Tavner and F. Spinato, "Reliability of different wind turbine concepts with relevance to offshore application," In Proceedings of the European Wind Energy Conference, Brussels, Belgium, 31 March–3 April 2008.

[36] J. Chai, G. An, Z. MA, and X. Sun, "A study of fault statistical analysis and maintenance policy of wind turbine system," In International Conference on Renewable Power Generation (RPG 2015); Institution of Engineering and Technology: Stevenage, UK, 2015; p. 4

[37] "Estimation of Turbine Reliability Figures within the DOWEC Project. 2002," Available online: http://autodocbox.com/Electric_Vehicle/87260830-Estimation-of-turbine-reliability-figures-within-the-dowec-project.html , Nov. 2021.

[38] N. E. Carlstedt, "Driftuppföljning av Vindkraftverk: Arsrapport" 2012: >50 kW. 2013. Available online: http://www.vindstat.nu/stat/Reports/arsrapp2012.pdf , Nov. 2021.

[39] G. J. Herbert, S. Iniyan, and R. Goic, "Performance, reliability and failure analysis of wind farm in a developing Country," Renew. Energy 2010, 35, 2739–2751.

[40] Matumiya et al., "Commitee for Increase in Availability/Capacity Factor of Wind Turbine Generator System and Failure/Breakdown Investigation of Wind Turbine Generator Systems," Summary Report, New Energy Industrial Technology Development Organization: Kanagawa, Japan, 2004.

[41] S. Faulstich, M. Dursterwitz, B. Hahn, K. Knorr, and K. Rohrig, "Wind energy Report Germany 2008: Written within the Research Project Deutscher Wind monitor," German Federal Ministry for the Environment Nature Conversation and Nuclear Safety: Bonn, Germany, 2009.

[42] C. I. Crabtree, D. Zappalà, and S. I. Hogg, "Wind energy: UK experiences and offshore operational challenges," .Proceedings of the Institution of Mechanical Engineers, Part A: Journal of Power and Energy. 2015; 229(7): 727-746. doi:10.1177/0957650915597560

[43] C. Su, Y. Yang, X. Wang, and Z. Hu, "Failures analysis of wind turbines: Case study of a Chinese wind farm," 2016 Prognostics and System Health Management Conference (PHM-Chengdu), Chengdu, 2016, pp. 1-6.

[44] B. Hahn, M. Durstewitz, and K. Rohrig, "Reliability of wind turbines - Experience of 15 years with 1 500 WTs," Wind Energy: Proceedings of the Euromech Colloquium, pp. 3 29-332, Springer-Verlag, Berlin.

[45] F. Spinato, P. J. Tavner, G. J. W. V. Bussel, and E. Koutoulakos, "Reliability of wind turbine subassemblies," in IET Renewable Power Generation, vol. 3, no. 4, pp. 387-401, Dec, 2009.

[46] S. Ozturk, V. Fthenakis, and S. Faulstich, "Failure Modes, Effects and Criticality Analysis for Wind Turbines Considering Climatic Regions and Comparing Geared and Direct Drive Wind Turbines," Energies 2018, 11, no.9, 2317. Doi: 10.3390/en1109231.

[47] J. Carroll, A. McDonald, and D. McMillan, "Reliability Comparison of Wind Turbines With DFIG and PMG Drive Trains," in IEEE Transactions on Energy Conversion, vol. 30, no. 2, pp. 663-670, June 2015, doi: 10.1109/TEC.2014.236724.

[48] M. Boettcher, J. Reese, and F. W. Fuchs, "Reliability comparison of fault-tolerant 3L-NPC based converter topologies for application in wind turbine systems," IECON 2013 - 39th Annual Conference of the IEEE Industrial Electronics Society, Vienna, 2013, pp. 1223-1229.

[49] H. Ben Abdelghani, A. Bennani Ben Abdelghani, F. Richardeau, J. Blaquière, F. Mosser, and I. Slama-Belkhodja, "Fault tolerant-topology and controls for a three-level hybrid neutral point clamped-flying capacitor converter," in IET Power Electronics, vol. 9, no. 12, pp. 2350-2359, May.2016.

[50] W. Chen, E. Hotchkiss, and A. Bazzi, "Reconfiguration of NPC multilevel inverters to mitigate short circuit faults using back-to-back switches," in CPSS Transactions on Power Electronics and Applications, vol. 3, no. 1, pp. 46-55, Mar 2018.

[51] A. Gaillard, P. Poure, and S. Saadate, "Reconfigurable control and converter topology for wind energy conversion systems with switch failure fault tolerance capability," 2009 IEEE Energy Conversion Congress and Exposition, San Jose, CA, 2009, pp. 390-397.

[52] N. M. A. Freire and A. J. M. Cardoso, "A Fault-Tolerant Direct Controlled PMSG Drive for Wind Energy Conversion Systems," in IEEE Transactions on Industrial Electronics, vol. 61, no. 2, pp. 821-834, Feb. 2014.

[53] G. Chen and X. Cai, "Reconfigurable Control for Fault-Tolerant of Parallel Converters in PMSG Wind Energy Conversion System," in IEEE Transactions on Sustainable Energy, vol. 10, no. 2, pp. 604-614, April 2019.

[54] A. Stabile, J. O. Estima, C. Boccaletti, and A. J. Marques Cardoso, "Converter Power Loss Analysis in a Fault-Tolerant Permanent-Magnet Synchronous Motor Drive," in IEEE Transactions on Industrial Electronics, vol. 62, no. 3, pp. 1984-1996, March 2015.

[55] I. Jlassi and A. J. M. Cardoso, "Fault-Tolerant Back-to-Back Converter for Direct-Drive PMSG Wind Turbines Using Direct Torque and Power Control Techniques," in IEEE Transactions on Power Electronics, vol. 34, no. 11, pp. 11215-11227, Nov. 2019.

[56] S. Xu et al. "Fault-Tolerant Control of ANPC Three-Level Inverter Based on Order-Reduction Optimal Control Strategy under Multi-Device Open-Circuit Fault," Sci Rep **7,** 14447 (2017), doi: 10.1038/s41598-017-15000-9.

[57] R. Katebi, J. He, and N. Weise, "An Advanced Three-Level Active Neutral-Point-Clamped Converter With Improved Fault-Tolerant Capabilities," in IEEE Transactions on Power Electronics, vol. 33, no. 8, pp. 6897-6909, Aug. 2018.

[58] S. Li and L. Xu, "Strategies of fault tolerant operation for three-level PWM inverters," in IEEE Transactions on Power Electronics, vol. 21, no. 4, pp. 933-940, July 2006.

[59] S. Ceballos, J. Pou, J. Zaragoza, E. Robles, J. L. Villate, and J. L. Martin, "Soft-Switching Topology for a Fault-Tolerant Neutral-Point-Clamped Converter," 2007 IEEE International Symposium on Industrial Electronics, Vigo, 2007, pp. 3186-3191.

[60] U. Choi, K. Lee, and F. Blaabjerg, "Diagnosis and tolerant strategy of an open-switch fault for t-type threelevel inverter systems," IEEE Trans. Ind. Appl., vol. 50, no. 1, pp. 495–508, Jan./Feb. 2014.

[61] A. L. de Lacerda and E. R. C. da Silva, "Study of failures in a three-phase active neutral point clamped rectifier: Short-circuit and open-circuit faults," 2015 IEEE Energy Conversion Congress and Exposition (ECCE), Montreal, QC, 2015, pp. 4773-4780.

[62] A. V. Rocha et al., "A new fault-tolerant realization of the active three-level NPC converter," 2014 IEEE Energy Conversion Congress and Exposition (ECCE), Pittsburgh, PA, 2014, pp. 3483-3490.

[63] J. Li, A. Q. Huang, Z. Liang, and S. Bhattacharya, "Analysis and design of active NPC (ANPC) inverters for fault-tolerant operation of high-power electrical drives," IEEE Trans. Power Electron., vol. 27, no. 2, pp. 519– 533, Feb. 2012.

E-learning Personalization in Midwifery and Maritime: A Machine Learning Approach for Intelligent Recommender Systems

Alexandros Bousdekis
Department of Business Administration
School of Business, Athens University of Economics and Business
Athens, Greece
e-mail: albous@mail.ntua.gr

Stavroula Barbounaki
Merchant Marine Academy of Aspropyrgos
Aspropyrgos, Greece
e-mail: sbarbounaki@yahoo.gr

Stefanos I. Karnavas
Merchant Marine Academy of Oinousses
Oinousses, Greece
e-mail: stefkarnavas@gmail.com

***Abstract*—The lockdown due to the pandemic of COVID-19 led to an unprecedented impact on education. Higher education institutions were forced to shift rapidly to distance and online learning. On the one hand, this fact revealed the weaknesses of adoption and utilization of e-learning strategies and technologies, but, on the other hand, it resulted in a digital revolution in education. However, the wide adoption of e-learning strategies and technologies and the complete transformation of the physical learning process to a virtual one pose the challenge of personalization of the learning process. This paper proposes a recommender system for supporting the professors in higher education of midwifery and maritime in understanding their students' needs so that he/she adapts the e-learning process accordingly. To do this, it utilizes learning profile theory and it implements k-means clustering and Bayesian Networks (BN) The proposed approach was applied to a maritime educational institution.**

Keywords-learning profiles; learning styles; higher education; k-means clustering, Bayesian network; classification.

I. INTRODUCTION

In this paper, we propose a recommender system for supporting the professors in higher education of midwifery and maritime in understanding their students' needs so that he/she adapts the e-learning process accordingly. This research work extends our previous work [1] in the following directions: (i) We incorporated X-means clustering for dynamically creating the clusters corresponding to learning profiles; (ii) we enriched the model with the learning profiles; (iii) we applied our proposed approach in an additional higher education institution for further validation.

According to a European Commission's report on digital skills in education in 2013, an average of 65% of students in EU countries never used digital textbooks, exercise software, broadcasts/podcasts, simulations or learning games [2]. Since then, higher education institutions have shown a persistent concern with enhancing students' academic performance through the use of innovative technologies that offer new ways of delivering and producing university education [3]. From an economic point of view, the industry of e-learning has developed considerably in the last decade. The market of e-learning all over the world will be over 243 billion dollars in 2022 [4].

The pandemic of COVID-19 led most of the governments around the world to impose lockdown, social/physical distancing, avoiding face-to-face teaching-learning, and restrictions on travelling and immigration [4]. It caused the closing of classrooms all over the world and forced 1.5 billion students and 63 million educators to suddenly modify their face-to-face academic practices [4]. This closure led to an unprecedented impact on education. Higher education institutions were forced to shift rapidly to distance and online learning. On the one hand, this fact revealed the weaknesses of adoption and utilization of e-learning strategies and technologies [5] [6]; but, on the other hand, it resulted in a digital revolution through online lectures, teleconferencing, digital open books, online examination, and interaction at virtual environments [7].

E-learning is the use of new multimedia technologies and the Internet to improve the quality of learning by facilitating access to resources and services, as well as remote exchange and collaboration [8] [9]. It has a great potential from the educational perspective and it has been one of the main research lines of educational technology in the last decades [4]. Particular attention has been given on understanding the adoption factors related to e-learning services satisfaction and acceptance by students and tutors [6] [8] [10].

However, the wide use of e-learning due to COVID-19 demonstrated inequalities as a result of previously underestimating the potential of e-learning and its exclusion from the digital education projects of educational institutions [4]. A considerable amount of literature has investigated inequalities between developed and developing countries [3] [11]. However, the wide adoption of e-learning strategies and technologies and the complete transformation of the physical learning process to a virtual one pose the challenge of personalization according to different learning profiles [12], a research area rather underexplored. E-learning provides people with a flexible way to learn allowing learning on demand and reducing the associated costs [8]. E-learning personalization is emerged as a major challenge [12] [13], especially in today's fast adoption of this alternative way of learning.

Despite the large amount of research works dealing with learning profiles in physical classrooms, these models should be further investigated and validated in the virtual classrooms, during the e-learning process. To this end, the contribution of e-learning to several learning factors according to the learning profiles has the potential to reveal the acceptance of e-learning by different learning profiles

and to result in e-learning process personalization in order to mitigate the respective inequalities.

The objective of the current paper is to develop an intelligent recommender system for supporting the professors in higher education in understanding their students' needs so that he/she adapts the e-learning process accordingly. In addition, the proposed recommender system is able to classify new records (i.e., students) to the appropriate learning profiles, e.g., in order to support the organization of the class groups. The proposed approach was applied to a maritime educational institution. The rest of the paper is organized as follows: Section II presents the related work on methods and approaches for evaluating students' acceptance of the e-learning process as well as learning profile models for learning personalization. Section III describes the research methodology and the proposed approach for the development of an intelligent recommender system for e-learning process personalization. Section IV presents the results from the adoption of the proposed methodology in the maritime and the midwifery education. Section V discusses the results and the implications of the proposed methodology. Section VI concludes the paper and outlines our plans for future work.

II. Related Work

In this section, we present the related work in order to present the current status in the literature and to identify the challenges and the research gaps on e-learning personalization. Section II.A reviews related research works on e-learning acceptance, and Section II.B reviews works related to learning personalization with a focus on learning profiles and their applicability to the e-learning process.

A. E-learning Acceptance

Existing literature is quite rich on evaluating students' experience, satisfaction and acceptance of the e-learning process. In general, earlier studies focused more on content, customization and technology, while more recent studies focused on students' attitude and interaction, expectations, acceptance and satisfaction [10] [14]. To this end, there is an emerging trend towards the identification of the key factors for the adoption of e-learning strategies and technologies.

Several studies have used the original version of the classic model, the DeLone & McLean (D&M) IS Success Model [15] to measure and evaluate the success of e-learning systems [16]-[18]. Holsapple and Lee-Post [16] introduced the E-Learning Success Model, which posits that the overall success of an e-learning initiative depends on the attainment of success at each of the three stages of e-learning systems development: system design, system delivery, and system outcome. Lin and Lee [17] proposed a research model to examine the determinants for successful use of online communities based on structural equation modelling (SEM) approach. The analytical results showed, among others, that system quality, information quality and service quality had a significant effect on member loyalty through user satisfaction and behavioural intention to use the online community. Lin [18] examined the determinants for successful use of online learning systems. The results showed that system quality, information quality, and service quality had significant effects on user satisfaction.

The use of virtual learning environments in addition to classroom study (blended learning), were surveyed by [19]. They concluded that the students' performance of the virtual learning environment support had better results than those having only face to face learning. The identified key satisfaction factors are information quality, system quality, instructor attitude toward e-learning, diversity in assessment, and learner perceived interaction with others. The authors in [8] identified clear governance structure and the need of organized distribution of planning responsibilities and implementation as the main adoption factors. In [20], the authors concluded that perceived usefulness, ease of use, perceived enjoyment, network externality factor, system factor, individual factors, and social factors are the main e-learning acceptance predictors. Student interface, learning community, content, and customization as well as ease of use of web courses have also been identified to have a significant impact on e-learning acceptance [21] [22].

In [23], the authors concluded that student e-learning adoption and attitudes in the university context are academic achievements mediated by digital readiness and academic engagement. In [24], the authors proposed an e-learning tools acceptance model in order to examine the level of acceptance and critical factors of virtual learning tools among university students in developing countries. Results confirm a strong relation between the perceived usefulness and the instructor preparation and autonomy in learning, as well as between the ease of use and the perceived self-efficacy perception. The reseach work of [25] developed a Technology Acceptance Model (TAM) for e-learning. The results indicated that system quality, computer self-efficacy, and computer playfulness have a significant impact on perceived ease of use of e-learning system. Furthermore, information quality, perceived enjoyment, and accessibility were found to have a positive influence on perceived ease of use and perceived usefulness of e-learning system.

The authors in [26] applied process mining methods in order to discover students' self-regulated learning processes during e-learning. They identified a high presence of actions related to forum-supported collaborative learning among the students who finally passed the exams and an absence of those in their failing classmates. The research work of [6] concluded that the main factors affecting the usage of e-learning are: technological factors, e-learning system quality factors, trust factors, self-efficacy factors and cultural aspects. Therefore, apart from the challenges related to the technological infrastructure, change management, course design, computer self-efficacy and financial support are also issues of outmost importance.

B. Learning Personalization

Learning personalization is an important topic in educational sciences. Since different people learn in different ways, it is important to create and adapt the e-learning in order to maximize and speed up the learning process [12]. The need to adapt teaching strategies to the student's preferences is a reality in classrooms, be they physical or

virtual [27] [28]. However, this does not mean that a method should be created for each student in a classroom, but that the best form of interaction for each of them be identified, building groups of learners with common characteristics [29]. Learning styles are cognitive, affective and psychological traits that determine how a student interacts and reacts in a learning environment [30]. The idea is to identify the marked characteristics of a given learner so that these traits influence his learning process.

Several learning profile models have been developed in the literature, such as the Myers-Briggs Type Indicator – MBTI, Kolb's Experiential Learning Model, the Hermann Brain Dominance Instrument (HBDI), the Dunn and Dunn Model, the Felder-Silverman Model, and the Honey and Mumford Model [27] [28]. With the wide adoption of e-learning strategies and technologies, there is the need for applying and validating learning profile models in the digital and online learning era. For example, in [12], the authors investigated the e-learning personalization aiming at keeping students motivated and engaged. To that end, they proposed the use of k-means algorithm to cluster students based on 12 engagement metrics divided into two categories: interaction-related and effort-related. The research work of [28] presented the architecture of a system that realizes an evaluation of learning profiles based on categories of student preferences. The profile models were built according to categories of student preferences based on the proposal of learning styles put forward by [30].

III. Research Methodology

In this section, we present the adopted research methodology for e-learning personalization forming the basis for the development of an associated intelligent recommender system. The methodology consists of six steps that are described in the following sub-sections: (A) Data Collection; (b) Learning Profile Model Selection; (C) Classification for Structuring the Learning Profiles; (D) Validation of Learning Profiles Classification; (E) Modelling the Relationships between Learning Profiles and E-learning Preferences; and, (F) Predicting the Class Attribute of E-learning Impact.

A. Data Collection

The data was collected in the form of an online questionnaire of 80 questions addressed to students of higher educational institutions. Each question was in the form of Likert scale (1: Strongly Disagree – 5: Strongly Agree) and it was related to one out of the four learning styles as defined by the Honey and Mumford Model [31]: *activist*, *reflector*, *pragmatist*, and *theorist*. For example, in an ideal scenario that a student has answered 5: Strongly Agree to all the questions matching to the "activist" learning profile and 1: Strongly Disagree to all the others, he/she is classified as "activist".

B. Learning Profile Model Selection

As it was mentioned, the selected learning profile model is the Honey and Mumford Model [31], which includes four learning styles: activist, reflector, pragmatist, and theorist. However, this classification is usually not straightforward since most of the students belong to a mixture of learning styles, meaning that they incorporate characteristics from more than one profile [31]. The main characteristics of these four learning profiles are described below [32].

Activist refers to an individual's preference for active involvement in the learning activity (through problem solving, discussion, creating their own models). Activists are enjoy new experiences and are not willing to participate in repeated tasks. They prefer brainstorming as a format of discussion. Therefore, the teaching and learning activities that are effective for this group need to provide new experiences, problem-based learning, games and group research. The teaching and learning activities that are not effective for this group are one-way lecture, passive learning, learning that involves many mixed and unarranged data, repeating the same activity.

Reflector prefers learning by watching and thinking. The reflector responds more positively to learning activities where there is time to observe, reflect and think and work in a detailed manner. Reflectors like to collect and analyse data and are careful at making decisions. They do not like to become leaders. The teaching and learning activities that are effective for this group need to be stimulating and to provide them with time to think before reacting and to provide conclusions without pressure. The teaching and learning activities that are not effective for this group are placing them in the role of leader or having them perform in front of people. They experience stress if required to perform immediately after a brief instruction

Pragmatist wants to know how to put what they are learning into practice in the real world. They experiment with theories, ideas, and techniques and take the time to think about how what they've done relates to reality. Pragmatists prefer to come up with new ideas, and solving problems especially for real life situations. The teaching and learning activities that are effective for this group are demonstrating practical techniques, providing them with the opportunity to express what they have learned and focusing on the practical issues. Learning methods that are not related to immediate need and performance with no clear practice or outline are not suitable for this group.

Theorist seeks to understand the theory behind the action. They follow models and reading up on facts to better engage in the learning process. Theorists are quite objective, and they do not enjoy things that are subjective. They prefer to make conclusions based on evidence, data analysis and logic. They have clear minds. The teaching and learning activities that are effective for this group are providing them with time to organise their feelings and to ask questions and process the methodology, assumption or logic in detail. The teaching and learning activities that are not effective for this group are learning that involves emotion, feelings, and activities that are unstructured.

C. Classification for Structuring the Learning Profiles

The classification of the student to the learning profiles is not straightforward since they may have characteristics of

more than one profile. Therefore, according to the given answers, the k-means clustering algorithm was applied in order to assign the respondents to 4 clusters ($k=4$) matching to the aforementioned learning profiles.

k-means clustering is a method of vector quantization that aims to partition n observations into k clusters in which each observation belongs to the cluster with the nearest mean (cluster centers or cluster centroid) [33]. k-means clustering minimizes the within-cluster variances (squared Euclidean distances). Given a set of observations $(\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n)$, where each observation is a d-dimensional real vector, k-means clustering aims to partition the n observations into k ($\leq n$) sets $\boldsymbol{S}=\{S_1, S_2, \ldots, S_k\}$ so as to minimize the within-cluster sum of squares (WCSS) (i.e., variance). Formally, the objective is to find:

$$\arg\min_{\mathbf{S}} \sum_{i=1}^{k} \sum_{\mathbf{x}\in S_i} \|\mathbf{x}-\boldsymbol{\mu}_i\|^2 = \arg\min_{\mathbf{S}} \sum_{i=1}^{k} |S_i| \operatorname{Var} S_i \qquad (1)$$

where $\boldsymbol{\mu}_i$ is the mean of points in S_i. This is equivalent to minimizing the pairwise squared deviations of points in the same cluster:

$$\arg\min_{\mathbf{S}} \sum_{i=1}^{k} \frac{1}{2|S_i|} \sum_{\mathbf{x},\mathbf{y}\in S_i} \|\mathbf{x}-\mathbf{y}\|^2 \qquad (2)$$

The equivalence can be deduced from the identity:

$$\sum_{\mathbf{x}\in S_i} \|\mathbf{x}-\boldsymbol{\mu}_i\|^2 = \sum_{\mathbf{x}\neq\mathbf{y}\in S_i} (\mathbf{x}-\boldsymbol{\mu}_i)(\boldsymbol{\mu}_i-\mathbf{y}). \qquad (3)$$

Because the total variance is constant, this is equivalent to maximizing the sum of squared deviations between points in different clusters (Between-Cluster Sum of Squares, BCSS), which follows from the law of total variance.

D. Validation of the Learning Profiles Classification

As it has been already mentioned, students may belong to a mixture of learning styles, in the sense that that they may incorporate characteristics from more than one profile. In order to validate the k-means clustering results of the previous step, i.e., the classification to the aforementioned 4 discrete learning profiles, we implement the X-means clustering algorithm.

In contrast to the k-means clustering algorithm which requires the number of clusters k to be supplied by the user and its search is prone to local minima, X-means searches the space of cluster locations and number of clusters to optimize the Bayesian Information Criterion (BIC) [34][35]. In this way, it can identify additional clusters representing mixtures of learning profiles. After comparing these results with the ones derived from the k-means clustering algorithm of the previous step, the domain expert is able to validate the applicability of Honey and Mumford Model according to the similarity of the two resulting sets of clusters. Moreover, they are able to select whether they will be based upon a pre-defined model of learning styles or they will create dynamically learning styles in order to tackle with the mixtures of learning profiles that most often exist in reality.

The X-means clustering algorithm starts with k equal to the lower bound of the given range and adds centroids until the upper bound is reached. During this process, the centroid set that achieves the best score is recorded. The algorithm consists of two operations repeated until completion:

- ***Improve Parameters***: which runs conventional k-means to converge.
- ***Improve Structure***: which finds out if and where new centroids should appear by splitting centroids. In this operation, the advantages of two splitting approaches are combined: (i) One at a time: picking one centroid, producing a new centroid nearby, and running k-means to completion; (ii) Half the centroids: Gaussian mixture model identification and heuristics criteria for assessing the usefulness of splitting.

The selection of the k values is performed according to the BIC. Given the data D and a family of alternative models M_j, X-means adopts the posterior probability $P(M_j|D)$ to score the models derived from k-means clustering. To approximate the posterior probabilities until normalization, the following formula is used [36][37]:

$$BIC(M_j) = \hat{l}_j(D) - \frac{p_j}{2}\log R \qquad (4)$$

where $\hat{l}_j(D)$ is the log-likelihood of the data according to the j-th model and taken at the maximum-likelihood point, and p_j is the number of parameters in M_j. This is also known as Schwarz criterion.

The maximum likelihood estimation for the variance, under the identical spherical Gaussian assumption is:

$$\hat{\sigma}^2 = \frac{1}{R-K}\sum_i \left(x_i - \mu_{(i)}\right)^2 \qquad (5)$$

The point probabilities are:

$$\hat{P}(x_i) = \frac{R_{(i)}}{R}\frac{1}{\sqrt{2\pi}\hat{\sigma}^M} e^{-\frac{1}{2\hat{\sigma}^2}\|x_i-\mu_{(i)}\|^2} \qquad (6)$$

The log-likelihood of the data is:

$$l(D) = \log\prod_i P(x_i) = \sum_i \left(\log\frac{1}{\sqrt{2\pi}\sigma^M} - \frac{1}{2\sigma^2}\|x_i-\mu_{(i)}\|^2 + \log\frac{R_{(i)}}{R}\right) \qquad (7)$$

Focusing on the set D_n which belong to centroid n and plugging in the maximum-likelihood estimates yield:

$$\hat{l}(D_n) = -\frac{R_n}{2}\log 2\pi - \frac{R_n M}{2}\log(\hat{\sigma}^2) - \frac{R_n - K}{2} + R_n\log R_n - R_n\log R \qquad (8)$$

The BIC is used both globally, when X-means selects the best model, and locally, in all the centroid split tests. Finally,

the algorithm generates the number of the clusters X, corresponding to the unsupervised set of learning profiles, as well as the students assigned to each cluster. Comparing these results to the ones derived from the k-means clustering with the pre-defined 4 learning profiles, the consistency of the Honey and Mumford Model is evaluated in a data-driven way.

E. Modelling the Relationships between Learning Profiles and E-learning Preferences

Subsequently, the proposed approach models the relationships between the learning profiles and e-learning contribution to learning factors as derived from the questionnaire. The input to this step is typically the outcome of the step "C: Classification for structuring the learning profiles" provided that the step "D: Validation of the learning profiles classification" provides acceptable results. Alternatively, the user of the recommender system may prefer to use the outputs of the X-means clustering algorithm, instead of the ones of the k-means, in order to apply an unsupervised approach. The latter is especially useful when the X-means algorithm does not approach the results of the k-means.

To do this, a Bayesian Network (BN) is applied aiming at identifying these causal and uncertain relationships. A BN, also known as belief network, is defined as a pair $B = (G, \Theta)$. $G = (V, E)$ is a Directed Acyclic Graph (DAG) where $V = \{v_1, \ldots, v_n\}$ is a collection of n nodes, $E \subset V \times V$ a collection of edges and a set of parameters Θ containing all the Conditional Probabilities (CP) of the network [38]. Each node $v \in V$ of the graph represents a random variable X_V with a state space $\boldsymbol{X}_V$ which can be either discrete or continuous. An edge $(v_i, v_j) \in E$ represents the conditional dependence between two nodes $v_i, v_j \in V$ where v_i is the parent of child v_j. If two nodes are not connected by an edge, they are conditional independent. Because a node can have more than one parent, let π_v the set of parents for a node $v \in V$.

Therefore, each random variable is independent of all nodes $V \setminus \pi_v$. For each node, a Conditional Probability Table (CPT) contains the CP distribution with parameters $\theta_{xi|\pi i} := P(x_i|\pi_i) \in \Theta$ for each realization x_i of X_i conditioned on π_i. The joint probability distribution over V is visualized by the BN and can be defined as:

$$P(X_1, \ldots, X_n) = \prod_{i=1}^{n} P(X_i|\pi_i) \quad (9)$$

With BN, inference for what-if analysis can be supported, either top-down (predictive support) or bottom-up (diagnostic support). If a random variable which is represented by a node is observed, the node is called an evidence node; otherwise, it is a hidden node [39]. Based on the learning profiles derived from the questionnaire, a BN with two layers was developed: at the top layer (i.e., learning profiles), there are 4 parent nodes matching to the respective clusters of students.

At the bottom layer (i.e., e-learning contribution to learning factors), there are 9 child nodes referring to 9 e-learning factors grouping the questions. In this way, the model identifies the preferences of each learning profile by assessing the impact of e-learning on the learning process of each profile. Therefore, according to the learning profile, the user is able to select the appropriate learning strategies aiming at personalizing the e-learning process.

The e-learning factors are constructed based on the groupping of the various questions of the questionnaire. Below, we describe their meaning:

F1: Comprehension: the level of comprehension of the course content with e-learning.

F2: Content digestion: the satisfaction by the content presented in comparison to the contents of the course.

F3: Knowledge generalization: the capability of understanding practical examples and how they support the overall concepts and theories.

F4: Guidance and monitoring: The level to which the interaction between the tutor and the students facilitates guidance and monitoring of students' preformance.

F5: Tutor communication: the level to which the communication between the tutor and the students is efficient.

F6: Active participation: The level to which e-learning enables the active participation of the students, e.g. by posing questions, participating to discussions, etc.

F7 Effective learning: The level to which the learning procedure is performed in an effective way.

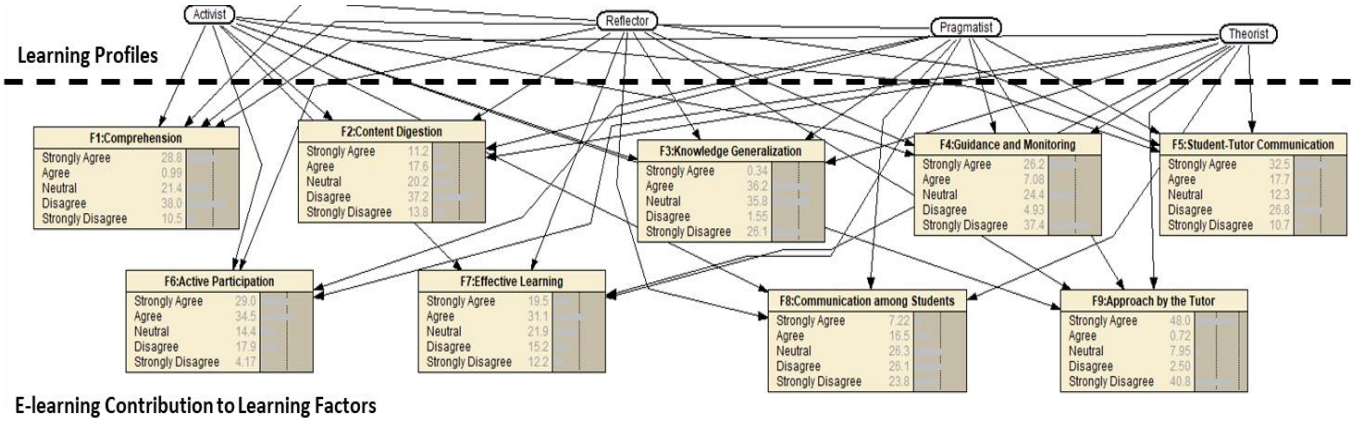


Figure 1. The Bayesian Network structure for modelling the relationships between learning profiles and e-learning contribution to learning factors for the maritime institution.

F8 Communication among students: The degree to which e-learning includes discussions among students and teamwork.

F9 Approach by the tutor: the extent to which the tutor applies pedagogical approaches that are personalized to the learning needs of the students.

F. Predicting the Class Attribute of E-learning Impact

At any time, the user of the recommender system is able to make queries in order to investigate particular relationships along with their associated CPTs. Moreover, the model incorporates a Naïve Bayes classifier for predicting the class attribute of a learning profile as soon as new records of students' responses are inserted into the database.

Naïve Bayes classifier is highly scalable, requiring a number of parameters linear in the number of variables in a learning problem. Maximum-likelihood training can be done by evaluating a closed-form expression, which takes linear time, rather than by expensive iterative approximation as used for many other types of classifiers [40]. Prediction of the class attribute can be performed even if the questionnaire is not completely answered.

IV. RESULTS

The proposed approach was applied on a dataset of 524 students: 268 from a maritime higher educational institution and 256 students of a midwifery department in a University in Greece. Following the research methodology described in Section III, the data was analyzed both as separate cases, one case in maritime students and one case in midwifery students, and as a whole. The implementation and execution of the experiments were performed using the sklearn.cluster library of Python [44] for the k-means clustering algorithm and the BN functionalities of the pgmpy (Probabilistic Graphical Models using Python) package [45].

Section IV.A presents the results from the dataset of maritime students. Section IV.B presents the results from the dataset of the midwifery students.

A. Dataset from Maritime Institution

The transformation of maritime from highly labour- to capital-intensive industry contributed to the presence of tertiary education in maritime studies [41]. However, the learning process in maritime education faces additional challenges due to the structure of their programs, the tendency of undergraduate students to combine studies and work, the internationalization, specialization, and standardization [42][43]. These make maritime education an interesting case study for the validation of e-learning process personalization.

TABLE I. CPs OF THE E-LEARNING CONTRIBUTION TO LEARNING FACTORS GIVEN THE LEARNING PROFILES IN MARITIME

	E-learning contribution	Learning profile	CP
Highest CPs	F1={Neutral}, F2={Agree}, F3={Disagree}, F4={Agree}, F5={Strongly Disagree}, F6={Disagree}, F7={Neutral}, F8={Strongly Disagree}, F9={Agree}	Activist	0.386
	F1={Disagree}, F2={Disagree}, F3={Agree}, F4={Strongly Disagree}, F5={ Disagree}, F6={Agree}, F7={Neutral}, F8={Neutral}, F9={Disagree}	Theorist	0.295
Lowest CPs	F1={Stongly Agree}, F2={Disagree}, F3={Strongly Agree}, F4={Neutral}, F5={ Disagree}, F6={Strongly Disagree}, F7={Neutral}, F8={Strongly Disagree}, F9={Disagree}	Reflector	0.081
	F1={Agree}, F2={Strongly Disagree}, F3={Agree}, F4={Strongly Disagree}, F5={Strongly Agree}, F6={Neutral}, F7={Agree}, F8={Agree}, F9={Strongly Disagree}	Activist	0.056

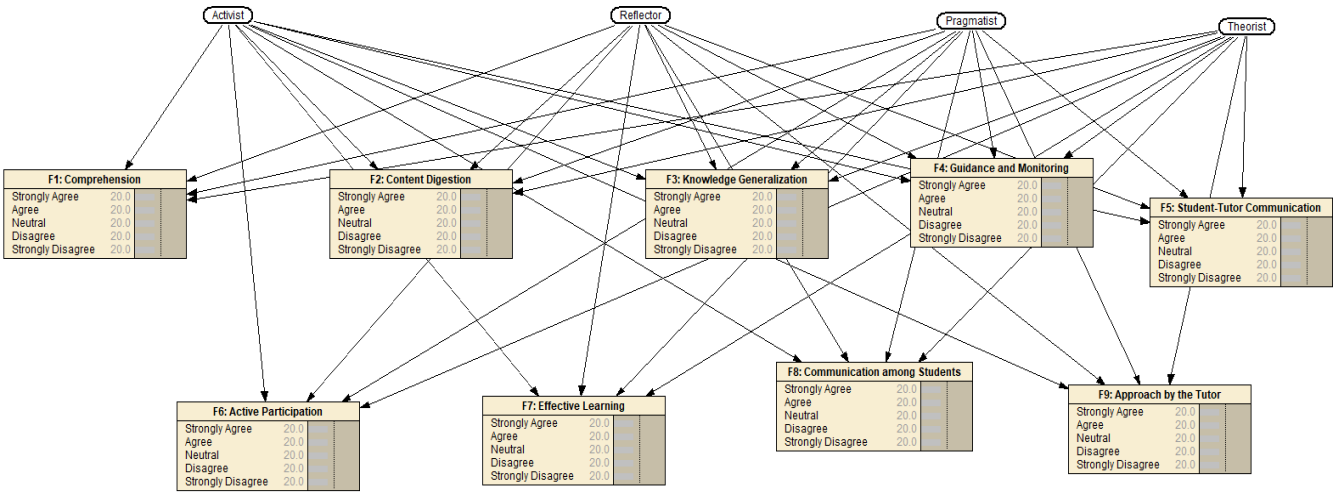


Figure 2. The Bayesian Network structure for modelling the relationships between learning profiles and e-learning contribution to learning factors for the midwifery institution.

After having structured the learning profiles of the respondents, the BN is created and the CPTs are calculated, as shown in Figure 1. Table I presents the highest and the lowest CPs of the e-learning contribution to learning factors given the learning profiles for the maritime institution. Therefore, the highest CP is the one of a student being activist given the answers of the second row that is 38.6%. The lowest CP is the one of a student being activist given the answers of the last row that is 5.6%. According to the queries posed by the user, various calculations can be done. As already mentioned, the model can also serve as a classifier for predicting the class attribute of learning factors as soon as new records of students are received and classified through the k-means clustering algorithm.

In order to evaluate its classification effectiveness, we inserted additional records, derived from more questionnaires addressed to students of the maritime educational institution, and we created the confusion matrix according to Table II in order to estimate the precision and the recall of the classifier using (5) and (6) [46].

The Precision results are quite satisfactory, while the Recall results can be further improved. We should also take into account that modelling human behavior, such as the learning process, has a high degree of uncertainty [47]. Moreover, the BN model sticks to the initially identified relationships, i.e., the ones that have been mined during the model training. Therefore, when new relationships, not previously identified, are added, they are not classified correctly. These records include values that are not frequent, so they are not critical for decision making

TABLE II. CONFUSION MATRIX

	Predicted Positive	Predicted Negative
Actual Positive	True Positive (TP) = 31	False Negative (FN) = 6
Actual Negative	False Positive (FP) = 4	True Negative (TN) = 22

$$Precision = \frac{TP}{TP + FP} = \frac{31}{31 + 4} = 88.57\% \quad (10)$$

$$Recall = \frac{TP}{TP + FN} = \frac{31}{31 + 6} = 83.78\% \quad (11)$$

B. *Dataset from Midwifery Institution*

Evidence within higher education clearly identifies that the academic and personal development of students is enhanced by engagement with the academic and non-academic life of college [48][49]. It is necessary for the development of important capabilities including critical thinking, problem-solving, team work, and written and oral communication skills, all of which are essential midwifery graduate competencies for practice in modern, dynamic and complex healthcare services [50].

After having structured the learning profiles of the respondents, the BN is created and the CPTs are calculated, as shown in Figure 2. Table III presents the highest and the lowest CPs of the e-learning contribution to learning factors given the learning profiles for the midwifery institution. Therefore, the highest CP is the one of a student being pragmatist given the answers of the second row that is 37.8%.

The lowest CP is the one of a student being activist given the answers of the last row that is 6.2%. According to the queries posed by the user, various calculations can be done. As already mentioned, the model can also serve as a classifier for predicting the class attribute of learning factors as soon as new records of students are received and classified through the k-means clustering algorithm. The confusion matrix is presented in Table IV.

TABLE III. CPS OF THE E-LEARNING CONTRIBUTION TO LEARNING FACTORS GIVEN THE LEARNING PROFILES IN MIDWIFERY

	E-learning contribution	Learning profile	CP
Highest CPs	F1={Disagree}, F2={Agree}, F3={ Neutral }, F4={Agree}, F5={Strongly Agree}, F6={Neutral}, F7={Agree}, F8={ Disagree}, F9={Neutral}	Pragmatist	0.378
	F1={Agree}, F2={Neutral}, F3={Disagree}, F4={Neutral}, F5={ Agree}, F6={Strongly Agree}, F7={Disagree}, F8={Neutral}, F9={Agree}	Reflector	0.198
Lowest CPs	F1={Disagree}, F2={Strongly Agree}, F3={Neutral}, F4={Strongly Disagree}, F5={Agree}, F6={Strongly Agree}, F7={Strongly Disagree}, F8={Strongly Agree}, F9={Strongly Disagree}	Theorist	0.078
	F1={Strongly Disagree}, F2={Agree}, F3={Agree}, F4={Disagree}, F5={Neutral}, F6={Agree}, F7={Neutral}, F8={Disagree}, F9={ Disagree}	Activist	0.062

TABLE IV. CONFUSION MATRIX

	Predicted Positive	Predicted Negative
Actual Positive	True Positive (TP) = 35	False Negative (FN) = 6
Actual Negative	False Positive (FP) = 7	True Negative (TN) = 24

$$Precision = \frac{TP}{TP + FP} = \frac{35}{35 + 7} = 83.33\% \quad (12)$$

$$Recall = \frac{TP}{TP + FN} = \frac{35}{35 + 6} = 85.37\%\% \quad (13)$$

V. DISCUSSION

E-learning provides people with a flexible way to learn allowing learning on demand and reducing the associated costs. The wide adoption of e-learning strategies and technologies and the transformation of the physical learning process to a virtual one pose the challenge of personalization according to different learning. E-learning personalization is emerged as a major challenge, especially in today's fast adoption of this alternative way of learning. The proposed

approach was proved to be effective in classifying the students' learning profiles in order to adapt the learning process and to provide personalized recommendations for the learning style. It validated in two diverse educational institutions and the results show a stability in students' classification according to their answers to the questionnaire. The proposed approach can work either with a pre-defined learning profiles model or with a learned set of learning profiles.

In the first case (i.e., pre-defined learning profiles model), the proposed approach was based on the Honey and Mumford Model, which includes four learning styles: activist, reflector, pragmatist, and theorist. Therefore, the k-means clustering algorithm takes as input k=4. This case refers to a supervised learning approach. It should be noted that the learning profiles model can be a different one. Then, X-means clustering algorithm validates the assumption about the number of the clusters that was derived from the learning profiles model. The user may define a threshold of outliers that is acceptable for the model validation. If this threshold is not exceeded, the learning profiles and the learning factors along with their values feed into the BN. The latter is able to perform predictions and to provide useful information upon user's queries.

In the second case (i.e., learned set of learning profiles), the clusters corresponding to learning profiles can be created dynamically directly by the X-means clustering algorithm. This case refers to an unsupervised learning approach in which the learning profiles are not known in advance, and the resulting clusters should be interpreted by a domain expert. Then, the BN is structured accordingly with its nodes of the upper layer corresponding to the resulting number of clusters by the X-means algorithm.

In both cases, the proposed approach takes advantage of machine learning algorithms in order to form the basis for a recommender system capable of supporting personalized teaching and learning procedures according to the students' learning profile.

VI. Conclusions and Future Work

During the last years, e-learning has been gaining an increasing attention in higher education. Especially during the last months, higher education institutions were forced to shift rapidly to distance and online learning. On the one hand, this fact revealed the weaknesses of adoption and utilization of e-learning strategies and technologies, but, on the other hand, it resulted in a digital revolution in education. A major challenge was to apply e-learning strategies and technologies for supporting e-learning personalization. In this paper, we proposed an intelligent recommender system for e-learning process personalization.

The proposed approach is based on the Honey and Mumford Model of learning profiles and utilized k-means clustering, X-means clustering, and BNs in order to classify the students to learning profiles and to reveal relationships with the contribution of e-learning to several learning factors. The proposed approach was applied to maritime and midwifery education. We validated the model in terms of its precision and recall in predicting the learning profile when new records are inserted into the database.

Regarding our future work, we plan to incorporate additional learning factors with respect to the e-learning impact. Moreover, we plan to apply more machine learning and data analytics methods, with an emphasis on fuzzy methods, in combination with different learning profile models. Finally, we will plan to expand our research to various universities in order to obtain more generalized results.

References

[1] S. Karnavas, A. Bousdekis, S. Barbounaki, and D. Kardaras, "An Intelligent Recommender System for E-learning Process Personalization: A Case Study in Maritime Education", in The Ninth International Conference on Data Analytics (DATA ANALYTICS) 2020, ThinkMind Library (pp. 84-89), 2020.

[2] D. Gubiani, I. Cristea, and T. Urbančič, "Introducing e-learning to a traditional university: a case-study." in Qualitative and Quantitative Models in Socio-Economic Systems and Social Work, pp. 225-241, Springer, Cham, 2020.

[3] H. J. Kim, A. J. Hong, and H. D. Song, "The roles of academic engagement and digital readiness in students' achievements in university e-learning environments," Int. J. of Educ. Tec. in Hig. Educ., vol. 16, no. 1, pp. 21-29, 2019.

[4] J. Valverde-Berrocoso, M. D. C. Garrido-Arroyo, C. Burgos-Videla, and M. B. Morales-Cevallos, "Trends in Educational Research about e-Learning: A Systematic Literature Review (2009–2018)," Sust., vol. 12, no 12, pp. 5153, 2020.

[5] N. Kapasia, P. Paul, A. Roy, J. Saha, A. Zaveri, R. Mallick, and P. Chouhan, "Impact of lockdown on learning status of undergraduate and postgraduate students during COVID-19 pandemic in West Bengal, India," Child. and Youth Serv. Rev., vol. 116, pp. 105194, 2020.

[6] M. A. Almaiah, A. Al-Khasawneh, and A. Althunibat, "Exploring the critical challenges and factors influencing the E-learning system usage during COVID-19 pandemic," Educ. and Inf. Tech., vol. 1, 2020.

[7] T. Gonzalez et al., "Influence of COVID-19 confinement in students performance in higher education," arXiv preprint arXiv:2004.09545, 2020.

[8] W. M. Al-Rahmi, N. Alias, M. S. Othman, A. I. Alzahrani, O. Alfarraj, A. A. Saged, and N. S. A. Rahman, "Use of e-learning by university students in Malaysian higher educational institutions: a case in Universiti Teknologi Malaysia," IEEE Acc., vol. 6, pp. 14268-14276, 2018.

[9] E. R. Vershitskaya, A. V. Mikhaylova, S. I. Gilmanshina, E. M. Dorozhkin, and V. V. Epaneshnikov, "Present-day management of universities in Russia: Prospects and challenges of e-learning," Educ. and Inf. Tech., vol. 25, no. 1, pp. 611-621, 2020.

[10] W. A. Cidral, T. Oliveira, M. Di Felice, and M. Aparicio, "E-learning success determinants: Brazilian empirical study," Comp. & Educ., vol. 122, pp. 273-290, 2018.

[11] E. Vázquez-Cano, M. León Urrutia, M. E. Parra-González, and E. López Meneses, "Analysis of interpersonal competences in the use of ICT in the Spanish University Context," Sust., vol. 12, no. 2, pp. 476, 2020.

[12] A. Moubayed, M. Injadat, A. Shami, and H. Lutfiyya, "Student engagement level in e-learning environment: Clustering using k-means," Amer. J. of Dist. Educ., pp. 1-20, 2020.

[13] J. Shailaja and R. Sridaran, "Taxonomy of e-learning challenges and an insight to blended learning," in 2014 International Conference on Intelligent Computing Applications,pp. 310-314, IEEE, 2014.

[14] H. J. Chen, (2020). "Clarifying the impact of surprise in e-learning system design based on university students with multiple learning goals orientation," Educ. and Inf. Tech., pp. 1-20, 2020

[15] W. H. Delone and E. R. McLean, "The DeLone and McLean model of information systems success: a ten-year update," J. of Man. Inf. Syst., vol. 19, no. 4, pp. 9-30, 2003.

[16] C. W. Holsapple and A. Lee-Post, "Defining, assessing, and promoting e-learning success: An information systems perspective," Dec. Sci. J. of Innov. Educ., vol. 4, no. 1, pp. 67-85, 2006.

[17] H. F. Lin and G. G. Lee, "Determinants of success for online communities: an empirical study," Beh. & Inf. Tech., vol. 25, no. 6, pp. 479-488, 2006.

[18] H. F. Lin, "Measuring online learning systems success: Applying the updated DeLone and McLean model," Cyberpsyc. & Beh., vol. 10, no. 6, pp. 817-820, 2007.

[19] D. Stricker, D.,Weibel, and B. Wissmath, "Efficient learning using a virtual learning environment in a university class," Comp. & Educ., vol. 56, no. 2, pp. 495-504, 2011.

[20] Y. M. Cheng, "Antecedents and consequences of e-learning acceptance," Inf. Sys. J., vol. 21, no. 3, pp. 269-299, 2011.

[21] Y. S. Wang, "Assessment of learner satisfaction with asynchronous electronic learning systems," Inf. & Manag., vol. 41, no. 1, pp. 75-86, 2003.

[22] H. M. Selim, "An empirical investigation of student acceptance of course websites," Comp. & Educ., vol. 40, no. 4, pp. 343-360, 2003.

[23] H. J. Kim, A. J. Hong, & H. D. Song, "The roles of academic engagement and digital readiness in students' achievements in university e-learning environments," Int. J. of Educ. Tech. in High. Educ., vol. 16, no.1, pp. 21, 2019.

[24] A. Valencia-Arias, S. Chalela-Naffah, and J. Bermúdez-Hernández, "A proposed model of e-learning tools acceptance among university students in developing countries," Educ. and Inf. Tech., vol. 24, no. 2, pp. 1057-1071, 2019.

[25] S. A. Salloum, A. Q. M. Alhamad, M. Al-Emran, A. A. Monem, and K. Shaalan, "Exploring students' acceptance of e-learning through the development of a comprehensive technology acceptance model," IEEE Acc., vol. 7, pp. 128445-128462, 2019.

[26] R. Cerezo, A. Bogarín, M. Esteban, and C. Romero, "Process mining for self-regulated learning assessment in e-learning," J. of Comp. in High. Educ., vol. 32, no. 1, pp. 74-88, 2020.

[27] C. Heaton-Shrestha, C. Gipps, P. Edirisingha, and T. Linsey, "Learning and e-learning in HE: the relationship between student learning style and VLE use," Res. Pap. in Educ., vol. 22, no. 4, pp. 443-464, 2007.

[28] L. A. Zaina and G. Bressan, "Classification of learning profile based on categories of student preferences," in 2008 38th Annual Frontiers in Education Conference, pp. F4E-1, IEEE, 2008.

[29] N. A. Fabio, J. A. Self, and S. P. Lajoie, "Modeling the process, not the product, of learning," Comp. as Cogn. Tools, vol. 2, pp.133-162, 2000.

[30] R. M. Felder, and R. Brent, "Understanding student differences," J. of Eng. Educ., vol. 94, no. 1, pp. 57-72, 2005.

[31] C. A. Lowery, "Adapting to student learning styles in a first year electrical/electronic engineering degree module," Eng. Educ., vol. 4, no. 1, pp. 52-60, 2009.

[32] R. Efe, S. Gonen, A. K. Maskan, and M. Hevedanli, "Science student teachers preferences for ways of learning: Differences and similarities," Educ. Res. Rev., vol. 6, no. 2, pp. 201-207, 2011.

[33] P. Honey and A. Mumford, "The learning styles helper's guide," Maidenhead: Peter Honey Publications, 2000.

[34] H. P. Kriegel, E. Schubert, and A. Zimek, "The (black) art of runtime evaluation: Are we comparing algorithms or implementations?," Knowl. and Inf. Sys., vol. 52, no. 2, pp. 341-378, 2017.

[35] N. Zendrato, H. W. Dhany, N. A. Siagian, and F. Izhari, "Big data Clustering using X-means method with Euclidean Distance," in Journal of Physics: Conference Series (Vol. 1566, No. 1, p. 012103). IOP Publishing, 2020.

[36] D. Pelleg and A. W. Moore, "X-means: Extending k-means with efficient estimation of the number of clusters," in ICML (Vol. 1, pp. 727-734), 2000.

[37] R. E. Kass and L. Wasserman, "A reference Bayesian test for nested hypotheses and its relationship to the Schwarz criterion," J. Americ. Stat. Ass., vol. 90, no. 431, pp. 928-934, 1995.

[38] J. Pearl, "Probabilistic reasoning in intelligent systems: networks of plausible inference," Elsevier, 2014.

[39] T. D. Nielsen and F. V. Jensen, "Bayesian networks and decision graphs," Springer Science & Business Media, 2009.

[40] T. Hastie, R. Tibshirani, and J. Friedman, "The elements of statistical learning: data mining, inference, and prediction," Springer Science & Business Media, 2009.

[41] A. A. Pallis and A. K. Ng, "Pursuing maritime education: an empirical study of students' profiles, motivations and expectations," Marit. Pol. & Manag., vol. 38, no. 4, pp. 369-393, 2011.

[42] Y. Y. Lau and A. K. Ng, "The motivations and expectations of students pursuing maritime education," WMU J. of Marit. Aff., vol. 14, no. 2, pp. 313-331, 2015.

[43] X. Chen, X., Bai, and Y. Xiao, "The application of E-learning in maritime education and training in China," TransNav: Int. J. on Mar. Nav. & Saf. of Sea Transp., vol. 11, no. 2, 2017.

[44] F. Pedregosa et al., "Scikit-learn: Machine learning in Python," J. of Mach. Learn. Res., vol. 12, pp. 2825-2830, 2011.

[45] A. Ankan and A. Panda, "pgmpy: Probabilistic graphical models using python," in Proceedings of the 14th Python in Science Conference (SCIPY 2015), Citeseer, vol. 10, 2015.

[46] C. Goutte and E. Gaussier, "A probabilistic interpretation of precision, recall and F-score, with implication for evaluation," in European conference on information retrieval, pp. 345-359, Springer, Berlin, Heidelberg, 2005.

[47] A. A. Chaaraoui, P. Climent-Pérez, and F. Flórez-Revuelta, "A review on vision techniques applied to human behaviour analysis for ambient-assisted living," Exp. Sys. with Appl., vol. 39, no. 12, pp. 10873-10888, 2012.

[48] P. Buckley and P. Lee, "The impact of extra-curricular activity on the student experience," Act. Learn. Hig. Educ., vol. 22, no. 1, pp. 37-48, 2021.

[49] E. R. Kahu and K. Nelson, "Student engagement in the educational interface: understanding the mechanisms of student success," Hig. Educ. Res. Dev., vol. 37, no. 1, pp. 58-71, 2018.

M. Clynes, A. Sheridan, and K. Frazer, "Student engagement in higher education: A cross-sectional study of nursing students' particpation in college-based education in the republic of Ireland," Nur. Educ. Tod., vol. 93, 104529, 2020.

Virtual Team Leadership and Operation in the Automotive Industry: Profile of a Research Case Study

Anatoli Quade
School of Business
University of Gloucestershire
Cheltenham, UK
E-Mail: Anatoli.quade@arcor.de

Martin Wynn
Computing and Engineering
University of Gloucestershire
Cheltenham, UK
E-Mail: MWynn@glos.ac.uk

David Dawson
School of Business
University of Gloucestershire
Cheltenham, UK
E-Mail: DDawson@glos.ac.uk

Abstract – **The impact of digitalisation on industry and society at large is huge and has parallels with the advent of the internet more than twenty years ago. In the automotive sector, companies are also confronted with the implications of the so-called megatrends of connected car, autonomous vehicles, sharing/subscription, and electrification, which are challenging current business models and working practices. This has brought about new approaches to project management practices, notably those relating to collaborating over distance between and within dispersed teams. Researchers and practitioners have started to think more comprehensively about the complexity of projects with virtual teams, and how best to manage them. This article is the result of the distillation of relevant literature relating to virtual teams and the analysis of in-depth interviews undertaken with industry experts. It puts forward a model (V-CORPS) for virtual team leadership and management. The authors believe these results can be of value in providing guidance for practitioners working in virtual teams, and as an analytical framework for further research studies in this field.**

Keywords – Project management; virtual teams; virtual leadership; German automotive industry; V-CORPS model

I. INTRODUCTION

Recent research set out a provisional model for virtual project leadership in the automotive industry [1]. Here, this model – the V-CORPS model (Virtual – Create the team; Organise the team; Relationship building; Performance evaluation; Sign-off and closure) - is developed in more detail, and particular focus is put on the underpinning research process. The model is of particular relevance because of the globalisation of the automotive industry and the dramatic changes in underlying business models and ways of working that are impacting the companies in this sector. This has brought working in virtual teams to the fore, presenting new challenges for project management, such as projects being led from a distance, with dispersed team members. This has given rise to the concept of "virtual leadership" (or "e-leadership"), which focuses on the social influencing capabilities of leaders of virtual teams.

Jugdev et al. [2] concluded that project management can be seen as a holistic discipline for achieving organisational efficiency, effectiveness, and innovation. Team leading plays a key role here. An examination of the extant literature on virtual leadership reveals issues relating to project complexity, social process, value creation, conceptualisation, and practitioner development [3]. Virtual teams face a number of issues that can impede effective project delivery – different time zones, different cultures, lack of face-to-face meetings, reduced productivity and increased miscommunication [4].

The research project reported on here had the goal of rethinking project management leadership for dispersed teams in the automotive industry, looking particularly at team leading from a distance and its influence on team members. As recently noted in the National Instruments Research Handbook [5] "within the next 10 years, we will see remarkable change in the automotive industry from improved engine efficiency to autonomous vehicles to electrification" and virtual project management will likely be of increasing importance in an industry undergoing rapid and radical change. Deloitte [6] see this as consisting of four main trends - **C**onnected car, **A**utonomous vehicles, **S**haring/subscription, and **E**lectrification - for which the acronym CASE is often used. This is leading to major changes in many aspects of the industry's operations, where issues need to be resolved in parallel and at speed, often in different geographical locations. Effective operation through virtual teams will become of increasing significance.

This paper is structured around five main sections. Following this Introduction, Section 2 outlines the research methodology and positions the two research objectives addressed in this article. Section 3 then reports the critical success factors (CSFs) drawn from current literature relevant to the research aim. Subsequently, Section 4 discusses the development of the initial V-CORPS framework, which was mainly based on concepts from the extant literature. Section 5 then outlines how the model was further developed, enhanced, and validated through a series of expert interviews carried out between October 2020 to April 2021. The final section provides an overall conclusion to the issues discussed in the paper and suggests how the model could be further developed and enhanced.

II. RESEARCH METHODOLOGY

Research design represents the structure that guides the appropriate research methods for the execution of data collection, and the subsequent analysis of the gathered data. In an initial stage, available literature in the automotive industry and in other industry sectors was investigated to ascertain current thinking on the leading and management of virtual

teams working on specific projects. Concepts and ideas from other disciplines were evaluated and adopted if deemed of value for leading virtual teams in the automotive industry. This was an integrative review [7] which aimed to synthesize areas of conceptual knowledge that could contribute to a better understanding of virtual team leadership and management, and lead to the development of an operational model.

An integrative literature review can provide an overview of the literature in a given field, encompassing the foremost ideas, models and debates, especially the concept that is not explicitly stated before – in this case the dynamics of virtual team leadership and management. It can provide the basis for a summary of the existing evidence concerning this theme and identify gaps in the current literature that may highlight possible areas for further investigation. It can also help build a framework or model for new research activities. This is particularly suitable when the research area is in the early stages of development, where key questions remain unanswered and an accurate picture of current thinking and evidence to date is required to promote the development of new models or methods.

The review of existing literature allowed the identification of critical success factors for the successful leading of virtual teams, and the construction of a provisional model for virtual team leading and management, which has now been progressed through primary research based on in-depth interviews with industry practitioners. A model of virtual project leadership in the automotive industry does not yet exist, and this research aimed to address this gap in the literature and in practice.

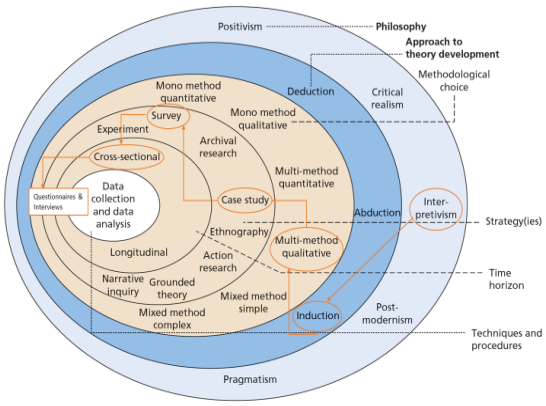


Figure 1. The research onion (based on Saunders et al. [8])

In terms of the methodological choices made for the primary research undertaken in the study, the "research onion" developed by Saunders et al. [8] provides a useful structure and guide to record elements of the methodology. The onion is separated into layers, each relating to an element of research method (Figure 1). The outer layer of the research onion refers to the philosophical position of the researcher, which in this case is interpretivist. This position allows the integration of humanistic qualitative methods and interests in a study [9], this being appropriate for research that considered the opinions and experiences of experts in the automotive field.

The next layer of the research onion indicates the approach used for theory development. Here, the inductive approach was chosen. Inductive reasoning is used when collective observations and experiences, including knowledge attained from other individuals and working practices, are combined to establish a "general truth" or acknowledged fact. The inductive approach was adopted in the analysis of empirical data, which was appropriate here because the model was built upon existing methods, experiences, and working practices relating to virtual team building and leading in the automotive industry.

TABLE I. ROLES AND RESPONSIBILITIES OF INTERVIEWEES

1.	**Head of Product Innovation:** 25 years of work experience as a director of product development, with different teams and 22 product patents.
2.	**Head of Project Management:** 26 years of work experience as project manager in the automotive industry.
3.	**President EMEA:** 15 years of work experience as a Plant and Project Manager
4.	**Vice President:** 21 years of work experience as a Product costing analyst
5.	**Project Manager:** 7 years of work experience as a Project Manager
6.	**Product Manager:** 6 years of work experience as a Product Manager in China
7.	**Head of Product Innovation:** 18 years of work experience as Project and Product Manager.
8.	**Vice President Business Development:** 20 years of work experience in sales and project development.
9.	**President and CEO:** 30 years of work experience in Product development and strategic Project Management.
10.	**Project Manager**: 15 years of work experience as a Project Manager in the automotive industry.
11.	**Senior Key Account Manager:** 15 years of work experience in the sales sector.
12.	**Product Certification Manager:** 19 years of work experience in programme management and product certification in the European and Japanese regions.
13.	**Agile Coach:** 22 years of work experience in Project Management.
14.	**Project Manager:** 10 years of work experience in the automotive area.
15.	**COO EMEA Region:** 20 years of work experience in automotive engineering, and 10 years as a Managing director.
16.	**Development Director:** 22 years of work experience in automotive engineering, and 16 years product development responsible
17.	**Industry Representative in EMEA and CIS region:** 17 years of work experience, and 13 years in project management
18.	**Product Manager:** 7 years of work experience as a Product Manager in CIS region

The aim was to create a model for virtual team building and leadership. A qualitative multi-methods investigation was pursued, which has its challenges, as different methodological traditions bring with them different communication traditions that are associated with different technical, rhetorical, and aesthetic criteria and norms. In an initial survey of views and perspectives, eighteen senior staff (Table 1) from a single automotive supplier were requested to complete a questionnaire, containing questions regarding the initial model and some open-ended questions. This was followed-up by semi-structured interviews to clarify the findings from the questionnaire. Yin [10] considers the interview as an important source for data collection, although the way in which an interview is conducted can be structured in several different formats. Here, the first step of this process involved the completion of a questionnaire by the interviewee, while

the second step used the questionnaire responses as the basis for discussion during the face-to-face interviews.

Using feedback from the initial round of interviews, the V-CORPS model was developed and enhanced. The model and the proposed activities and actions were then fed back to the interviewees, allowing the model to be refined and validated in a second round of interviews. In this sense, this was a research case study focusing on a single automotive supplier in Germany. The timeline was cross-sectional (commonly referred to as a "snapshot") and provided an insight into the current automotive working environment. The company is a part of a supply chain of a larger group, and is primarily responsible for the **E**urope, **M**iddle **E**ast, and **A**frica (EMEA) region. In 2018, the company had approximately 50,000 employees distributed over 98 production sites in 25 countries worldwide, with a turnover of €7.5 billion, as well as an annual investment in research and development of circa €1.2 billion. The company's language is English, and it works as a matrix organisation across a large geographical area and thus leadership over distance is of the utmost importance.

The company, in 2021, is engaged in more than twenty projects operating across the EMEA region, and with the increasing complexity of vehicles resulting in more complex and extensive projects, an expansion of different suppliers for vehicle projects overall is required for the OEMs. A centralised project work scheme is more time efficient for the OEMs, so that suppliers and sub-suppliers will be able to work according to the same project standards. This time efficiency is necessary due to the rigid time limits of development projects, which usually last for approximately 2 to 2.5 years, as both the complexity of the vehicles and component requirements have increased significantly.

In this context, the research objectives (ROs) addressed in the research and reported on here are:

RO1. To review existing literature on virtual leadership and virtual teams and identify critical success factors for the e-leadership of virtual teams in the automotive industry.

RO2. To develop a new operational model for the e-leadership of virtual teams that minimises personal contact and optimises project outcomes in the automotive industry.

III. CRITICAL SUCCESS FACTORS

Project management has become more versatile and complex in terms of people and project leading over the past few decades, especially when project teams are geographically dispersed. This has been done with the support of a variety of project management methods and concepts and the use of faster and cheaper communication technology, which have collectively facilitated the achievement of project goals and milestones more effectively. Whether these methods would also work for virtually managed teams in the automotive industry is a gap in the literature. A review of the extant literature suggests a number of factors as critical to the building and leadership of virtual teams. These may be seen as key concepts emerging from the integrative literature search on project management and team development, which the authors have considered of particular relevance to virtual team leadership and management. They are taken from the literature on both the automotive industry and other different industry sectors, and the relevant elements of project management methodologies.

These CSFs are as follows:

Build trust: A number of authors, including Maes and Weldy [11] and Ford et al. [12], have emphasised that trust between leaders and their team members, as well as amongst team members themselves, is the most important aspect for leading from a distance, and that it is possible to see trust as a key starting point for working with virtual teams. The building of trust is a pre-requisite for team cohesion, and the gaining of trust is part of social influence for distance-led team members, as discussed by Scheunemann and Bühlmann [13]. It is a major challenge in overcoming distance and time barriers and winning over team members. Building trust is an essential and challenging aspect for leading, and this is highlighted in the literature [11] [14]. Ford et al. [12] describe trust as the key to a capable virtual team.

Create a team structure: A team operating virtually, at a distance, needs to be underpinned and supported by a clear team structure. A team structure can engender intra-team communications and foster a collective, shared approach to the working behaviour of the team. This structure can be viewed as a contract for team members that allows them to pursue individual and project objectives effectively. Klitmøller and Lauring [15] found that communication and knowledge sharing were more challenging in a virtual team environment than with face-to-face counterparts, and that a clear team structure was essential in overcoming these challenges.

Overcome cultural and language barriers: The avoidance of the possible negative impact of cultural differences is a necessary preventive measure to mitigate possible bias between the different team members. Nader et al. [16] note that cultural barriers are a serious impediment to the effectiveness of virtual teams. It is essential that the general understanding and respect of culture is recognised by the leader, and that neither origin nor gender plays a role in the team, with only ability and merit counting.

Language barriers are an important issue which cannot be underestimated. Due to the fact that the members of virtual teams often do not speak the same language, many companies opt for mutual understanding through English [13]. It is essential that the leader considers this issue and accommodates language differences during complex negotiations. Team members may need to develop agreed procedures for avoiding misunderstandings and time wasting through misinterpreted instructions or information.

Manage time and distance barriers: One of the most important pre-requisites for successful virtual working is the effective management of time and distance barriers. The "follow the sun methodology" allows the phased deployment of teams around the globe, and the increased use of collaboration and communication tools can facilitate more autonomous work, and yet also allow all team members to be

in one virtual space during critical situations. Effective communication across time and distance barriers is essential to give team members a form of security (the feeling that they are not alone) and can be seen as the "project life-blood" of the team. Layng [4] found that communication was a key factor in the success of virtual teams.

There is a range of available technologies to support communication and co-working in virtual teams, which have seen increased deployment in the lockdown periods brought in as a response to the coronavirus pandemic. In addition to standard phone, texting and email, there are more sophisticated messaging services like Microsoft Teams, WhatsApp and Facebook Messenger. Video conferencing and meeting tools such as Skype and Zoom support virtual meetings across time and distance boundaries, and many of the standard project and document management tools will be used by virtual teams. Similarly, if virtual teams are interacting with the customer, shared access to customer files (probably via a customer relationship management system) will be necessary. The use of the Cloud to provide shared access to these software systems is an option.

Influence through horizontal communication: Virtual teams are frequently multi-functional, composed of individuals and specialists drawn from different departments, with virtual leaders who often have no direct line management authority. Influencing skills are thus of particular importance, especially in virtual teams when there are limited opportunities for face-to-face meetings. The influencing of team members can take place through adopting elements of nonviolent communication (Observations, Feelings, Needs/Values, and Requests) to minimise escalation of disagreements and minor disputes among team members. Alistoun and Upfold [17] discussed how virtual team leaders can be trained to successfully influence team members, deploying computer-mediated communication, building trust, shortening subjective distance, sharing information, processing gains and losses, dealing with feelings of isolation, encouraging participation, and enhancing coordination and cohesion. If the leader can appear to communicate on the same hierarchical level as team members (horizontal communication), the leader is seen to be on the "same wavelength" as the team members, only revealing their true hierarchical position in urgent or emergency situations. Influencing team members is a topic which has an impact on team and work behaviour, and must be considered before and during the project, and constantly being improved upon by getting to know the team members.

To have social influence on team members, virtual team leaders need to use a range different communication technology to ensure a social presence [18]. The use of communication technology makes the virtual socialisation of team members possible, allowing leaders to assess their teams' capabilities, and receive, provide and accept feedback from their team members. For team members, it promotes a sense of connectedness to leaders, as well as allowing leaders to create a social presence [19].

These CSFs suggest the key issues for establishing a successful virtual team, but also indicate which factors are necessary for successful virtual leading. The tendency to work virtually is growing [10], and recent research reports an improvement in the effectiveness of virtual teams from less than 30% in 2006 [20] to 68% in 2016 [21].

IV. Building the V-CORPS Model

The automotive industry operates globally and working with virtual teams has become an inevitability. Building a team that has to work virtually requires the main focus to be on people. The integrative literature review suggests that virtual team development and leadership can usefully be based on the team development stages defined by Tuckman [22] and Tuckman and Jensen [23] for small co-located teams.

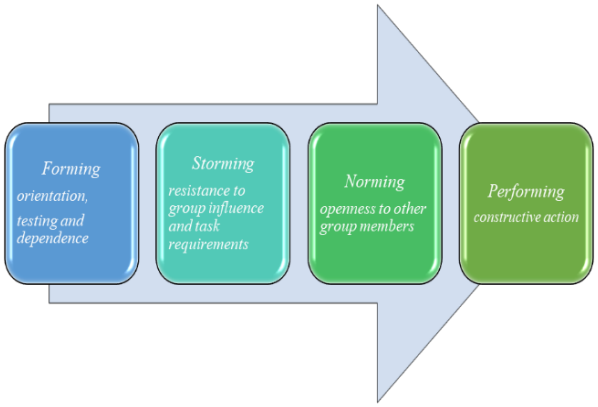


Figure 2. Stages of team development (after Tuckman [22])

The four stages depicted in Figure 2 can be seen as the group developmental process for interpersonal relationships between team members of co-located teams. In the first part of the model's development, interpersonal relationships and task activities are considered, resulting in a four-stage model in which each stage needs to be successfully navigated in order to reach effective group functioning [22].

Table II. Tuckman and Jensen's Group Structure and Task Activities [25]

Stages	Group structure	Task activity
Forming	Testing and dependence	Orientation of the task
Storming	Intragroup conflict	Emotional response to task demands
Norming	In-group feeling and cohesiveness development; new standards evolve, and new roles are adopted	Open exchange of relevant interpretations; personal opinions are expressed
Performing	Roles become flexible and functional; structural issues have been resolved; structure can support task performance	Interpersonal structure becomes the tool of task activities; group energy is channelled into the task; solutions can emerge

Adjourning	Anxiety about separation and termination; sadness; feelings; towards the leader and group members	Self-evaluation

Tuckman and Jenson [23] added one further stage – Adjourning - as a separate and distinct final stage in which separation of team members would be considered. The stages of development are not seen as a process, but more as a life cycle (Figure 3) for spin-off and reintegration of team members. Tuckman and Jensen [23] found that in groups where substantial amounts of activity take place, interpersonal relationships are developed, and group dissolution becomes an extremely important issue for many of the members. The authors developed the model to indicate, for each stage, a description of their associated group structures and task activities (Table 2). The group structures were seen as "the pattern of interpersonal relationships - the way members act and relate to one another", whilst task activities were "the content of interaction as related to the task at hand" [25].

Tuckman's model has some limitations. It was developed with therapeutic groups in mind, and its interpretation and application in other working environments is challenging. Cassidy [26] notes that the Storming stage in particular may not be clearly defined for practitioners outside of therapeutic groups. It is thus difficult to apply directly to daily working lives and needs to be customised for individual team development situations. The model also does not consider how team personnel may change over time and the steps that must be taken to introduce and integrate new team members, which is particularly challenging when a project is at an advanced stage [27]. The objective of the research, therefore, was to attempt to adapt this model to the automotive industry, and at the same time to interweave the CSFs discussed above into a new adapted framework, customised for this industry sector.

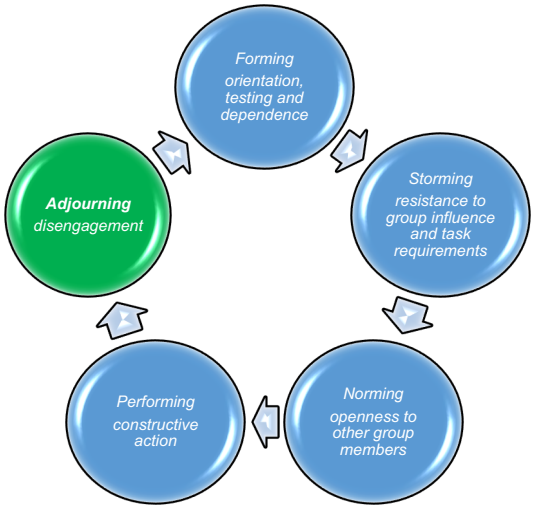


Figure 3. Tuckman and Jensen's Group Development Lifecycle [23]

The initial framework (Table 3) looked to build upon the CSFs identified in the literature and incorporate some of the thinking evident in Tuckman's model. In addition, elements of project management methodologies were incorporated into the five-stage model, which also takes into consideration a number of management challenges for virtual teams – such as differences in employment and occupational health legislation across different countries, norms regarding social interaction, a lack of mutual knowledge of context and access to dispersed knowledge, stress and fatigue issues, and data security [28] [29].

It is important to note the differences between co-located and virtual teams, and how they communicate to reach their goals. As pointed out by Berry [30], a co-located team is a group of individuals who interact interdependently and who are brought together or come together voluntarily to achieve certain outcomes or accomplish particular tasks and are able to have face-to-face conversations or meetings at any time. Virtual teams could theoretically comprise the same individuals as co-located teams, with the premise of working over the world and communicating through the use of information and communications technology. Virtual team members consist of individuals spread across geographies, cultures and time zones.

Managing virtual teams is different to, and more complex than, managing face-to-face teams. Virtual teams are groups of individuals that still share most of the characteristics and dynamics found in traditional teams. The challenge for virtual teams is in cultural differences, mentalities, work-settings etc., which are of significance for the virtual leader when influencing team members from a distance. Cortellazzo et al. [31] state that when focusing on behavioural norms, it is particularly important for virtual teams to have a clear definition of the norms pertaining to their use of communication tools, through which information flows and activities are performed. Berry [30] suggests that the effective management of virtual teams requires knowledge and understanding of the fundamental principles of team dynamics, regardless of the time, space, and communication differences between virtual and face-to-face working environments.

These considerations and the CSFs discussed above underpinned the development of the initial 5-stage model for virtual leadership and management of virtual teams. Using indicators evident in the existing literature, some initial key activities were assigned to each cell in the 5x5 matrix (Table 3). The stages in the model are outlined below.

Creating the team: To support virtual team members in achieving a high level of performance, some key considerations need to be taken into consideration in the creation of the team. The choice of the appropriate team members is vital – not only those that have the relevant work experience for project requirements, but also those that are able to work remotely, being self-motivated and independent. The project manager has to make a pre-analysis of the team members and speak to their line mangers to get an impression

TABLE III. THE INITIAL V-CORPS MODEL

CSF/ V-CORPS Stage	Creation	Organisation	Relationship Building	Performance Evaluation	Sign-off & Closure
Build trust	First impressions – preferably via a face-to-face meeting – are important in building trust	Clearly define project tasks and responsibilities and assign roles for individual team members	Conduct the "Big Five" analysis of each team member Offer support in critical situations	Performance evaluation underlines mutual dependence of team members in achieving successful project outcomes	Acknowledgement of lessons learned and reflection on team leading can reinforce mutual trust and respect
Create team structure	Explain and apply corporate policies for team working Clarify expected outcomes	Define and agree terms and conditions, project rules and team composition	Introduce 'team working contract' and a team chat/forum to facilitate team communication	Highlight the importance of the team structure in achieving project success	Team dissolution. Creation of long-lasting relationships
Overcome cultural and language barriers	Establish whether any cultural or language barriers exist	Clarify support actions and steps to be taken in the event of language or cultural issues. Provide a common understanding of working posture and customer requirements	Equal treatment and support during breakdown of communication. Explain how and when to escalate properly to avoid time wasting	Stress the importance of a standard work-culture across the team. Ensure that team performance comes before individuality	Private contact data exchange (if desirable). Stay in touch with team members after project closure
Manage time and distance barriers	Investigate and evaluate implications of geographical differences and discuss how to overcome them	Define ways of working to accommodate time and distance issues. Establish technology platforms to be used for virtual team operations	Show dependencies between tasks and team members. Implement simulation procedures to avoid unnecessary product testing.	Review impacts of time and distance differences across the team Adjust working practices accordingly Provide appropriate training	Avoid anxiety about separation and project closure
Influence through horizontal communication	Round of interviews Project manager treats team members as equals	Highlight the importance of teamwork and the value of the project to the company	Intervene only when necessary, e.g., key decisions, supportive role, problem escalation	Create a relaxed environment while focusing the team on specific project milestones. Avoid coercion	Project evaluation. Encourage mutual support. Team members leave the project feeling appreciated

of their ability to work in a virtual environment. This pre-analysis is essential prior to taking the next steps of team member selection since virtual teams tend to be more sensitive to trust issues and the need for communication [32].

Caulat [20] concludes that people who are very process-oriented and structure-driven might be effective when managing the virtual process of communication between the members during a project but might find it challenging to facilitate and participate in virtual meetings where spontaneity is required.

Cross-cultural awareness is also necessary for team cohesion, influence, and trust promotion. It is essential that the project manager be in place as the first team-building measure, with an overview of team member actions and reactions, especially during the team creation period. The project manager can assess how team members score against the project CSFs.

Building trust, as Seshadri and Elangovan [33] note, is an interpersonal challenge faced by managers to foster collaboration with team members through communication and building relationships. Caulat [20] argued that, by working with cultures as diverse as Japanese, Indian, Swedish, and Russian, she realised that cross-cultural awareness may help in understanding each other, but that it is certainly not sufficient for establishing a sound basis for the development of trust within the team. Although the pre-investigation of team members is essential, it is the first meeting where the project manager meets his team face-to-face, and can leave a positive, lasting impression, which can establish the tone and *modus operandi* for future project procedures [34].

Organising the team: Maintaining a uniform team structure before and during the project is an essential factor in avoiding time-consuming discussions regarding the *modus operandi* of the team.

The organisation of virtual team structures needs special consideration, not only for the establishment of working procedures, but also regarding social aspects, and the avoidance of miscommunication or misunderstandings which can affect the entire team's behaviour. It is essential to sensitise each team member to the potential impact of social behaviour. This structure is significant in facilitating communication and knowledge sharing, which is more challenging than with face-to-face counterparts [17].

A clear organizational structure is also of particular importance when dealing with a complicated project environment that includes challenges in language, political climates, organisational policies, time zones, and cultures [35]. To counteract these challenges, it is essential to outline the CSFs for the project through the organisation stage and discuss each of them with the team members, to define rules for working with each other. The project manager may need to act as a moderator between the team members and intervene in critical situations (e.g., escalations between team members).

It is also essential to consider the language skills of the team members before and during the project process because virtual workers with low language proficiency invoke apprehension and uncertainty in individuals [36]. The organisational structure can be used as the framework, within which issues can be tackled and team cohesion enhanced, and through which the project manager can discuss and explain what he/she expects from team members.

Relationship building: The team organization structure provides the starting point for relationship building between the project manager and the team members. Building relationships is the foundation of all teamwork, especially for virtual teams, and can help counteract the multiple negative aspects of working over distance [4]. It is necessary to confront prejudices about the working performances of the different nationalities of team members.

It is advisable that communication between the individual team members takes place at least two weeks before the start of the project [4], as this will, in the best case, enable the group to become more socially grounded through a personal meeting or by participating in "virtual water cooler communication", thereby increasing their loyalty to the group [37] [38]. This will support relationship building and similarities between the team members can be found before the project starts. It is important for virtual leading teams to create a social environment to promote team cohesion, which will be established through interpersonal challenges for the project manager and ensure that team members communicate with each other, build relationships and foster trust [32]. This builds commonalities, which creates sympathy, trust and encourage team spirit.

In the relationship building phase, a number of techniques can be used, such as Goldberg's Big Five model [39] for assessing and understanding personality traits. Project managers can try to analyse themselves and the team members to find out what kind of leadership is right for each member, and how to employ the right team member in the right position. This model is also useful for relationship building between team members, for working from a distance and improving mutual influencing effectiveness. The leader must not neglect the social behaviour of the team members, and one possible tactic here is to book a short slot at the beginning of each team meeting to speak about non-project themes. This gives an added value of trust, which can greatly improve team effectiveness and relationship building.

Performance evaluation: Leading a team during a project is an evolving and ongoing process. It is essential to update the team regularly and be responsible for enabling communication. The more team members are up to date, the better their performance is, and the fewer miscommunications and misunderstandings there are. It is advisable to try to bring more personality and dependency to the virtual world. It is also important to make clear to team members that their performance levels depend on each other, and to get them to consider what kind of impact their performance has on project outcomes and the company. The quality and effectiveness of information exchange also impacts on team performance – used correctly, it can empower individuals, alter behaviour, and help develop a cohesive team.

The same is true for decision-taking, where team performance counts. Care taken by the project manager (for example in including all team members in certain decisions) can enhance the overall performance of the whole team. In virtual teams, language and mental barriers must be considered. Shared understanding of key decision options is important. Horizontal communication is essential, where team members get the feeling that they are on the same working level and can contribute to a discussion and decision.

Sign-off and closure: The bonding between team members during the project phases can create a form of psychological contract, which will reflect the social team influence of the project manager, and that of the team members themselves. The dissolution of this contract is a key element of the project sign-off and closure stage, and it is an important aspect for the possible future creation of new virtual teams. King [40] defines a psychological contract as an individual's belief in the perception of reciprocal obligations between that person and another party. For working in a virtual team, this can be considered as a contract between team members, which is unofficial, but essential for the project.

The disbanding of the psychological contract will likely involve a meeting between the project manager and the entire team on site when project completion meetings can be held with each team member. Project disbandment can be done in a virtual way, but psychological effectiveness, in terms of the appreciation of individual team members, is not as valuable as when there is a local presence face-to-face. In the final discussion, both positive and negative aspects of the project can be reviewed, and the further growth of the team in subsequent projects can be discussed. The project manager should also have their team ready at the end of the project to give some reflection and feedback on the project management process, so that negative aspects can be aired and reviewed.

V. Model Development and Validation

The V-CORPS model was developed and refined in a series of stages. A questionnaire was designed to reflect the initial V-CORPS model and emailed to eighteen experts, all with relevant experience in project management (Table 1). The questionnaire contained eight questions relating to the initial model and respondents were asked to give their views on its contents. Some used a Likert scale [8] allowing respondents to register their level of agreement or

disagreement with certain statements relating to the model. The questionnaire concluded with questions aimed at identifying whether or not some CSFs were necessary for the final build of the model, and to further understand the value of these CSFs in the experts' opinion. This feedback provided an indicator of support or otherwise for the general direction of travel embodied in the model, and of the possible future development of the activities at each stage (Table 4). It gave a clear picture of the views of the experts regarding the challenges of virtual project management, and the obstacles of virtual team building and leadership in the automotive industry. In addition, it also became clear that secondary aspects such as capacity bottlenecks and time-critical project milestones are particularly important in a virtual environment and can undermine project success. The answers in the questionnaire also highlighted how strongly the experts felt a duty towards their team members and their commitment to playing the main role in a functioning project.

TABLE IV. INITIAL QUESTIONNAIRE FINDINGS: ORGANISATION STAGE

Question	Outcome	Finding	Research action
Organisation Stage			
1.	94.4% positive feedback 5.6% negative feedback	A clearly-defined role in a team for each team member promotes team cohesion and trust improvement	Definition of roles in a team is a part of trust building which must be implemented in the model
2.	100% positive feedback	Adherence to project rules and procedures is important	To be implemented in the model and considered as a critical factor
3.	100% positive feedback	Team has to know how to escalate when necessary	The model must indicate that the leader needs to clarify the escalation process
4.	94.4% positive feedback 5.6% negative feedback	The definition of working guidelines is important in virtual teams	This must be evidenced in the model
5.	88.3% positive feedback 5.9% no opinion 5.9% negative feedback	Emphasising the team is one unit is important for the team in the early stages	Clarify regarding this CSF in the model

Overall, the experts showed a keen interest on the V-CORPS model, and this provided a basis for subsequent discussion in one-to-one interviews to flesh out further suggestions and recommendations relating to each cell in the matrix – trying to garner as many ideas as possible for activities that would typically be required across a virtual project. This produced a significant number of new ideas and activities for each cell in the matrix, which were recorded in the model, cell by cell. In a second round of interviews, this material was then presented back to interviewees for discussion and prioritisation, thereby providing the basis for a filtering process (Figure 4), which identified the activities that were generally supported by the experts for each stage in the model. The final operational V-CORPS model (Table 5) was again returned to interviewees for final comment and ratification.

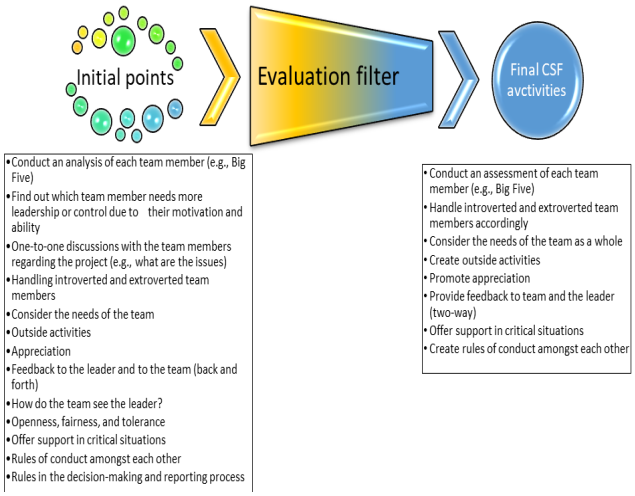


Figure 4. Example of activity reduction in the relationship Building Trust stage for the Building Trust CSF

This represented the final step in a research process that comprised six interlinked steps, each of which was undertaken sequentially (Figure 5). Research was based on a survey conducted in a single company, but due to the selection of experts (from line management to the CEO) a wide range of data was obtained. The interviews were analysed through data reduction and coding. These were summarised in the form of statements and implemented in the model's development. The conceptual framework was validated through a survey and semi-structured interviews with eighteen experts. These data were used to create a preliminary operational V-CORPS model, which was then tested and validated via a follow-up survey with six experts.

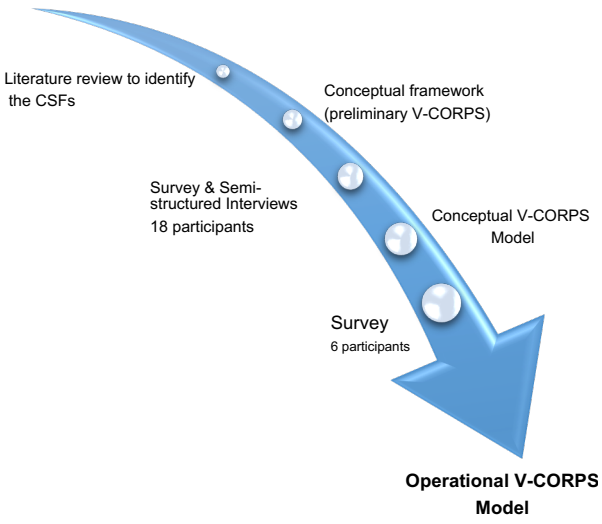


Figure 5. The V-CORPS model development and evaluation research process

VI. CONCLUSION

The research project reported upon here has now successfully addressed the two main research objectives noted above. Analysis of the existing literature was used as the basis for the development of five CSFs for virtual teams. Then, by taking some of the principles and concepts from Tuckman's staged model for group facilitation [22], and combining them with the CSFs, a new model for virtual team operation in the digital era was developed, refined, and validated. The CSFs have relevance to each of the five stages on the V-CORPS model, which can be used as a guideline and point of departure for those assembling and leading virtual teams. This will require new ways of working and a change in management culture and can best be viewed as what Chan Kim and Mauborgne [41] call "non-disruptive creation." They suggest that, compared with disruptive change in business models, "non-disruptive creation opens a less threatening path to innovation for established companies", and that "it doesn't directly challenge the existing order or the people who make their livelihoods based on it" (p.9).

All stages on the V-CORPS model are important, but as Tuckman [25] concluded in his studies of group development, the outcomes from the performance evaluation stage will be critical to final project results. This means, from the leader's perspective, it is necessary to bring the team to the most effective performance level to fulfil the project requirements. Team creation, team organisation and relationship building are all of significance in supporting and progressing this objective. It is also important that virtual teams are equipped with the process capability and technology support to respond to changes quickly and effectively [42].

This research clearly has its limitations. It is based primarily on a set of interviews from one company in the automotive supply chain, and project realities in other industry sectors will vary in some regards. This suggests testing and development of the model in other corporate environments, in which virtual projects play a significant role, would be worthwhile. This could include not only other automotive companies, but also other industries involved in product development using globally dispersed resources. As globalisation and the widening application of digital technologies changes working practices, frameworks like the V-CORPS model that provide guidance on the process and people aspects of change management will be of increasing relevance and value.

TABLE V. THE VALIDATED V-CORPS MODEL

CSF/ V-CORPS Stage	Creation	Organisation	Relationship Building	Performance Evaluation	Sign-off & Closure
Build trust	• Get first impressions (Face-to-face or video meeting) • Be prepared to answer questions from the team • Facilitate introductions to each other (What am I good at?) • Get to know each other through games (Common first task) • Offer help as requested	• Get a first impression from the respective team members and show appreciation • Show trust in the team and the project • Find out each team member's expectations • Create an outage plan • Promote mutual assistance in the team • Connect through social media • Create a WhatsApp group	• Conduct an assessment of each team member • Handle introverted and extroverted team members accordingly • Consider the needs of the team as a whole • Create outside activities • Promote appreciation • Provide feedback to team and the leader (two-way) • Offer support in critical situations • Create rules of conduct amongst each other	• Outline the importance of reliability between team members and the dependency on performance • Review the concerns of team members identified in the relationship phase (appreciation) • Create commitment • Show the consequences of subtasks for the whole team • Create compatibility • Use Agile project planning tools	• Acknowledge lessons learnt, reporting results into the company • Recognize the team's achievements and celebrate them • Get constructive feedback in both directions (360 ° feedback, no negative emotions) • Leave a positive impression • Admit mistakes openly • Use virtual meetings & Big data analyses
Create team structure	• Address corporate policies • Develop vision and mission • Explain project scope • Address project guidelines and values • Highlight the common goal and the expectations of the team and the leader • Create a team charter, explaining why people are there and what to expect	• Assign roles for individual members • Discuss working methods • Determine meeting culture • Address consequences in case of non-compliance • Provide understanding of exceptions for special cases • Highlight importance of completing tasks on time and communicating when not • Set technology standards, procedures, and guidelines. • Organise standard IT training programme	• Agree working practices • Establish communication channels and technical team communication • Create on-site visits (if possible) • Create regular meetings (15 min stand up meeting if possible) • Promote mutual support • Create virtual face-to-face meetings and a chat/forum • Call instead of email! • Create rules for the decision-making and reporting processes • Present the effectiveness through cloud computing	• Highlight the importance and the effectiveness of the project structure • Create project goals and regular effectiveness reviews • Create measurement and key indicators for the team • Consider retrospective issues (are there points that have crept in that stand in the way) • React to changes quickly! • Reduce reactive work	• Ensure smooth team dissolution • Create a long-lasting relationship • Write letters of recommendation • Show consideration for each individual team function • Build / enlarge network
Overcome cultural and language barriers	• Identify cultural or language barriers and address them • Ensure common language • Organize cultural training – and use as an ice breaker • Build confidence from the first stage	• Observe company guidelines and policies • Define support actions and the steps to be taken should an issue arise • Identify and support language weaknesses • Take away the fear of communicating with the leader • Explain work organization correctly • Emphasise that private views have no place in professional life • Create cross-cultural seminars • Active mobile use	• Develop corporate identity • Intervene in an emergency • Create consequences for certain actions • Appoint a mediator who understands the culture and will support you as a coach • Exemplify professional and equal treatment in the team	• Stress the importance of work-culture • Ensure that performance comes before individuality • Avoid pack formation in the team • Promote team spirit across locations • Characterize discipline • Respect Individuality	• Exchange (if desirable) private contact data • Stay in touch with team members after project end • Create an individual farewell according to cultural customs and norms • Maintain on-going network • Stay connected through social media
Manage time and distance barriers	• Introduce and regulate technology deployment • Investigate means of communication (are we up to date?) • Define communication channels (usual times) • Address hardware and software resources for the team • Develop regular exchange of information (weekly feedback) • Define team planning (how do I use the team members to save my time)	• Define the use of working tools for the project • Define the preparatory work to avoid time loss due to time differences (e.g., Follow the Sun methodology) • Define processes, methods, documents, with templates as appropriate • Make special tools available for certain functions • Define working time windows • Analyze distribution of working hours • Define time for communication • Organize training if necessary • Offer help (mutually)	• Show dependencies between tasks and team members • Ensure that team members can rely on one another • Develop regular results review • Increase promotion of understanding • Offer support with prioritization • Underline that the behaviour of an individual influences team results • Create flowcharts and solution paths • Present Big Data analyses	• Demonstrate ways of working that avoid wasting time • Specify tools for teamwork in detail • Emphasise that methods and procedures must be known and followed • Work on training courses and problems in a team • Work in a way that promotes quality • Present the effectiveness of cloud computing	• Avoid anxiety about separation and ending • Promote contact retention • Offer perspective wherever possible • Give self-confidence • Give outlook for possible future projects and cooperation
Influence through horizontal communication	• Create round of interviews • Champion working in a collaborative way • Develop flat hierarchy • Distribute roles and responsibilities • Show openness for social media • Present Agile Methods	• Highlight the importance of teamwork (One team = One unit) • Develop equality (equal treatment for all team members) • Create understanding and acceptance by the team that the leader has overall responsibility • Develop 'acting as a team' (get through success and failure together) • Give positive feedback where possible • Address errors openly	• Intervene only when crucial e.g., key decision making, supportive roles, and escalating situations • Maintain self-control • Keep calm (no heated actions) • Create a pastoral care function • Be consistent in using the virtual meetings technology	• Create an environment where the team can focus on project milestones • Avoid increased workload • Always motivate to communicate • Establish an environment based on trust • Function as a role model	• Ensure team members leave the project feeling appreciated • Encourage mutual respect after project completion

REFERENCES

[1] A. Quade, M. Wynn, and D. Dawson, "Collaboration through virtual teams: towards an operational model for virtual project leadership in the automotive industry", the Tenth International Conference on Advanced Collaborative Networks, Systems and Applications, COLLA 2020, Porto, Portugal.

[2] K. Jugdev, J. Thomas, and C. Delisle, "Rethinking Project Management: Old Truths and New Insights", International Project Management Journal, vol. 7, no. 1, pp. 36-43, 2001.

[3] P. Svejvig and P. Andersen, "Rethinking project management: A structured literature review with a critical look at the brave new world", International Journal of Project Management, vol. 33, no. 2, pp. 278-290, 2015. doi:10.1016/j.ijproman.2014.06.004.

[4] J. Layng, "The Virtual Communication Aspect: A Critical Review of Virtual Studies Over the Last 15 Years", Journal of Literacy and Technology, vol. 17, no. 3, pp. 172-218, 2016.

[5] National Instruments, Automotive Research User Handbook: prototypes and testbeds to validate the transportation solutions of tomorrow, National Instruments, 2018.

[6] Deloitte, CASE automotive trends and insights, 2021. Available at: https://www2.deloitte.com/us/en/pages/manufacturing/topics/case-automotive-industry.html [Retrieved August, 2021]

[7] R.J. Torraco, "Writing integrative literature reviews: Using the Past and Present to Explore the Future". Human Resource Development Review, vol. 15, no. 4, pp. 404–428. doi: 10.1177/1534484316671606

[8] M. Saunders, P. Lewis, and A. Thornhill, *Research Methods for Business Students*, 8th ed., Pearson Education Limited, 2018

[9] D. Layder, Understanding Social Theory, 2nd ed., 2006. doi:10.4135/9781446279052

[10] R. K. Yin, Applications of Case Study Research, 3rd ed., SAGE Publications Inc., 2012

[11] J. Maes and T. Weldy, "Building Effective Virtual Teams: Expanding OD Research and Practice", Organization Development Journal, vol. 36, no. 3, pp. 83-90, 2018.

[12] R. Ford, R. Piccolo, and L. Ford, "Strategies for building effective virtual teams: Trust is key", Business Horizons, vol. 60, no. 1, pp. 25-34, 2017.

[13] Y. Scheunemann and B. Bühlmann, "It's Time for a Virtual Team Driver's License". Evernote + Google, 1 – 23, 2018. Available at: https://evernote.com/c/assets/marketing/virtual-team-management/virtual-team-whitepaper-r3.pdf. [Retrieved May, 2020]

[14] V. Seshadri and N. Elangovan, "Role of Manager in Geographically Distributed Team", Journal of Management, vol. 6, no. 1, pp. 122-129, 2019. Available at: http://www.iaeme.com/JOM/issues.asp?JType=JOM&VType=6&IType=1 [Retrieved May, 2020].

[15] A. Klitmøller and J. Lauring, "When global virtual teams share knowledge: Media richness, cultural difference and language commonality", Journal of World Business, vol. 48, no. 3, pp. 398-406, 2013. doi:10.1016/j.jwb.2012.07.023.

[16] A. Nader, A. Shamsuddin, and T. Zahari, "Virtual Teams: a Literature Review", Australian Journal of Basic and Applied Sciences, vol. 3, no. 3, pp. 2653-2669, 2009.

[17] G. Alistoun and C. Upfold, "A Proposed Framework for Guiding the Effective Implementation of an Informal Communication System for Virtual Teams", Proceedings of the European Conference on Information Management & Evaluation, 2012.

[18] A. Walvoord, E. Redden, L. Elliott, and M. Coovert, "Empowering followers in virtual teams: Guiding principles from theory and practice", Computers in Human Behavior, vol. 24, no. 5, pp. 1884-1906, 2008. doi:10.1016/j.chb.2008.02.006.

[19] L. Cowan, "e-Leadership: Leading in a Virtual Environment – Guiding Principles For Nurse Leaders", Nursing Economics, vol. 32, no. 6, pp. 312-322, 2014.

[20] G. Caulat, Virtual leadership. 360 The Ashridge Journal, pp. 6-11, 2006.

[21] C. Solomon, Trends in Global Virtual Teams, 2016. Available at: http://cdn.culturewizard.com/PDF/Trends_in_VT_Report_4-17-2016.pdf [Retrieved May, 2020].

[22] B. W. Tuckman, "Developmental Sequence in Small Groups", Psychological Bulletin, vol. 63, no. 6, pp. 384-399, 1965.

[23] B. W. Tuckman and M. Jensen, "Stages of Small-Group Development Revisited", Group & Organization Studies, vol. 2, no. 4, pp. 419-427. 1977.

[24] D. A. Bonebright, D. A. (2010), "40 years of storming: a historical review of Tuckman's model of small group development", Human Resource Development International, vol. 13, no. 1, pp. 111-120, 2010. doi:10.1080/13678861003589099

[25] B. W. Tuckman, "Developmental sequence in small groups", Group Facilitation: A Research and Applications Journal, vol. 3, pp. 66-81, 2001.

[26] K. Cassidy, "Tuckman Revisited: Proposing a New Model of Group Development for Practitioners", Journal of Experiential Education,vol. 29, no. 3, pp. 413-417. 2007

[27] T. Rickards and S. Moger, "Creative Leadership Processes in Project Team Development: An Alternative to Tuckman's Stage Model", British Journal of Management, vol. 11, no. 4, pp. 273-283, 2000. doi:10.1111/1467-8551.00173

[28] P. Pyöriä, "Managing telework: risks, fears and rules", Management Research Review, vol. 34, no. 4, pp. 386-399, 2011. doi:10.1108/01409171111117843.

[29] J. MacDuffie, "HRM and Distributed Work", The Academy of Management Annals, vol. 1, no. 1, pp. 549-615, 2007. doi:10.1080/078559817.

[30] G. Berry, "Enhancing Effectiveness on Virtual Teams: Understanding Why Traditional Team Skills Are Insufficient", Journal of Business Communication, vol. 48, no. 2, pp. 186-206, 2011. doi:10.1177/0021943610397270.

[31] L. Cortellazzo, E. Bruni, and R. Zampieri, "The Role of Leadership in a Digitalized World: A Review", Front Psychol, vol. 10, pp. 1-21, 2019 doi:10.3389/fpsyg.2019.01938.

[32] B. Rosen, S. Furst, and R. Blackburn, " Overcoming Barriers to Knowledge Sharing in Virtual Teams", Organizational Dynamics, vol. 36, no. 3, pp. 259-273, 2007. doi:10.1016/j.orgdyn.2007.04.007.

[33] V. Seshadri and N. Elangovan, "Role of Manager in Geographically Distributed Team", Journal of Management, vol. 6, no. 1, pp. 122-129, 2019. Available at: http://www.iaeme.com/JOM/issues.asp?JType=JOM&VType=6&IType=1 [Retrieved May, 2020].

[34] I. Zigurs, "Leadership in Virtual Teams: Oxymoron or Opportunity?", Organizational Dynamics, vol. 31, no. 4, pp. 339 – 351, 2003.

[35] D. Barnwell, S. Nedrick, E. Rudolph, M. Sesay and W. Wellen, "Leadership of International and Virtual Project Teams", International Journal of Global Business, vol. 7, no.2, pp. 1-8, 2014.

[36] T. Neeley, P. Hinds, and C. Cramton, "The (Un)Hidden Turmoil of Language in Global Collaboration",

Organizational Dynamics, vol. 41, no. 3, pp. 236-244, 2012. doi:10.1016/j.orgdyn.2012.03.008.

[37] A. Akkirman and D. Harris, "Organizational communication satisfaction in the virtual workplace", The Journal of Management Development, vol. 24, nos 5/6, pp. 397-409, 2005.

[38] H. Duckworth, "How TRW Automotive helps global virtual teams perform at the top of their game", Global Business and Organizational Excellence, vol. 28, no. 1, pp. 6-16, 2008. doi:10.1002/joe.20237.

[39] L. Goldberg, "An Alternative 'Description of Personality': The Big-Five Factor Structure", Journal of Personality and Social Psychology, vol. 56, no. 6, pp. 1216 – 1229, 1990.

[40] J. King, "White-collar reactions to job insecurity and the role of the psychological contract: Implications for human resource management", Human Resource Management, vol. 39, no. 1, pp. 79-92, 2000.

[41] W. Chan Kim and R. Mauborgne, "Nondisruptive Creation: Rethinking Innovation and Growth", Sloan Management Review, Spring, 2019. Available at: https://sloanreview.mit.edu/article/nondisruptive-creation-rethinking-innovation-and-growth/ [Retrieved April 25, 2021].

[42] G. Schmidt, "Virtual Leadership: An Important Leadership Context", Industrial and Organizational Psychology, vol. 7, no. 2, pp. 182-187, 2015. doi:10.1111/iops.12129.

Becoming a Smart City: A Textual Analysis of the US Smart City Finalists

Jasmine DeHart
School of Computer Science
University of Oklahoma
Norman, Oklahoma, USA
dehart.jasmine@ou.edu

Oluwasijibomi Ajisegiri
School of Computer Science
University of Oklahoma
Norman, Oklahoma, USA
oluwasijibomi.ajisegiri@ou.edu

Greg Erhardt
Department of of Civil Engineering
University of Kentucky
Lexington, KY, USA
greg.erhardt@uky.edu

Jamie Cleveland
Duke Energy One
Charlotte, NC, USA
jamie.cleveland@duke-energy.com

Corey E. Baker
Department of Computer Science
University of Kentucky
Lexington, KY, USA
baker@cs.uky.edu

Christan Grant
School of Computer Science
University of Oklahoma
Norman, Oklahoma, USA
cgrant@ou.edu

***Abstract*—The term "smart city" is widely used, but there is no consensus on the definition. Many citizens and stakeholders are unsure about what a smart city means in their community and how it affects cost and privacy. This paper describes how city planners and companies envision a smart city using data from the 2015 Smart City Challenge. We use text analysis techniques to investigate the technology and themes necessary for creating a smart city using surveys, document similarity, cluster analysis, and topic modeling from the seven finalists from the 2015 Smart City Challenge Applicants. With this investigation, we find that smart city requests include various technologies, and the goal of smart cities is to enhance and connect the communities to improve the lives of its' citizens. On average, aspiring smart cities requested 12 new or improved technologies. We also find that two of the seven studied smart city applications center privacy in their proposals. The analysis within gives governments and citizens a common interpretation of a smart city.**

***Index Terms*—smart city; privacy; networks; text analysis.**

I. Introduction

The concept of a "smart city" has recently led people, cities, and governments to pursue idyllic improvements to municipal infrastructure. Each stakeholder may have different expectations for how their city should invest in improvements. Currently, no standard definition for a smart city exists causing variable expectations of residents, city governments, and other community stakeholders.

Citizens have an expectation of privacy, affordability [1], and timely and interactive *information* from a smart city [2]. While innovations in technology continue, citizens are critical about how unvetted smart cities can violate intrinsic rights [3]. People are inventing methods to disguise themselves from surveillance systems using fashionable masks [4]. Citizens also depend on other products to curtain themselves from other devices, such as smart speakers [5][6]. Recent studies have shown that some popular smart technologies, such as smart thermostats, may not provide stated benefits [7]. However, laws are being proposed and passed to ensure the responsibility of the city or company protects the privacy of the citizens [8][9]. Significant costs are incurred when deploying sensors equipped with 5G or WiFi connectivity due to data subscription fees [10][11]. The transformation into a smart city is expensive (e.g., between $30 Million and $40 Billion), and only a few cities are able to obtain the resources required for upgrades [1].

In this paper, we study the finalist applications from the 2015 Smart City Challenge [12] to understand what types of technologies cities requested along with the funding requirements needed to bring smart cities to fruition. Figure 1 describes the main concerns of the citizens and city governments when envisioning smart cities according to the Smart City Challenge applications.

In Section II, we describe a survey to understand the perceptions of smart cities in relation to privacy and cost. In Section III, we perform a detailed textual analysis of the submitted smart city applications. We then propose solutions to the cost and privacy issues in Section IV-A and Section IV-B, respectively. Furthermore, we describe a case study of a privacy-enabled low-cost smart city technology implemented in a U.S. city in Section IV-C. Finally, we summarize our findings in Section V.

II. Surveying perspectives of Smart Cities

We deployed a survey (IRB #13565) to learn about the current understanding of people tangentially involved with smart city implementation or governance. The survey was compromised of eighty-eight questions and had a respondent size of six participants. The average time to complete the

This material is based upon work supported by the National Science Foundation under Grant No. 1952181.

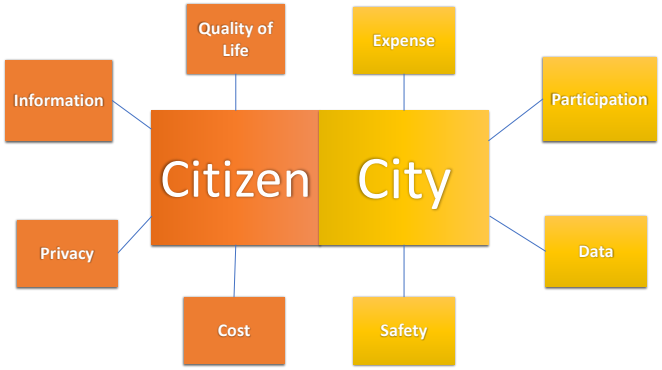


Fig. 1. Citizen and City concerns with Smart City Technology and Services. Citizens concerns centers around information, privacy, costs, and quality of life. City concerns in regards to Smart Cities focus on expenses, safety, data, and participation of the community.

survey was twenty-seven minutes. The participants were able to complete the survey online and it did not require the participants to answer every question. The survey focused on participant's knowledge of privacy, their respective government/companies' involvement with developing smart cities, and the current cost of data collection. Participants were asked about projected costs spent on technologies, knowledge of smart city efforts, and understanding of privacy expectations (see Table I).

Respondents had insightful answers when asked to define smart cities. One participant highlighted "pressing issues for its residents and businesses" and defined a smart city as:

> "One that employs technologies to improve services to the community and/or make government operations more efficient and effective. A truly smart city/community should also be targeting the most important and pressing issues for its residents and businesses not just applying technology for technology's sake." – Survey Respondent

Another respondent stated that a smart city is:

> "A city whose residents are connected by technology, high-speed broadband, providing services online and interactively, tele-health services, using IOT and AI in traffic management, air quality management, parking, waste management, public safety, utilities, autonomous vehicles, etc." – Survey Respondent

Beyond understanding what people defined as a smart city, we wanted to gain insight into the privacy concerns in potential and deployed smart cities. When asked to define privacy, participants highlighted the need to be able to revoke access to their data.

> "...I would include the ability to control or at least delete personal data as well that has been collected especially if the data has become obsolete or inaccurate." — Survey Respondent

When asked about what data privacy protection methods would help improve their willingness to take part in city data sharing, several participants stated they would like the "ability to review" any data collected by the city concerning them. If data is collected anonymously, there is an inherent difficulty when designing systems to review personalized data requests. To solve this, respondents suggest block chain or smart contract techniques to provide anonymous keys to support audit requests. Analyzing the current results, we found common concerns around the concept of privacy. The words *personal, private, uninvited surveillance and protect* are the noticeably frequent words used to describe and articulate how privacy is visualized for both pedestrians and companies. The survey further asks about data sharing. Participants were asked if they were comfortable sharing their data for the development and enhancement of smart cities. However, the results show participants are skeptical about sharing their data with smart cities. The reasons provided by the participants included possible increased policing in under-served communities, vulnerability to data leakage, and not being aware of the purpose of data collection.

TABLE I. THE SURVEY IRB #13565 WAS COMPROMISED OF MULTIPLE-CHOICE QUESTIONS AND SHORT ANSWERS. WE ADD SELECTED QUESTIONS FROM SURVEY RELATED TO DEFINING A SMART CITY, PERSPECTIVES OF PRIVACY, TECHNOLOGY, AND SPENDING.

Number	Question
1	How would you define a smart city/community?
2	How would you define privacy?
3	What data privacy protection methods would increase your willingness to share data with the city?
4	Would you be comfortable sharing personal data within these smart communities?
5	What makes you feel uncomfortable with sharing your personal data within smart communities?
6	How do you use the pedestrian counting data – for what purpose(s)?
7	How much do you spend annually on pedestrian counting data?
8	Where are the locations you need to have pedestrian counting data?

When asked about pedestrian counting for marketing and economic development, some of the pain points concerning pedestrian counting include cost and frequency of pedestrian counting, while privacy is the most valued feature with simplicity as the second most valued. According to the preliminary survey responses, companies spend between $11,000 - $20,000 annually on counting pedestrians. The average amount companies spend on data collection for traffic counts is approximately $25,000 annually; while the maximum allotted amount for traffic counting is $150,000. We also found that, although companies use pedestrian counting for marketing, economic development, safety, and infrastructure development; some of the pain points or challenges concerning pedestrian counting are cost and frequency of pedestrian counting. The most common places for pedestrian counting include intersections, downtown, or shopping areas.

III. Analyzing Smart City Finalist Applications

In 2015, the United States Department of Transportation announced the Smart City Challenge, which asked cities in the US to create an integrated, smart, and efficient transportation system built on data, applications, and technology in an effort to improve the lives of their citizens [12]. The Smart City Challenge received 78 applicants describing what a smart city looked like for their community. From this challenge, the seven cities chosen as finalist include: Columbus (Ohio), Austin (Texas), Denver (Colorado), Kansas City (Missouri), Pittsburgh (Pennsylvania), Portland (Oregon), and San Francisco (California). Figure 2 displays U.S. cities that are applicants 2015 Smart City Challenge Applicants, of these, the red circles denote the seven finalists (the circle area denotes the population size).

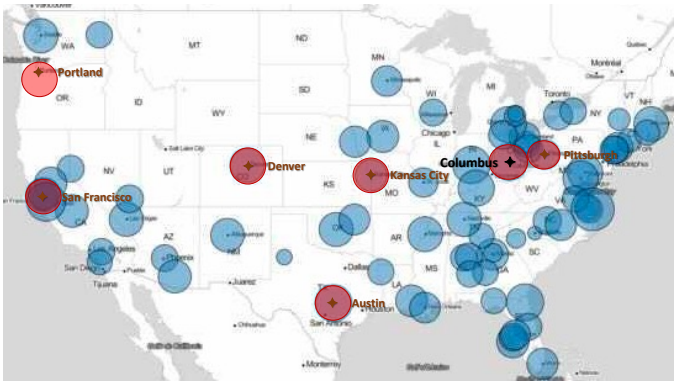


Fig. 2. Smart City Challenge Applicant locations in the United States; red circles denoted the seven finalists.

In an effort to understand the needs and wants of smart cities, we evaluate the finalist from the Smart City Challenge. We perform text analysis methods from each submitted application. We first describe the document preprocessing to transform the PDFs into a usable format (Section III-A). We then perform document similarity, where documents refer to the finalist applications, and we analyze overlap in the application requests (Section III-B). Next, we performed cluster analysis to group the finalist applications by themes according to word usage in each document (Section III-C). Additionally, we performed topic modeling to derive the dominant themes present across the documents and provide insights on what a "smart" city is compromised of (Section III-D). Furthermore, we provide details on the requested technology (Section III-E) and privacy mechanisms (Section III-F) for the Smart City Challenge finalists considered for implementation.

A. Document Preprocessing

Each finalist document was downloaded from the Smart City Challenge website where their vision statements were made publicly accessible as a PDF file [13]. We extracted the textual content from the files with Python code using the PyPDF2 PDF manipulation library [14].

Figure 3 shows the distribution of word tokens across all documents with a truncated tail. The documents are cleaned by

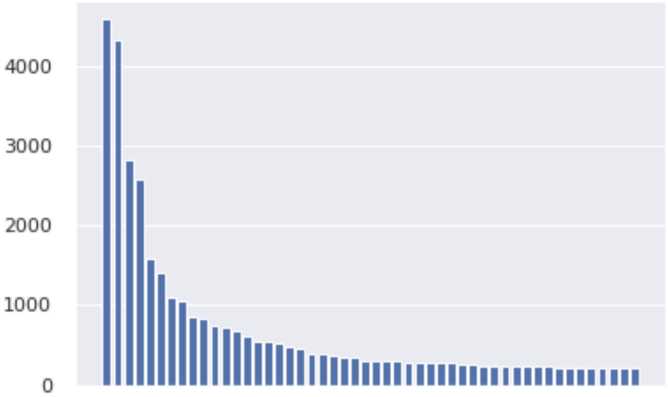


Fig. 3. Token Frequency Distribution across the Smart City Finalist Corpus. The higher frequency tokens are conjunctions and overly common words.

removing stopwords and alphanumeric text, then those words are stemmed. The words are further processed and embedded using natural language processing tools. Stopwords are derived from a list of typically infrequent words or misspellings (e.g. "asd", "buisness") or overly common words (e.g. "the", "a", "is") and overly specific city names. We calculate tf-idf scores [15] with

$$tf_{t,d} = \frac{f_{t,d}}{\sum_{t' \in d} f_{t',d}}, \quad (1)$$

where t, d are the term and document, respectively, and the ratio $f_{t,d}$ is the frequency of a term t in document d. We multiply the tf score for each term by an idf term to account for words that appear in each document. The inverse document frequency is given by

$$idf_t = \log \frac{(1+n)}{(1+df_t)} + 1, \quad (2)$$

for each term t, where n is the total number of documents, and df_t is the number of documents where t appears. We the value $tf_{t,d} * idf_t$ to remove additional terms that add little to no meaning to the content of the topics and themes. The repetition and frequency of unimportant words can influence the text analysis results.

The process of removing alphanumeric terms can alleviate typos as well as unsupportive words. Another pre-processing method we used was stemming. We used the Porter Stemmer to remove the endings of words to set them to the root [16]. When using this Stemmer, you will notice endings such as "ing", "ed", and "es" being removed. With this collection of processed documents, we create a corpus which is used in the analysis steps. With the use of Text Analysis, we can extract the text from these documents to create machine-readable information to perform machine learning tasks to better understand the content.

B. Document Similarity

With a cleaned corpus, we can gain insights of the documents by comparing them with similarity metrics. The document similarity can be calculated by comparing vectorized

documents with the Euclidean distance measure. When two documents are compared, the Euclidean distance score between them acts as a proxy for the similarity of the documents. These distances are visualized in Figure 4.

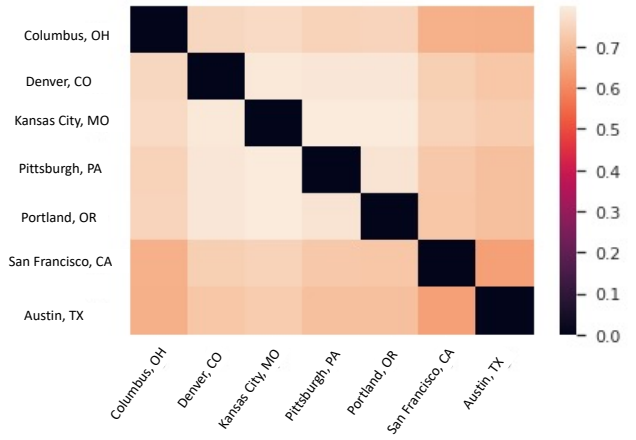


Fig. 4. Similarity Matrix for Finalist Documents; darker shades describe the similarity strength of the Smart City Finalist Applications.

Figure 4 shows relations by varying the intensity of the colors between the range of 0 to 1. The stronger correlations are noted in the darker shades with the values being closer to zero. Weak correlations are shown in the lighter shades with the values being closer to one. From Figure 4, we observe that the documents with the strongest similarity are `San Francisco`, `Columbus`, and `Austin`. The applications for the cities of `San Francisco` and `Austin` show the closest similarity among all documents. The moderate color variation across the corpus insists that the documents have similar content, but also distinctive features as shown in the additional analyses below.

C. Cluster Analysis

Cluster Analysis was performed to group similar documents together. The documents that are found in the same cluster are more similar than those in the other clusters. The cluster analysis was completed using K-Means clustering [17]. K-Means is an iterative centroid based clustering method that creates groups based on closeness or similarity. It uses expectation maximization to place the centroids at an optimal location in the data space such that similar documents are in a cluster and dissimilar documents are not clustered. For the K-Means algorithm, we must define a k value, which is the number of clusters the K-Means Model should produce. To obtain the k value, we evaluated the elbow of the corpus by fitting the model to various values of k between two and six. This elbow analysis of a corpus helped us determine the optimal number of clusters for the respective corpus [18]. The optimal k value was found when the cluster number is set at 4. The corpus was then passed into the K-Means Model to cluster the documents.

To create this visualization, we used Principal Component Analysis (PCA). PCA is traditionally used as a dimension reductionality method. We employ PCA to create a visualization that helps us understand the clusters – we choose the first two principal components as the axes of a two-dimensional plane. This cluster visualization is shown in Figure 5. From this visualization, we notice four distinctive clusters.

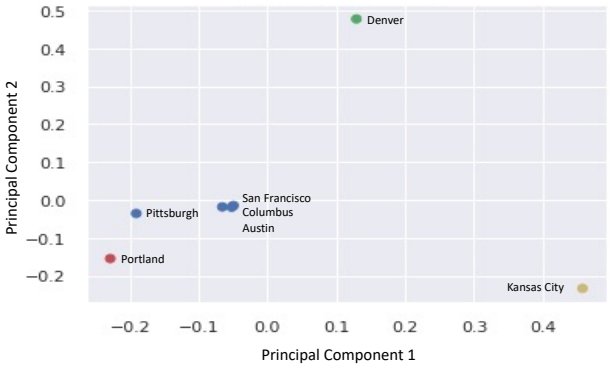


Fig. 5. Two Component PCA for visualizing K-Means Clustering for Smart City Challenge Finalists.

The cities of `Denver`, `Portland`, and `Kansas City` are individual clusters which imply that they differ from one another as well as the large cluster. The larger cluster is comprised of `Pittsburgh`, `San Francisco`, `Columbus`, and `Austin`. The content in these documents are closer in similarity. The centroid of this cluster is positioned on `Columbus`. We also see that there is heavy overlap in `San Francisco`'s and `Austin`'s applications, which are also present in the similarity matrix described in Section III-B.

D. Topic Modeling

To start the Topic Modeling process, we begin to define phrases and vocabulary from the corpus. When building the word dictionary for this model, we choose the words that appear in more than two documents but less than 90% of all documents. After this process is complete, we create a Latent Dirichlet Allocation (LDA) Topic Model [19]. LDA can produce weighted topics based on the analysis of the corpus. Our corpus consists of seven documents and a vocabulary size of 2,282. With this generative probabilistic model, we are able to derive themes and topics that are representative of the corpus. Topics are represented as a list of weighted terms of which we take the top-k. Our LDA model creates 3 topics that are used to discover themes for our documents. In Table II, the three topics are displayed with their respective words and themes.

The terms denoted in gray have little to no contribution to the theme of the cluster. These terms are used based on their assigned weight from the output values of the LDA model and the assigned topics. From Table II, we see that Topic #1 includes four cities (`Columbus, Kansas City, San Francisco, Austin`), Topic #2 includes two cities (`Denver and Pittsburgh`), and Topic #3 includes one city (`Portland`). These applications focus on several topics, but the overall similarity of the document content allowed the model to group and create topics. The documents were

TABLE II. TOPICS AND THEMES DERIVED FROM THE LDA MODEL. THE GROUPS ARE LISTED WITH ASSOCIATED CITIES, TOPICS, AND THEMES. THE TOPICS LISTED CONTAIN THE TOP TEN WORDS.

#	Cities	Topics	Theme
1	Columbus, Kansas City, San Francisco, Austin	grant, proposal, event, digital, automated, university, demonstration, automated vehicle, deploy, tool	Autonomous Technology and Tools
2	Denver, Pittsburgh	component, grant, department transportation, university, benefit, consortium, efficiency, foundation, percent, avenue	Building Partnerships and Infrastructure
3	Portland	device, efficiency, equity, percent, market place, university, cloud, engineering, payment, benefit	Connecting the Collegiate Experience to the City

assigned to these groups by their dominant topic. The themes derived from these groups encompass the various focuses a smart city can have. From these themes, it is implied that cities can become "smarter" with the use of autonomous technology, building partnerships and infrastructure, and connecting to the local universities in the city.

E. Technology Enhancements

To define the essence of a Smart City, we establish the universal technologies requested by smart cities. We introduce definitions needed to build a basis for understanding the foundation of the technologies requested by these Smart City finalists. These definitions provide a foundation to understand the types of connectivity and technology smart cities need to be operational and effective. There are additional technologies, networks, and sensors not mentioned that smart cities can implement in their community.

Many cities are interested in *Dedicated Short-Range Communications* (DSRC), which allows vehicles to communicate with each other and other road users directly. It is a wireless communication technology that can function properly without involving cellular or other infrastructures. It can save lives by cautioning drivers of a looming, threatening situation or occurrence in time to take necessary actions to help evade the situation.

Cities are also interested in technologies that improve efficiency for travelers. *Traffic Signal Priority* (TSP) can be defined as technological set of operational improvements to shorten the wait time at traffic signals for vehicles and prolong the time for green light signals. This can be done by using the existence of vehicular locations and wireless communication to extend the time of the green light at a traffic signal. TSP can be implemented at street intersections. Additionally, pedestrian counters can be implemented in these intersections as well. *Pedestrian counters* can be defined as an electronic device that is used to classify, count, and measure pedestrian traffic amongst along a particular road. These counters can be used to measure the direction of the traffic by time and location. With this technology, corporations can find peak traffic times, identify entry and exit points of travelers, and set travel management protocols. Smart kiosks can serve as a gateway for pedestrian counting as well. A *smart kiosk* is an information kiosk that detects and tracks prospective clients and sends/stores information about these prospects as data for usage [20]. These kiosks can serve as a medium among the citizens, the city, and additional technologies. *Smart parking* technologies can be defined as a strategy that infuses the use of technology to inform citizens about free and occupied parking spaces over the Web or mobile apps. Simultaneously, it can use minimal resources there by reducing time and consumption of fuel.

Electric transportation is any vehicle whose propulsion and accessory systems are powered exclusively by a zero-emissions electricity source. Electric transportation vehicles have rechargeable batteries. The E-bikes use rechargeable batteries battery mounted on the bike frame, and electric bus' battery is under the hood or protective barrier. Cities are interested in planning charging stations to support electric vehicles. Expanding from electronic transportation, smart cities are also interested in implementing autonomous vehicles.

Similarly, cities are interested in promoting *autonomous transportation*, or vehicle that drive with minimal human intervention. Also called driver-less or self-driving vehicles; autonomous transportation requires detailed real-time environmental sensing for detection and classification of surrounding objects along navigation pathways. Cities should also understand the evolving regulations of transportation governing automated vehicles. These electric and autonomous vehicles can include cars, scooters, bikes, and buses [21][22].

TABLE III. REQUESTED TECHNOLOGIES FROM SMART CITY CHALLENGE FINALIST. THE TECHNOLOGIES ARE LISTED IN DESCENDING ORDER. TECHNOLOGIES CAN BE REQUESTED BY ALL CITIES.

Technology request	**Number of Cities**
Smart Traffic Signals	7
Web Applications	7
Electric Vehicle Charging Station	7
Use of Sensors	7
Use of WiFi/Communications	7
Use of Cameras	7
Autonomous Vehicles	6
Connected Vehicles: DSRC technology	5
Smart grid	3
Use of GPS	3
Kiosks	3
Use of Cellphone signals	3
Autonomous home delivery	3
Smart Parking	3
Bike and/or pedestrian Counters	2
Electric Bus	2
Information screens for bus stops	2
Road condition monitors	2
SMART roadside lights	2
Traffic Management Centers	1
Universal smart access card	1
Bike sharing	1
Transportation Hubs	1
Interactive Voice Response	1
Smart Pedestrian Guides	1

In Table III, we display the requested technology for the cities. Among the seven finalists, 25 technologies were requested. The average city requested 12 technologies to be used in their smart city. The amount of technology requested by the city could vary depending on the population as seen in Figure 6.

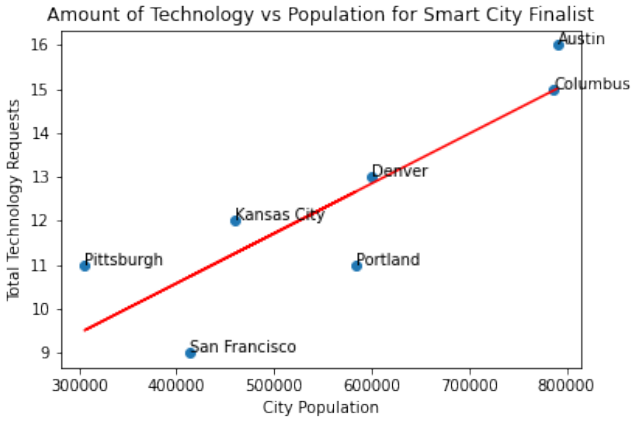


Fig. 6. Comparing the city's population size with the amount of technology requested. A linear regression line shows the projected fit for the cities.

The winning city, `Columbus (Ohio)`, requested a total of 16 technologies to implement their smart city. Following close behind is the city of `Austin, TX` with a total of 15 technology requests. The remaining cities had 13 requests (`Denver, CO`), 12 requests (`San Francisco, CA`), 11 requests (`Kansas City, MO and Pittsburgh, PA`), and 9 requests (`Portland, OR`). To integrate these technologies, the cities use sensors, video, Global Positioning Systems (GPS), and radio signals from pedestrians, vehicles, and equipment. These cities also use these video and GPS feeds for license plate recognition and to track crime-related incidents. The goal of becoming a smarter city revolves around connecting communities to opportunities, decreasing health disparities, reducing air pollution, and increasing the mobility of citizens by relieving congestion of roadways. Assisting low socioeconomic and disabled citizens has risen to the forefront of smart city development strategies. In an effort to make these advancements more inclusive of those communities, smart cities have proposed the use of:

- *Smart kiosks* enable advanced payment options by incorporating additional features, such as braille and voice feedback
- *Electronic signs* can provide visual and audio cues to pedestrians crossing intersections
- *Autonomous car sharing* allows commuters first and last mile transportation with a reduction in costs
- *Information screens* provide real-time transportation updates through audio and video

With the incorporation of these additional technologies, we see these cities becoming more inclusive and smarter for all. On top of an already costly smart city, these specialized technologies introduce additional expenses tied to continuous maintenance for supporting the aforementioned technologies.

F. Privacy Considerations in Smart Cities

A major concern for citizens in literature is understanding how increased technologies in cities will affect their privacy [3] [4][8][9]. Furthermore, cities will become a 24-hour hub for collecting information about the mobility and efficiency of transportation, but also personally identifying information of its' travellers [20]. In the Smart City Challenge [12], we examine how applicants describe risks and mitigation strategies for the deployment of technologies to their cities. From these concerns, we focus on the risks associated with residents and visitors of the city. The main concerns listed by the cities include data sharing, individual privacy, system security, data privacy, and data management.

In Table IV, we reviewed these Smart City Finalist proposals and assessed a score based on a Likert Scale (Excellent, Average, Poor) from the five central themes found in the documents: Data Sharing, Individual Privacy, System Security, Data Privacy, & Data Management. From the proposal and discussion, a city will receive a rating for all five privacy categories:

- Excellent: The proposal has thorough discussion about the risks and mitigation strategies related to topic and a solid plan of action.
- Average: The proposal has moderate to little discussion about the risks and mitigation strategies related to topic and a general plan of action.
- Poor: The proposal has little to no discussion about the risks and mitigation strategies related to topic and no plan of action.

We give each of the five categories a definition to describe their clarity of the topic in the documents. **Data management** outlines access control procedures, storage schema, and storage policies for smart city data and databases. **Data privacy** entails the encryption of items in the data, and what information is stored from the citizens and anonymization schemes. **Data sharing** includes the procedures and policies of which the smart city data will be shared with organizations, entities, or the public. **Individual privacy** focuses on the protection of citizens in the city. This protection could include, but not limited to, encryption schemes, consent documents, and privacy mitigation techniques. **System security** details the overall protection mechanisms for the smart city infrastructure.

Data sharing and data privacy concerns are addressed by the majority (4 of 7) of the cities. Strategies for addressing data sharing included access management, encryption, and anonymization. Individual privacy, system security, and data management categories are each addressed by three of the cities. Columbus is the only city without a risk analysis in their proposal. This city will develop their plan during the implementation of their city, but would this be enough? Immediately after winning, Columbus created the Smart City Program Office to assess possible risks and mitigate them. Of the finalists, none of these cities provide a detailed description

of the protection they will provide their citizens in their proposals. To mitigate the proposed risks these cities seek to: (1) implement standards from government and industry, (2) anonymize or mask sensitive personal data, and (3) partner with cyber-security experts and government.

TABLE IV. RATING OF PRIVACY DISCUSSION BY CITY. EACH CITY RECEIVES A RATING (POOR, AVERAGE, OR EXCELLENT) BASED ON FIVE CATEGORIES.

City	Data Sharing	Individual Privacy	System Security	Data Privacy	Data Management
Columbus, OH	Poor	Poor	Poor	Poor	Poor
Austin, TX	Poor	Poor	Excellent	Excellent	Excellent
Denver, CO	Poor	Poor	Poor	Average	Poor
Kansas City, MO	Poor	Average	Excellent	Poor	Poor
Pittsburgh, PA	Poor	Poor	Poor	Average	Poor
Portland, OR	Average	Poor	Poor	Average	Average
San Francisco, CA	Poor	Poor	Poor	Poor	Poor

Beyond security breaches and attacks, what protection will these cities use to ensure the privacy of those who want to remain anonymous in an "always on" city? Researchers have investigated the concerns of privacy leaks and the types of privacy leaks on social media [23]. These privacy leak concerns can be expected in a smart city where citizens are continually being monitored. To help cities protect their citizens, we propose the use of a visual mitigation library used for videos and images based on existing literature [24]. This work provides a foundation for several mitigation techniques used for social media networks; however these same technologies can be implemented to protect the citizens from surveillance concerns and privacy issues. Beyond the citizen's concern for anonymity or protection of minors, there is a concern for the type of information that is revealed in a public setting.

IV. Discussion

We provide essential interpretations and considerations for smart city infrastructure. Based on the Finalist's Applications, we propose the use of a low-cost and privacy-enabled smart city. Furthermore, we explore an existing smart city technology and provide discussion its' privacy-enabled and low-cost features.

A. Proposed Solution: Low-Cost Smart Cities

Smart City projects can be expensive to deploy and manage. Cities around the world such as San Diego, New Orleans, London, and Songdo have either proposed or invested in Smart City projects that cost between $30 Million and $40 Billion. In addition to the cost of deploying and maintaining the IoT devices themselves, a significant portion of the expense is a result of providing Internet connectivity via 5G or WiFi to those devices. These costs are a major barrier to the widespread deployment of Smart City technology and the social benefits that may ensue from that technology [25].

To alleviate the costs, opportunistic communication, such as Delay Tolerant Networks (DTNs) can be used as a backbone for Smart City communication to facilitate data that does not have real-time Quality of Service (QoS) constraints. DTNs traditionally provide opportunistic networking connections in areas with little to no infrastructure. Messages are delivered with some delay which is directly correlated with the layout, density, and mobility of nodes in the network [26][27]. Recognizing that some data are needed in real-time, edge-computing can be utilized as long as the placement of internet-connected nodes are optimized in the network. For data that can tolerate delays, the natural movement of people and vehicles through a city to transfer data between nodes. In this way, the citizens become an integral part of the smart city network itself.

In order for low-cost Smart Cities to flourish and DTNs as backbone to be practical, both the technology questions related to the devices and the network itself, as well the social aspects of how people and vehicles move through a city must be addressed. For almost 20 years there has been a substantial amount of research in opportunistic communications and delay tolerant networks; unfortunately real-world deployments traditionally fall short of their simulated counterparts [28]. Related efforts, [29][30][31][32][33][27][34][35][36], have proven the ability to deliver messages when connections are intermittent, but generally are limited to performing within simulation environments [37].

B. Proposed Solution: Privacy Mitigation Library

Cameras can be integrated into several requested technologies which makes it popular commodity. We consider how the privacy risks of cameras and video surveillance can be mitigated in smart city infrastructure. In a city where facial recognition systems are used can lead to privacy leaks due to individual privacy rights. Pedestrians carry identification, purchase items with their credit or debit cards, use physical keys to enter restricted areas, and virtual passcodes to access sensitive information. These types of sensitive content will be captured in those videos and image feeds [38][39], with the use of obfuscation we can ensure that content will not be leaked to others. Studies have shown that the use of obfuscation methods [40][41][42] can protect individual privacy. These obfuscation methods can include blurring, blocking, adversarial noise, or replacing items in visual content. Methods such as blurring and blocking alters the pixelation of the visual content to provide distortion to the human eye. These methods can be added on to objects, faces, and text in visual content. The technique of adversarial noise [43] adds a few pixels that can (1) impede a computer's ability to learn anything from the visual content even if it is in their possession, and (2) still allow the images to be visible to humans. To protect individuals identities, studies have suggested face swapping [44][45][46] which can switch detected faces of citizens with a pre-existing library of faces at their disposal.

To address this concern, we suggest the deployment of the *ViperLib*. This mitigation library will allow the Smart Cities to select from a library of mitigation techniques that can be integrated into their systems. As proposed by [24], mitigation techniques can be integrated into mobile applications, servers, IoT devices, and comprehensive systems. Techniques such as obfuscation (e.g., adversarial noise, blurring, blocking), interception, and blind vision can be integrated into this library easily ready for use. The library can also facilitate active engagement strategies for alerting authorized personnel about pertinent privacy concerns and suggesting the possible mitigation strategies for that visual content. These types of alerting strategies are similar to *Chaperone Bot or Privacy Patrol* from previous works [24]. These alerting systems can alert officials of private information that is displayed in public settings before the data is stored without additional protection. We hope that this can provide security to the data privacy and storage methods smart cities will implement. These alerting strategies are important to provide a human-in-the–the-loop system at various phases of the deployment and collection processes that these cities will have.

The *ViperLib* open-source library can be integrated into existing "off-the-shelf" packages. Citizens can select the privacy protection features that must be integrated into deployed systems. Such libraries can provide safety, security, and peace of mind to the citizens that reside in those areas.

C. Case Study: Deployed Technology in a Smart City

Smart city technology must be reliable, low-cost, and consider privacy to attract citizens to engage with those platforms. The Smart City Applications Platform (SCAP) is an example of a privacy-aware system coupled with reliable and effective management. It serves as a strong example for organizations to model pedestrian counting and computer vision technologies in smart cities. In this study, SCAP is deployed in city *C*. SCAP was created by a major utility company. This platform consists of a complete hardware and software solution which identifies various types of moving objects common to an outdoor urban environment such as bicycles, pedestrians, and scooters. At its core, SCAP is a Field Node with computer vision software that analyzes data from a high-definition camera on an edge compute device and transforms it into object count data (Figure 7A). The Field Node is available in a stand-alone enclosure or as an integrated subsystem of a digital information kiosk as seen in Figure 8. The Field Node kiosk is integrated in city *C*'s downtown infrastructure. This data can be uploaded to the Cloud (Figure 7B) as anonymized statistics after data analysis is complete (Figure 7C). The data can then be viewed in a portal or accessed via an Application Programming Interface (Figure 7D).

In order to provide data in as real-time of a manner as possible, the video analytics data is sent from the local device to the Cloud. Should the network connection be lost, data is queued in the Field Node compute device and transmitted once the network returns. This connection uses Message Queuing Telemetry Transport (MQTT) between the edge and cloud for

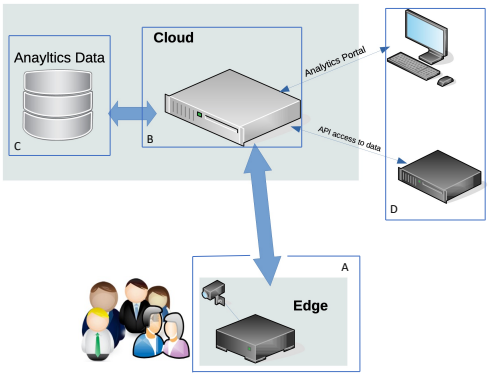


Fig. 7. High-level overview of the deployed Smart City Applications Platform (SCAP).

communication. MQTT is a standard publish and subscribe technology that uses machine-to-machine communication with low bandwidth requirements. The cloud database is set up in a cluster for backup and redundancy purposes. SCAP utilizes a cloud based user management system to control access to the Portal and to the Cloud API. In order to access any system information or data, a username and password must be created. The Platform is designed to utility-grade cybersecurity and network security standards. It is important to note that the SCAP software does not collect or record personally identifiable information, such as facial images, phone numbers or mobile phone MAC addresses. Rather anonymized target object count data is collected and provided to the user. Furthermore, all video is processed on a local computer and no images are recorded or stored ensuring piece of mind for citizens and visitors.

In consideration of robust physical security, the SCAP Field Node or digital kiosk features an enclosure with a specially keyed locking system. Both the incoming and outgoing data to the Field Node is encrypted. Through the monitoring and control software, licenses for the Field Nodes can be remote enabled or disabled. Each Field Node utilizes a compute device with storage capability. As a result, should the Field Node become compromised, the larger system is unaffected.

While the SCAP Field Nodes have the capability to work with a variety of wired and wireless data back-haul networks, the most common type is anticipated to be cellular. One of the major advantages of the SCAP is that it has low bandwidth requirements. This allows the use of the lower bandwidth CAT-M1 network when cellular communications are required. As the Smart City Applications Platform is still in its infancy and undergoing field trials, there will be ample opportunity to reduce the cost of both system deployment and operations. For example, the complexity of mounting the Field Node equipment to appropriate street furniture or buildings will be simplified and as system requirements are better understood, optimization of the Field Node components will allow for a reduction in the Bill of Material costs as well as annual operating costs.

(a) (b) (c)

Fig. 8. Field Node Designs provided by Smart City Applications Platform. (a) Field Node Integrated in a Kiosk, (b) Opened Field Node Kiosk deployed in the city, (c) Rendering of Field Node Integrated with a Light Pole (refer to Pole-Mountable Camera Support Structure, US Design Patent D902,985 S) [47]

V. Conclusion and Future Work

The finalist of the Smart City Challenge showcased what their city would look like throughout their application. We analyzed the applications to reveal the intricacies in the expectations of smart cities. When becoming a smart city, the city government and citizens could create multiple goals or milestones. From our analysis, an innovative smart city proposal would include creative approaches to deploying autonomous technology and tools such as drone delivery, building partnerships and infrastructure, and bridging the local collegiate experience in the smart city. This analysis also alluded that the typical smart city will require 12 new technologies on average to become a smart city, which is more than a city with smart technology. With creativity and development of smart city infrastructure, it is important to take cost and privacy in consideration. Using our survey and analysis, we find that privacy and cost can continue to be concerns for citizens and corporations in these environments. In the analysis of the Finalist's proposals, we find that the discussion of privacy and cost is not at the forefront of developer concerns; rather technological innovation. The winning city from the Smart City Challenge proposed innovative ways to develop their city, but showed less interest in privacy and cost than other applicants.

The analysis and evaluation of smart cities using the 2015 Smart City Challenge and deployed surveys are important to understand the needs and wants of smart cities, but also understand perspectives of individuals in those cities. These insights show the disconnect between citizens and organization who develop these smart cities. With the input of citizens for smart cities, the organizations will be able to create inclusive, adaptable and trusted relationships to aid in the acceptance and assimilation to the futuristic growth of the city.

In summary, this paper argued that smart cities have the capability to be both private and inexpensive in deployment and long-term sustainability. During planning and implementation of these cities, officials along with citizens should further consider the high cost and privacy concerns associated with their development choices. The need for privacy mitigation in smart cities extends from the protection of personally identifying information to the choice of anonymity and protect of minors. Beyond the deployment of the *ViperLib*, we proposed the use of DTNs to lower the cost of smart cities and allow citizens to assist in the transmission of data across the city. Deploying traditional IoT infrastructure is prohibitively expensive for most cities and expanded development introduces privacy risks. However, low-cost smart cities and privacy-enabled technologies can achieve the goals of smart cities while allowing citizens to feel secure and protected.

Future research considers the potential effects of security for cyber-physical systems in real IoT deployments. To do this, we will collaborate with Louisville, Kentucky, a Smart City Applicant, to discuss future strategies and deployment plans for *ViperLib* as part of NSF Grant (#1952181).

Acknowledgment

This material is based upon work supported by the National Science Foundation under Grant No. 1952181. Thank you Jason Clermont and our partners at the Louisville Downtown Partnership and Duke Energy One.

References

[1] Jasmine DeHart, Corey E Baker, and Christan Grant. Considerations for designing private and inexpensive smart cities. *The Sixteenth International Conference on Wireless and Mobile Communications*, 2020.

[2] César R Cortez and Victor M Larios. Digital interactive kiosks interfaces for the gdl smart city pilot project. 2015.

[3] Joshuas Emerson Smith. As San Diego increases use of streetlamp cameras, ACLU raises surveillance concerns, August 2019. last accessed on 09/01/2020.

[4] Adam Harvey. Cv dazzle: Camouflage from computer vision. *Technical report*, 2012.

[5] Jack Morse. There's a privacy bracelet that jams smart speakers and, hell yeah, bring it. last accessed on 09/01/2020.

[6] Yuxin Chen, Huiying Li, Shan-Yuan Teng, and Steven Nagels Zhijing Li. Wearable microphone jamming. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 1–12, 2020.

[7] Alec Brandon, Christopher M Clapp, John A List, Robert Metcalfe, and Michael Price. Smart tech, dumb humans: The perils of scaling household technologies. *Work*, 2021.

[8] Cory Doctorow. The case for ... cities that aren't dystopian surveillance states. *The Guardian*, January 2020. last accessed on 09/01/2020.

[9] Hannah Devlin. AI systems claiming to 'read' emotions pose discrimination risks. *The Guardian*, February 2020. last accessed on 09/01/2020.

[10] Josep Paradells, Carles Gómez, Ilker Demirkol, Joaquim Oller, and Marisa Catalan. Infrastructureless smart cities. Use cases and performance. *2014 International Conference on Smart Communications in Network Technologies, SaCoNeT 2014*, pages 1–6, 2014.

[11] Esther Max-Onakpoya, Oluwashina Madamori, Faren Grant, Robin Vanderpool, Ming-Yuan Chih, David K Ahern, Eliah Aronoll-Spencer, and Corey E Baker. Augmenting cloud connectivity with opportunistic networks for rural remote patient monitoring. In *2020 International Conference on Computing, Networking and Communications (ICNC)*, pages 920–926. IEEE, 2020.

[12] U.S. Department of Transportation. Smart city challenge, Jun 2017. last accessed on 09/01/2020.

[13] U.S. Department of Transportation. Smart city challenge vision statements, 04 2016.

[14] Mathieu Fenniak. Home page for the pypdf2 project, 12 2013.

[15] Karen Sparck Jones. A statistical interpretation of term specificity and its application in retrieval. *Journal of documentation*, 1972.

[16] Martin F Porter. An algorithm for suffix stripping. *Program*, 1980.

[17] James MacQueen. Some methods for classification and analysis of multivariate observations. In *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, volume 1, pages 281–297. Oakland, CA, USA, 1967.

[18] Ville Satopaa, Jeannie Albrecht, David Irwin, and Barath Raghavan. Finding a" kneedle" in a haystack: Detecting knee points in system behavior. In *2011 31st International Conference on Distributed Computing Systems Workshop*, pages 166–171. IEEE, 2011.

[19] David M Blei, Andrew Y Ng, and Michael I Jordan. Latent dirichlet allocation. *The Journal of Machine Learning Research*, 3:993–1022, 2003.

[20] Ruben Sánchez-Corcuera, Adrián Nuñez-Marcos, Jesus Sesma-Solance, Aritz Bilbao-Jayo, Rubén Mulero, Unai Zulaika, Gorka Azkune, and Aitor Almeida. Smart cities survey: Technologies, application domains and challenges for the cities of the future. *International Journal of Distributed Sensor Networks*, 15(6):1550147719853984, 2019.

[21] Mojdeh Azad, Nima Hoseinzadeh, Candace Brakewood, Christopher R Cherry, and Lee D Han. Fully autonomous buses: A literature review and future research directions. *Journal of Advanced Transportation*, 2019, 2019.

[22] Mark Campbell, Magnus Egerstedt, Jonathan P How, and Richard M Murray. Autonomous driving in urban environments: approaches, lessons and challenges. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 368(1928):4649–4672, 2010.

[23] Jasmine DeHart, Makya Stell, and Christan Grant. Social media and the scourge of visual privacy. *Information*, 11(2):57, 2020.

[24] Jasmine DeHart and Christan Grant. Visual content privacy leaks on social media networks. *arXiv preprint arXiv:1806.08471*, 2018.

[25] Oluwashina Madamori, Esther Max-Onakpoya, Christan Grant, and Corey Baker. Using delay tolerant networks as a backbone for low-cost smart cities. In *2019 IEEE International Conference on Smart Computing (SMARTCOMP)*, pages 468–471. IEEE, 2019.

[26] Ari Keränen, Jörg Ott, and Teemu Kärkkäinen. The one simulator for dtn protocol evaluation. In *Proceedings of the 2nd international conference on simulation tools and techniques*, page 55. ICST (Institute for Computer Sciences, Social-Informatics and Telecommunications Engineering), 2009.

[27] Pan Hui, Jon Crowcroft, and Eiko Yoneki. Bubble rap: Social-based forwarding in delay-tolerant networks. *IEEE Transactions on Mobile Computing*, 10(11):1576–1589, 2011.

[28] Corey E Baker, Allen Starke, Tanisha G Hill-Jarrett, and Janise McNair. In vivo evaluation of the secure opportunistic schemes middleware using a delay tolerant social network. In *2017 IEEE 37th International Conference on Distributed Computing Systems (ICDCS)*, pages 2537–2542. IEEE, 2017.

[29] Roy Cabaniss, Srinivasa S Vulli, and Sanjay Madria. Social group detection based routing in delay tolerant networks. *Wireless networks*, 19(8):1979–1993, 2013.

[30] Paolo Costa, Celicia Mascolo, Mirco Musolesi, and Gian Pietro Picco. Socially-aware routing for publish-subscribe in delay-tolerant mobile ad hoc networks. *IEEE Journal on Selected Areas in Communications*, 26(5):748–760, 2008.

[31] Elizabeth M Daly and Mads Haahr. Social network analysis for information flow in disconnected delay-tolerant manets. *Mobile Computing, IEEE Transactions on*, 8(5):606–621, 2009.

[32] Wang Gang, Wang Shigang, Liu Cai, and Zhang Xiaorong. Research and realization on improved manet distance broadcast algorithm based on percolation theory. In *2012 International Conference on Industrial Control and Electronics Engineering (ICICEE)*, pages 96–99. IEEE, 2012.

[33] Wei-jen Hsu, Debojyoti Dutta, and Ahmed Helmy. Csi: A paradigm for behavior-oriented profile-cast services in mobile networks. *Ad Hoc Networks*, 10(8):1586–1602, 2012.

[34] Anders Lindgren, Avri Doria, and Olov Schelen. Probabilistic routing in intermittently connected networks. In *Service Assurance with Partial and Intermittent Resources*, pages 239–254. Springer, 2004.

[35] Mirco Musolesi and Cecilia Mascolo. Car: context-aware adaptive routing for delay-tolerant mobile networks. *IEEE Transactions on Mobile Computing*, 8(2):246–260, 2009.

[36] Amit Kr Gupta, Jyotsna Kumar Mandal, and Indrajit Bhattacharya. Comparative performance analysis of dtn routing protocols in multiple post-disaster situations. In *Contemporary Advances in Innovative and Applicable Information Technology*, pages 199–209. Springer, 2019.

[37] Andreea Picu and Thrasyvoulos Spyropoulos. Dtn-meteo: Forecasting the performance of dtn protocols under heterogeneous mobility. *IEEE/ACM Transactions on Networking*, 23(2):587–602, 2014.

[38] Roberto Hoyle, Robert Templeman, Denise Anthony, David Crandall, and Apu Kapadia. Sensitive lifelogs: A privacy analysis of photos from wearable cameras. In *Proceedings of the 33rd Annual ACM conference on human factors in computing systems*, pages 1645–1648. ACM, 2015.

[39] Mohammed Korayem, Robert Templeman, Dennis Chen, David Crandall, and Apu Kapadia. Enhancing lifelogging privacy by detecting screens. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, pages 4309–4314. ACM, 2016.

[40] Tribhuvanesh Orekondy, Mario Fritz, and Bernt Schiele. Connecting pixels to privacy and utility: Automatic redaction of private information in images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8466–8475, 2018.

[41] Yifang Li, Nishant Vishwamitra, Bart P Knijnenburg, Hongxin Hu, and Kelly Caine. Blur vs. block: Investigating the effectiveness of privacy-enhancing obfuscation for images. In *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 1343–1351. IEEE, 2017.

[42] Terrance Edward Boult. Pico: Privacy through invertible cryptographic obscuration. In *Computer Vision for Interactive and Intelligent Environment (CVIIE'05)*, pages 27–38. IEEE, 2005.

[43] Alexey Kurakin, Ian Goodfellow, and Samy Bengio. Adversarial machine learning at scale. *arXiv preprint arXiv:1611.01236*, 2016.

[44] Iryna Korshunova, Wenzhe Shi, Joni Dambre, and Lucas Theis. Fast face-swap using convolutional neural networks. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3677–3685, 2017.

[45] Bingquan Zhu, Hao Fang, Yanan Sui, and Luming Li. Deepfakes for medical video de-identification: Privacy protection and diagnostic information preservation. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, pages 414–420, 2020.

[46] Sachit Mahajan, Ling-Jyh Chen, and Tzu-Chieh Tsai. Swapitup: A face swap application for privacy protection. In *2017 IEEE 31st International Conference on Advanced Information Networking and Applications (AINA)*, pages 46–50. IEEE, 2017.

[47] James Cleveland, Gregory S Tribbe, Louis Lombardi, Gilbert DeFreitas, and Peter Henderson. Pole-mountable camera support structure, November 24 2020. US Patent App. 29/713,374.

Towards Frictional 3D Object Shape Scanning and Reconstruction by Means of Vibrissa-like Tactile Sensors

Lukas Merker
Department of Mechanical Engineering
Technische Universität Ilmenau, Germany
e-mail: lukas.merker@tu-ilmenau.de

***Abstract*—Interacting with the environment, mobile robots could benefit from advanced tactile sensors complementing optical sensors, consequently gathering overlapping information. In the animal kingdom, there are numerous examples of tactile sensors just as the vibrissae of rats. These tactile hairs enable the animals, inter alia, to detect object shapes based on few contacts. Vibrissae themselves consist of dead tissue and, thus, all sensing is performed in the support of each vibrissa. This characteristic and simple measuring structure provides an inspirational framework for developing tactile sensor concepts. Within the present paper, we take advantage of a recently developed mechanical model of a vibrissa-inspired sensor for 3D object shape scanning and reconstruction. It consists of a cylindrical, one-sided clamped bending rod, which is swept along a 3D object surface undergoing large deflections. Instead of assuming an ideal contact (no frictional effects) as in previous publications, the contact model includes Coulomb's friction. Simulating frictional scanning sweeps, the focus is on both generating the support reactions (observables) at the base of the rod theoretically and subsequently using these quantities in order to reconstruct a sequence of contact points approximating the scanned object surface. Our investigation reveals that (of course) the generated support reactions are affected by friction, but (surprisingly) the reconstruction error seems to be largely invariant against friction.**

Keywords–Vibrissa; tactile sensor; surface sensing; surface reconstruction.

I. INTRODUCTION

Object shape detection and obstacle avoidance are key topics in mobile robotics [1][2]. In nature, almost all species solve these tasks by sensing overlapping information of the environment, e.g., combining information provided by the sense of vision and touch. However, in mobile robotics, the majority of data is frequently gathered solely relying on optical (vision) sensors – the NASA Rover Perseverance[1] of the 2020 Mars mission serves as a striking example. In some cases, completely dispensing with tactile sensing carries the risk of missing information under poor visibility and impedes the interaction with environmental objects. Therefore, advanced tactile sensors hold great potential to complement optical sensors. Developing tactile sensors, engineers often draw their inspiration from biology. Besides the human skin, another prominent and particularly well-researched tactile sense organ are the mystacial vibrissae in the snout region of the rat. Based on few contacts with an object of interest, these tactile hairs allow for the detection of the object's distance, orientation, shape and texture [3]–[5]. Moreover, vibrissae are used for sensing fluid flows [6].

[1]https://mars.nasa.gov/mars2020/spacecraft/rover/cameras/, June 09,2021

Basically, a vibrissa consists of a long and slender hair-shaft with no receptors along its length, which is conically and pre-curved shaped [7][8] and supported by the Follicle Sinus Complex (FSC) [9][10]. Making contact with an object of interest, mechanical stimuli are transmitted through the hair-shaft to the FSC, where the actual sensing is realized by a wide variety of mechanoreceptors. Despite the fact that it is not conclusively clarified how exactly animals manage to determine object features, e.g., object shapes [11], the remarkable structure of natural vibrissae has frequently been transferred into technical sensor concepts over the last decades. A variety of these approaches pursue the goal of scanning and reconstructing object shapes. Basically, these approaches all share a common structure: a slender, more or less flexible probe, mimicking the vibrissal hair-shaft, one-sided attached to some kind of measuring device, representing the FSC. Besides this basic structure, approaches differ considerably in modeling the probe and its support, as well as in the evaluated signals (observables) and the procedure of localizing contact points in space. For the process of object scanning, two different approaches have been established in literature [12]:

- *The tapping strategy*: The object is scanned by repeatedly pushing the probe against various points of the object. In doing so, the artificial vibrissa is retracted from the object right after the very first contact (small pushing angles). Consequently, the deformations of the probe remain small and a linear bending theory is sufficient to accomplish the localization of the contact point [13]–[16] based on measurements of the curvature or torque [16], angles and/or moments [13]–[15] at the support of the probe.

- *The sweeping strategy*: For object scanning, the probe is pushed against an object far beyond the very first contact, consequently undergoing large deformations and sliding over the objects surface [12][17]. Therefore, a highly flexible and elastic probe is mandatory.

The latter strategy is a particularly promising approach due to its passive feasibility [12], e.g., using the robot movement as an actuation without the need of repeatedly making and releasing contact. Moreover, due to the high flexibility of the used probe, the sensor design offers high collision robustness. Passing an object, the probe bends out of way, consequently sweeping along the object and continuously transducing signals to the measuring unit, see Figure 1. Due to these advantages, we focus on the sweeping strategy, which has received less attention in the literature than the tapping one [12].

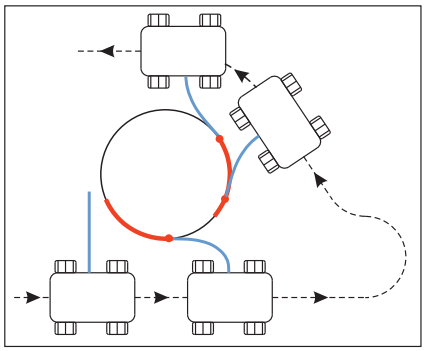


Figure 1. Schematic application example of a rover exploring a circular object adapted from [20].

In [18], the authors used a steel probe attached to a six axis hub load cell to measure the support reactions (three forces and three moments) at the base of the probe. The whole assembly was swept along several edged 3D objects while measuring the support reactions. During scanning, lateral slip of the artificial vibrissa was prevented by actively adjusting the scanning direction w.r.t. the surface normal, which was continuously determined exploiting the measured support reactions. An initial-value problem based on a non-linear bending theory was used to determine the contact position in space. Finally, the authors successfully reconstructed a cloud of contact points and, thus, provided a proof of concept. A similar approach was used in [17] considering the process of object scanning and reconstruction as a plane problem and focusing on an analytical treatment. The presented sensor concept consists of a one-sided clamped probe, which is swept along a 2D object contour. Just as in [18], the sensing concept is based on the support reactions, i.e., two forces and one moment at the clamping of the probe. By analogy with [18], the deformation of the probe is described using an arc-length parameterization of the rod axis and using non-linear Euler-Bernoulli bending theory to express its curvature. However, the major difference compared to [18] is that the problem is sub-divided into two (inverse) steps, which are both treated fully analytically and validated using an experimental setup:

- *Step 1*: simulating scanning sweeps in order to generate the support reactions (observables) at the base of the rod theoretically, assuming the object contour to be known, and
- *Step 2*: using the generated support reactions from the previous step in order to reconstruct a sequence of contact points, finally approximating the original object contour.

In contrast, [18] used a numerical approach and limited the consideration to Step 2.

In [12], a sweeping-based reconstruction algorithm was presented, which is based on repeatedly inferring from one contact point to the next one by continuous measurement of the moment and rotation angle at the base of the probe. This method dispenses the need of force measurements, but it is limited to tangential contacts along the probe and is not suitable for 3D reconstruction.

In [19], the authors modeled an artificial vibrissa using a multi-body system. There, the step of generating the support reactions in 3D space was included, but only for a preset (a priori known) impressed force. In addition, this force was restricted to be always perpendicular to the probe tangent at the contact point. Finally, focusing on the design of the probe considering pre-curved and tapered shapes, the reconstruction process was realized based on a neural network - in this way, unique mappings between the input and output signals are proven but without any analytical correlations.

Summarizing the mentioned approaches, the focus is frequently on the reconstruction of the contact point location in space based on different measured signals at the support of the probe (Step 2). Thus, a theoretical generation of the support reactions (Step 1) is rarely taken into account. However, including the latter process into consideration holds great potential to gain insights into the scanning process and to provide a simulation tool, allowing for parameter studies to investigate the sensor system without having to rely on a large number of experiments. For instance, the theoretical generation of support reactions allows us to provide a data basis for different predefined (well-known) friction parameters. In practice, such parameters are mostly unknown and hard to predefine. Therefore, following [17] and [1], we focus on both mentioned Steps 1 and 2. Our investigation differs from previous works in using the sweeping scanning strategy for 3D object reconstruction passively, i.e., without actively adjusting the scanning direction w.r.t. surface normals of the object [18]. Moreover, as far as the authors know, it is the first work to consider frictional effects in that context. To date, it is an open question how friction affects the 3D reconstruction result when using the sweeping scanning strategy. For instance, imagining two objects of the same shape made of different materials, it is unclear if different friction pairings (probe/object) might erroneously lead to different shape estimations.

In contrast to [18], our investigations are not limited to edged 3D objects. Instead, objects with wide-ranging curvature are scanned in the presence of lateral slip. The presented model differs from [19] in considering a reaction force resulting from the sweep of a rod along a mathematically described object surface instead of a preset impressed force. Unlike [12], our reconstruction method is suitable for 3D object shapes. Moreover, it is not limited to tangential contacts, but also includes tip contacts between the probe and the object. Finally, we present a novel approach to approximate the micro-mechanics of the contact during object scanning based on Coulomb's law of friction, which distinguishes the present paper from similar works.

The remainder of the paper at hand is structured as follows: In Section II, we present the mechanical model of the vibrissa-inspired sensor for 3D shape scanning and reconstruction, starting with the basic setup in Section II-A. In Section II-B, we focus on the contact mechanics during object scanning assuming Coulomb friction. After some preliminary geometrical considerations, we derive the deformation equations of the probe. Then, following the procedure of [17], we separately analyze the above-mentioned Steps 1 and 2, respectively. In Section III, we firstly demonstrate the general appearance of the simulated scanning sweeps. Subsequently, the simulation results of Steps 1 and 2 are analyzed with the overall goal of

clarifying the influence of the friction coefficient on both, the support reactions at the base of the probe and the reconstruction error. Finally, the results of the present paper are summed up and some future research subjects are identified in Section IV.

II. MODELLING

Within this section, we build up and mathematically describe the mechanical model of the vibrissa-like tactile sensor concept step by step, starting with the basic setup.

A. Basic setup

The mechanical model consists of two interacting components in a fixed Cartesian coordinate system (x, y, z) (global frame), see Figure 2:

1. a highly flexible probe, one-sided clamped at it's lower end ("foot", "base") with constraint direction $\vec{e}_z$;
2. a fixed 3D target object.

The probe is modeled as a circular cylindrical rod with an originally straight axial line. Its shape is characterized by the length L and a constant circular cross-section, resulting in a constant second moment of area I. Guided by the biological paragon vibrissa, the cross-sectional dimensions of the rod are extremely low compared to its length. Moreover, we assume the rod to consist of a homogeneous and isotropic, linear elastic Hooke's material. Based on these assumptions, the mechanical behavior of the rod is essentially determined by a constant Young's modulus E resulting in a constant bending stiffness EI. From the outset, we introduce the following units of measure in order to allow any kind of scaling [17]:

$$\begin{aligned} [length] &:= L, \\ [force] &:= \frac{EI}{L^2}, \\ [moment] &:= \frac{EI}{L}. \end{aligned} \tag{1}$$

Remark 1. *We point out that, e.g.,* $[lenght] := L$ *denotes that all lengths are measured in the unit and value of L.*

The object is assumed as a rigid body with a strictly convex, smooth surface $z = C(x, y)$. Within the present paper, we consider the example of an elliptic paraboloid

$$(x, y) \mapsto C(x, y) = a\,x^2 + b\,y^2 + h, \tag{2}$$

with $a, b > 0$ and $h \in (0, 1)$ and unit normal vector

$$\vec{n}(x, y) = \frac{1}{\sqrt{4a^2x^2 + 4b^2y^2 + 1}} \begin{pmatrix} -2ax \\ -2by \\ 1 \end{pmatrix} \tag{3}$$

The scanning sweep of the rod along the object's surface is realized by a kinematic drive, i.e., the clamping position $P_0(x_0, y_0, 0)$ of the rod (system input) is shifted incrementally along a straight trail in the x-y-plane. This process is considered quasi-statically. Interacting with the object, the rod gets bent in the inference of some unknown contact force $\vec{f}$ forming an elastic line in $\mathbb{R}^3$. In doing so, the strict convexity of the object ensures, that there is always only one contact point $P_1(\xi, \eta, \theta)$ between the rod and the object [21], see Figure 2.

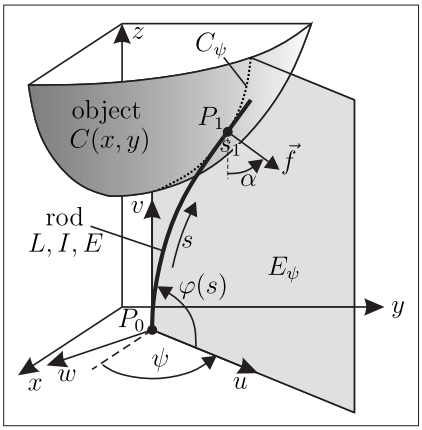


Figure 2. Mechanical model for object shape scanning and reconstruction rod in contact with an object's surface adapted from [1].

B. 3D frictional contact model

It was concluded in [20], that (based on the assumptions in Section II-A) the elastic line of the rod shrinks to one in some deformation plane E_ψ with (yet) unknown orientation ψ, see Figure 2. However, this plane bending working hypothesis was made assuming an ideal contact, i.e., in the absence of friction. Here, we now analyze the problem using a more realistic approach, roughly approximating the micro-mechanics of the contact using Coulomb's law of friction [22]. Therefore, we assume the constraining force $\vec{f}$, transmitted through the object contact as composed of two components, see Figure 3:

$$\vec{f} = \vec{f}_n + \vec{f}_t \tag{4}$$

The component $\vec{f}_n$ is aligned with the outer normal vector $\vec{n}_1$ of the surface C at P_1.

$$\vec{f}_n = |\vec{f}_n| \cdot \vec{n}_1, \tag{5}$$

In contrast, lying in the tangent plane at P_1 of C, the component $\vec{f}_t$ still has an unknown orientation. Both components are coupled by Coulomb's law of friction:

$$\vec{f}_t = -\mu |\vec{f}_n| \frac{\vec{v}}{|\vec{v}|}, \tag{6}$$

where $\mu = \frac{|\vec{f}_t|}{|\vec{f}_n|} = \tan(\zeta)$ is the coefficient of friction with friction angle ζ and $\vec{v}$ is the sliding velocity. Being restricted to a quasi-static model, we cannot make any statements about the sliding velocity $\vec{v}$. However, according to (6), the tangential force solely depends on the direction of sliding but not on the actual speed of sliding. Therefore, the basic idea within the present paper is to interpret a sequence of contact points $(P_{1,i})_{i\in\mathbb{N}}$ as the trajectory of P_1 on C, see Figure 3. Then, the direction opposing the sliding velocity at some point $P_{1,k}$ is approximated using the previous contact point $P_{1,k-1}$ in the following way:

$$\vec{f}_t^* = |\vec{f}_t^*| \cdot \vec{t}^*, \quad \text{with} \quad \vec{t}^* = \frac{P_{1,k-1} - P_{1,k}}{|P_{1,k-1} - P_{1,k}|} \tag{7}$$

The distance $P_{1,k-1} - P_{1,k}$ of two successive contact points decreases with decreasing step size of the system input and, thus, using tiny incremental steps, (7) approaches a tangent vector of C at $P_{1,k}$, pointing from $P_{1,k}$ to $P_{1,k-1}$.

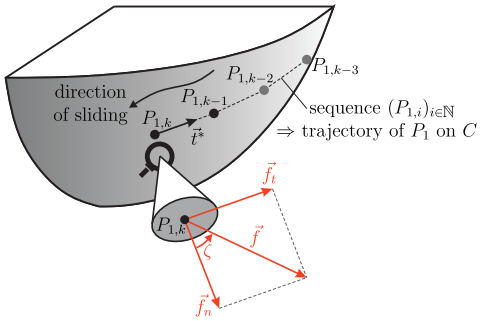


Figure 3. Sequence of contact points between rod and object, representing the trajectory of P_1 on C with the constraining force at $P_{1,k}$ (see enlarged section).

C. Preliminary geometrical considerations

Figure 4 shows the constraining force $\vec{f}$ acting at P_1 for both friction partners, i.e., for the object's surface in Figure 4(a) and for the bending rod in Figure 4(b). As can be seen from Figure 4(b), all forces are considered to act at the center of the rod's cross-section for sake of simplicity and, thus, their lines of action intersect the axial line of the rod. Otherwise, the contact forces might create moments about the axial line, possibly causing twist deformations, which are neglected in our theory. This simplification is justified by the extremely small cross-sectional dimensions of the rod under consideration (and the natural paragon vibrissa).

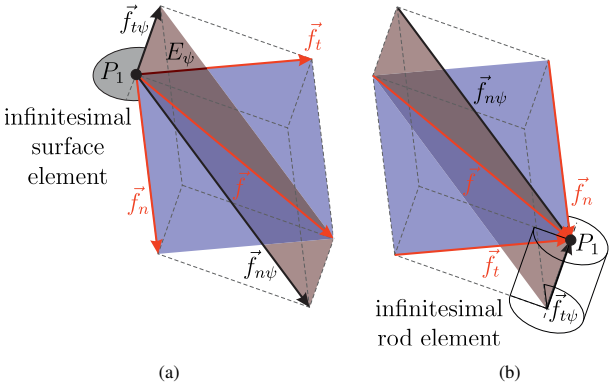


Figure 4. Components of the constraining force $\vec{f}$ at P_1: (a) infinitesimal surface element; (b) infinitesimal rod element.

Dropping twist deformations of the rod, Figure 2 and Figure 4 clarify the following geometrical relationships: The constraining force $\vec{f}$ defines the deformation plane E_ψ in such a way that the geometrical condition $\{P_0, P_1, \vec{e}_z, \vec{f}\} \in E_\psi$ is fulfilled [20]. Introducing the local coordinate system (u,v,w) (local frame), see Figure 2, we come up with the following transformation rules [20]:

$$\begin{pmatrix}\vec{e}_u\\ \vec{e}_v\\ \vec{e}_w\end{pmatrix} = \mathbf{T}(\psi)\cdot\begin{pmatrix}\vec{e}_x\\ \vec{e}_y\\ \vec{e}_z\end{pmatrix}, \qquad \begin{pmatrix}\vec{e}_x\\ \vec{e}_y\\ \vec{e}_z\end{pmatrix} = \mathbf{T}^{-1}(\psi)\cdot\begin{pmatrix}\vec{e}_u\\ \vec{e}_v\\ \vec{e}_w\end{pmatrix} \quad (8)$$

$$\text{with} \qquad \mathbf{T}(\psi) = \begin{pmatrix}\cos(\psi) & \sin(\psi) & 0\\ 0 & 0 & 1\\ \sin(\psi) & -\cos(\psi) & 0\end{pmatrix} \quad (9)$$

In the following, all considerations are restricted to the plane E_ψ (u-v-plane). This plane intersects the surface C in some intersection curve $C_\psi = E_\psi \cap C$, see Figure 2. The tangent and normal directions

$$\vec{t}_\psi = \vec{e}_w \times \vec{n}_1 \qquad \vec{n}_\psi = \vec{t}_{1\psi} \times \vec{e}_w \quad (10)$$

of C_ψ at P_1 are used to split the constraining force $\vec{f}$ into two orthogonal components $\vec{f}_{n\psi}, \vec{f}_{t\psi} \in E_\psi$, see Figure 4. The relation $\vec{f} = \vec{f}_n + \vec{f}_t = \vec{f}_{n\psi} + \vec{f}_{t\psi}$ is visualized by means of a rectangular cuboid in Figure 4. There, the purple diagonal plane is spanned by $\vec{f}_n$ and $\vec{f}_t$, the brown diagonal plane, namely E_ψ, is spanned by $\vec{f}_{n\psi}$ and $\vec{f}_{t\psi}$ and the line of intersection of these planes denotes the resulting force $\vec{f}$.

D. The plane elastic line

The analysis within this section is largely analog with [20] and therefore roughly outlined here. Within the deformation plane E_ψ (u-v-plane), the elastic line of the rod is described parameterizing the axis of the rod by means of its slope angle φ in dependence on its natural coordinate arc length s. Moreover, using Euler's constitutive law $\kappa(s) = m(s)$ in dimensionless representation (mind (1)) we finally end up in a system of Ordinary Differential Equations (ODEs) [20]:

$$\boxed{\begin{array}{ll} u'(s) = \cos(\varphi(s)) & \varphi'(s) = \kappa(s)\\ v'(s) = \sin(\varphi(s)) & \kappa'(s) = f\cos(\varphi(s) - \alpha)\end{array}} \quad (11)$$

In (11), f is the magnitude of $\vec{f}$ and α is defined as the signed angle between $\vec{f}$ and $-\vec{e}_v$, see Figure 2:

$$\alpha = \operatorname{atan2}(\vec{e}_u \cdot \vec{f}, -\vec{e}_v \cdot \vec{f}) \quad (12)$$

The ODE system (11) describes the plane elastic line of the rod. Depending on the corresponding problem, Boundary Conditions (BCs) or Initial Conditions (ICs) are to be adjoint in Section II-E and II-F, respectively.

E. Step 1: Generating the support reactions theoretically

The preliminary process of theoretically generating the support reactions at each particular clamping position P_0 reflects/replaces the actual experiment of sweeping the rod along the object while measuring the support reactions (observables). For Step 1, we assume

- the object surface C,
- the clamping position P_0,
- and the friction coefficient μ

to be known (preset) in advance. In this way, Step 1 is used as a valuable tool allowing for the generation of the support

reactions during scanning sweeps with different preset friction coefficients.

Following [1] and [20], we distinguish two different contact phases, see Figure 5:

- For tip contacts, the position of contact along the rod $s_1 = 1$ is known, but the angle $\varphi_1 = \varphi(1) > \tilde{\alpha}$ is unknown.
- In contrast, for tangential contacts, the position of contact along the rod s_1 is unknown, but instead, we have the angular relationship $\varphi = \tilde{\alpha}$.

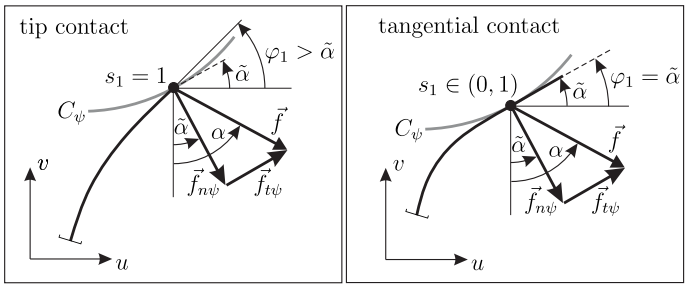


Figure 5. Comparison between tip and tangential contact with angular relationships within the deformation plane E_ψ.

Inspired by Figure 5, we define the angle $\tilde{\alpha}$ as the tangent slope angle of C_ψ at P_1:

$$\tilde{\alpha} = \text{atan2}(\vec{e}_u \cdot \vec{n}_\psi, -\vec{e}_v \cdot \vec{n}_\psi) \tag{13}$$

Consequently, we find the following BCs for tip (14) and tangential contacts (15), respectively:

$$\begin{array}{ll} u(0) = 0 & u(1) = u_1 \\ v(0) = 0 & v(1) = v_1 \\ \varphi(0) = \dfrac{\pi}{2} & \\ & \kappa(1) = 0 \end{array} \tag{14}$$

$$\begin{array}{ll} u(0) = 0 & u(s_1) = u_1 \\ v(0) = 0 & v(s_1) = v_1 \\ \varphi(0) = \dfrac{\pi}{2} & \varphi(s_1) = \tilde{\alpha} \\ & \kappa(s_1) = 0 \end{array} \tag{15}$$

In (14) and (15), u_1 and v_1 are the coordinates of P_1 w.r.t. the local frame. Finding equilibrium states of the rod in contact with the object is achieved using a Matlab algorithm. Basically, the algorithm repeatedly solves the Boundary-Value Problems (BVPs) (11)&(14) and (11)&(15) for each clamping position using shooting methods to determine the unknown parameters s_1, $|\vec{f}_n|$, u_1 and ψ.

Remark 2. *More specifically, assuming reasonable starting values for these unknown parameters, the used algorithm proceeds using* (2)–(15) *in the following way: For each clamping position* P_0*...*

(a) *... determine*
- *the orientation of the local frame using* (8)*,*
- *the contact point* P_1 *evaluating* (2) *and* (8)*,*
- *the coordinate* v_1 *using* (8)*,*
- *the normal force* $\vec{f}_n$ *using* (3) *and* (5)*,*
- *the tangential force* $\vec{f}_t$ *using* (6) *and* (7) *,*
- *the resulting force* $\vec{f}$ *based on* (4)*,*
- *the components* $\vec{f}_{n\psi}$ *and* $\vec{f}_{t\psi}$ *of* $\vec{f}$ *using* (10)*,*
- *the angles* α *and* $\tilde{\alpha}$ *using* (12) *and* (13)*.*

(b) *... try to find a solution for the BVPs* (11)&(14) *or* (11)&(15)*, respectively.*

(c) *... repeat A and B varying the parameters* s_1*,* $|\vec{f}_n|$*,* u_1 *and* ψ *until an equilibrium state has been found.*

Once all relevant parameters are known, the support reactions

$$f_{0u} = -f\sin(\alpha), \qquad f_{0v} = -f\cos(\alpha) \tag{16}$$

$$m_{0w} = f[u_1\cos(\alpha) + v_1\sin(\alpha)] \tag{17}$$

are calculated and expressed w.r.t. the global frame using (8):

$$\vec{m}_0 = m_{0x}\vec{e}_x + m_{0y}\vec{e}_y + m_{0z}\vec{e}_z \tag{18}$$

$$\vec{f}_0 = f_{0x}\vec{e}_x + f_{0y}\vec{e}_y + f_{0z}\vec{e}_z \tag{19}$$

Finally, this procedure results in a sequence of support reactions for an entire scanning sweep, starting from the very first contact between the rod and the object and terminating with the detachment of the rod.

F. Step 2: Reconstructing contact points

The process of reconstructing a sequence of contact points in space based on the support reactions of Section II-E reflects the actual sensor application. Of course, the shape of the object, as well as the friction coefficient, are unknown in this context. Instead, only the following quantities are assumed to be known in advance:

- the support reactions (18)&(19) generated in Step 1 (or measured using a real experiment)
- the clamping position P_0

However, compared to Step 1, the mechanical problem of Step 2 is relatively simple, as the orientation of the deformation plane E_ψ is directly evident evaluating

$$\psi = -\text{atan2}(m_{0x}, m_{0y}) \tag{20}$$

Then, expressing (18)&(19) w.r.t. the local frame using (8), f and α follow from (17)

$$f = \sqrt{f_{0u}^2 + f_{0v}^2}, \quad \alpha = -\text{atan2}(f_{0u}, f_{0v}) \tag{21}$$

and the curvature at P_0 writes

$$\kappa(0) = -m_{0w} \tag{22}$$

Using (22), we have the following ICs:

$$\begin{array}{l} u(0) = 0 \\ v(0) = 0 \\ \varphi(0) = \dfrac{\pi}{2} \\ \kappa(0) = -m_{0w} \end{array} \tag{23}$$

In contrast to Step 1, including a BVP, Step 2 is characterized by the IVP (11)&(23), which can be solved without shooting methods. Instead, it is integrated numerically using an event function, which cancels further computation if the termination condition $\kappa(s_1) = 0$ is fulfilled. This condition results from the single contact point scenario due to the modeling assumptions, in contrast to [21]. Then, the last step of the numerical integration includes the solutions for s_1, u_1, v_1 and φ_1, see Figure 5. Finally, the contact position in space is expressed with respect to the global frame using (8).

III. RESULTS & DISCUSSION

The following results are based on simulated scanning sweeps of the probe along a 3D object, which are illustrated in Section III-A. We analyze the theoretically generated support reactions at the base of the probe in Section III-B. In Section III-C, these support reactions are used to reconstruct sequences of contact points.

A. Simulated scanning sweeps

Within the present paper we use (2) with the parameters $a = 0.5$, $b = 1$ and $h = 0.4$ as an exemplary object surface. The geometry parameters are chosen based on preliminary studies. On the one hand, $a \neq b$ ensures, that the results are not restricted to surfaces of revolution. On the other hand, the chosen object distance h ensures the occurrence of both tip and tangential contacts. The scanning trail is assumed to be parallel to the x-axis. The clamping position $P_0(x_0, y_0, 0)$ of the rod is displaced along this trail decreasing the system input x_0 with constant y_0. In doing so, we simulated scanning sweeps for $y_0 = -0.4 : 0.2 : +0.4$ and $\mu = 0 : 0.1 : 0.4$ resulting in a total number of 25 scanning sweeps, of which four exemplary sweeps are shown in Figure 6. There, the scanning sweeps are represented by sequences of elastic lines (equilibrium states), where tip contacts are colored in blue and tangential ones in red for $s \in (0, s_1)$ and black for $s \in (s_1, 1)$. For reasons of clarity, only one in ten calculated deformation states is shown in Figure 6. Essentially, a scanning sweep may exhibit two fundamentally different characteristics, see Figure 6:

- For the special case $y_0 = 0$, the scanning trail entirely lies within a symmetry plane of the object (x-z-plane). Regardless of the friction coefficient, this arrangement results in a special case of an entirely plane scanning sweep, see Figure 6(a) ($\mu = 0$) and Figure 6(b) ($\mu = 0.4$). Such a plane scanning sweep always terminates with a "snap-off" (dynamical detachment) of the deformed rod from the object [23]. It is to be noted, that increasing friction coefficients result in slightly delayed "snap-offs" and, thus, increase the scanning range, compare Figures 6(a) and 6(b).
- For the more general case $y_0 \neq 0$, a completely different scanning behavior is observed: As seen in Figure 6(c) ($y_0 = 0.2$, $\mu = 0$) and Figure 6(d) ($y_0 = 0.2$, $\mu = 0.4$) the rod bends around the object and finally smoothly detaches from the object stress-free, i.e., without any snap-off. In general, for the case $\mu = 0$, the symmetry of the object w.r.t. y-z-plane results in a symmetric appearance of the elastic lines, see Figure 6(c). This symmetry is not maintained with increasing mu, see Figure 6(d). Instead, the orientation of the bending plane E_ψ in Figure 6(d) seems to lack behind the one in Figure 6(c) as a consequence of the frictional force.

Remark 3. *The reason for showing Figure 6 at the very beginning of this section is to create a figurative idea about the scanning process. However, it is important to highlight that Figure 6 actually results at the very end of the simulation process after performing Steps 1 and 2.*

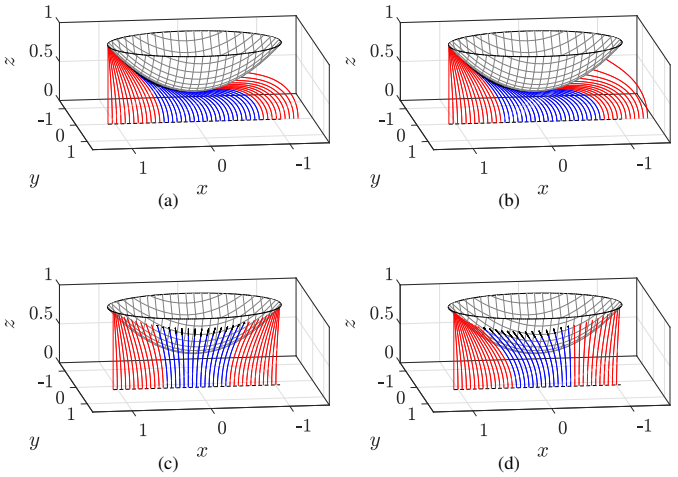


Figure 6. Exemplary scanning sweeps in negative x-direction represented by sequences of elastic lines: (a) $y_0 = 0$, $\mu = 0$; (b) $y_0 = 0$, $\mu = 0.4$; (c) $y_0 = 0.2$, $\mu = 0$; (d) $y_0 = 0.2$, $\mu = 0.4$.

B. Step 1: Generating the support reactions

Figures 7 and 8 show the components of the support reactions (18) and (19) resulting from ten exemplary scanning sweeps plotted against the system input x_0 in dependence on different friction coefficients μ. In both figures, the phase transitions from tip to tangential contact and vise versa are marked with an 'o'. Note that Figures 7 and 8 are to be read from right to left due to the scanning direction (negative x-direction). Figure 7 represents the plane special case ($y_0 = 0$) identified in Section III-A and, thus, the dark blue ($\mu = 0$) and yellow curves result from the scanning sweeps in Figures 6(a) and 6(b). Regardless of the friction coefficient, the components f_{0y}, m_{0x} and m_{0z} are zero during the entire scanning sweep and therefore obscure each other for different values of μ. This fact confirms the observation that for $y_0 = 0$ all scanning sweeps entirely take place in the x-z-plane. It is striking that the remaining components of the support reactions are affected by the friction coefficient, e.g., the component f_{0z} consistently increases with increasing μ. Moreover, comparing the values x_0 at the end of each scanning sweep (left side of each curve), confirms that an increasing coefficient results in a longer overall contact phase. It can be seen that the phase transitions are characterized by discontinuities (kinks). However, the friction coefficient seems to have little impact on these discontinuities and the location of the phase transitions in general.

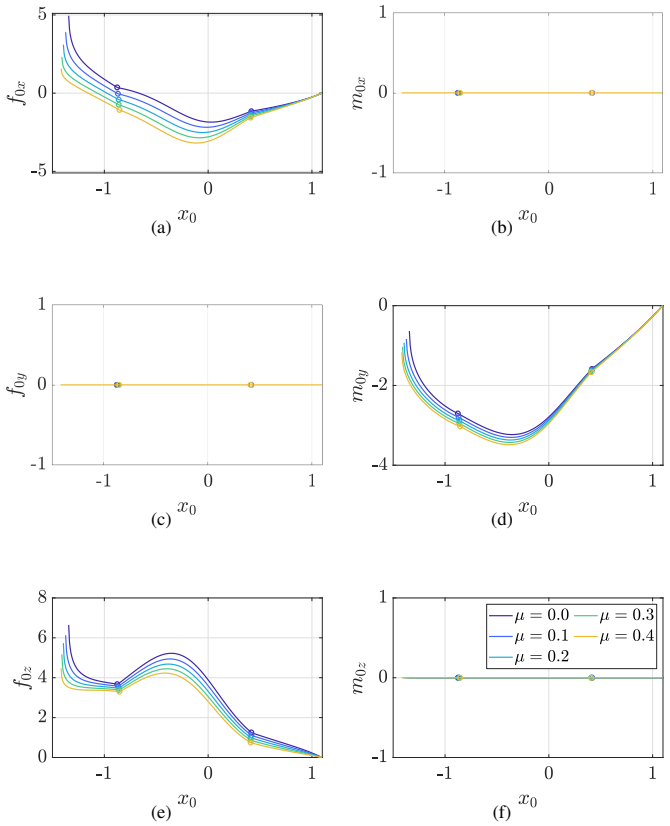


Figure 7. Components of the support reactions (observables) $\vec{f}_0$ and $\vec{m}_0$ resulting from simulated scanning sweeps with $y_0 = 0$ plotted against the system input x_0 in dependence on different preset friction coefficients μ.

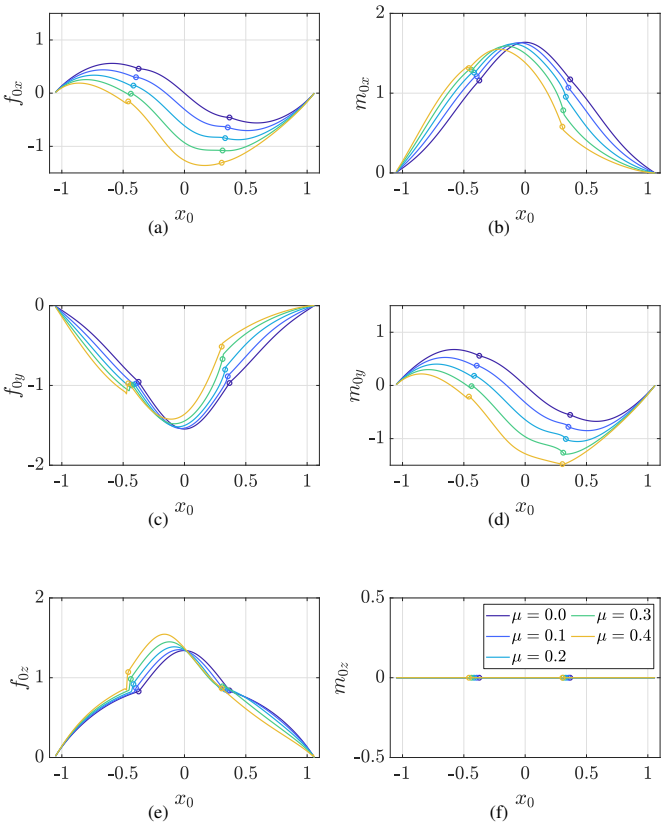


Figure 8. Components of the support reactions (observables) $\vec{f}_0$ and $\vec{m}_0$ resulting from simulated scanning sweeps with $y_0 = 0.2$ plotted against the system input x_0 in dependence on different preset friction coefficients μ.

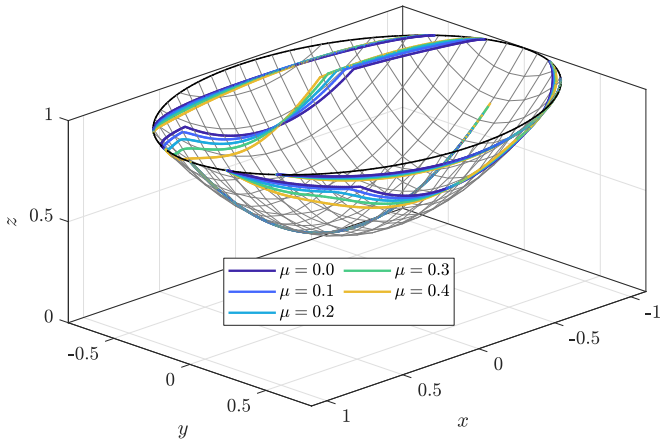
Exemplary considering the case $y_0 = 0.2$, Figure 8 represents the more general case $y_0 \neq 0$ of a scanning sweep, see Section III-A. Again, the dark blue ($\mu = 0$) and yellow curves correspond to the scanning sweeps in Figures 6(a) and 6(b), respectively. In contrast to Figure 7, only the component m_{0z} is zero during the entire scanning sweep due to the plane bending assumption (ignoring twist deformation), see Section II-C. The remaining components are strongly affected by the friction coefficient. The dark blue curves ($\mu = 0$) are vertical or point symmetric as a consequence of the object's symmetry. For increasing friction coefficients, the maximum values of the support reactions are increasingly shifted in the negative direction of x_0. However, for $y_0 \neq 0.2$, the friction coefficient has no impact on the position of the rod detachment. Moreover, all support reactions fall to zero at the end of the scanning sweep. This confirms the stress-free detachment of the rod from the object. Finally, compared to Figure 7, the phase transitions seem to be stronger affected by the friction coefficient. For instance, an increasing friction coefficient results in delayed phase transitions and causes discontinuities (small jumps). These jumps might result from a sudden change of the sliding trajectory (and, thus, the direction of the frictional force), when the contact point s_1, which is one in case of tip contact, suddenly begins to move along the rod axis in case of tangential contact.

Summarizing, the support reactions are strongly affected by the friction coefficient. Unfortunately, Figures 7 and 8 do not yet allow us to draw further conclusions about the object's shape. Therefore, we consider Step 2 in the next section.

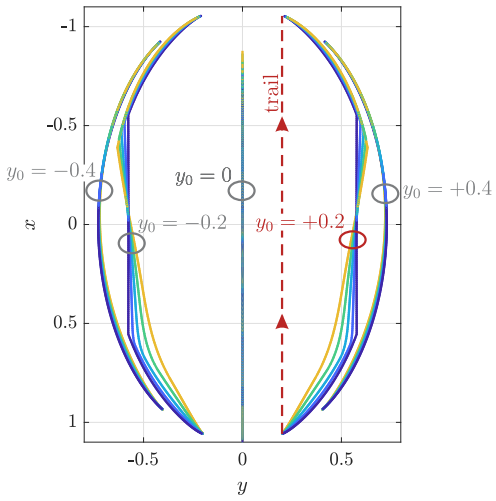
C. Step 2: Reconstructing contact points

Figure 9 shows 25 sequences of Reconstructed Contact Points (RCPs) based on the generated support reactions of 25 simulated scanning sweeps. The isometric view in Figure 9(a) shows the original object surface C alongside with the RCPs. Moreover, a top view with hidden object surface is given in Figure 9(b). In contrast to Figure 6, where only one in ten calculated deformation states is displayed, the point sequences in Figure 9 include the RCP of each single simulation step. Consequently, each sequence of RCPs rather appears as a line than a multitude of points. The high density of RCPs ensures a good approximation of the tangent direction in (7). Considering Figure 9 it becomes clear that the sequences of RCPs depend on both the scanning trail position y_0 and the friction coefficient μ. Mind: for each scanning trail $y_0 = -0.4 : 0.2 : +0.4$, five scanning sweeps with different friction coefficients $\mu = 0 : 0.1 : 0.4$ were simulated. This results in a bundle of five sequences of RCPs for each y_0, marked in Figure 9(b). For instance, considering Figure 9(b), scanning the object on the red scanning trail ($y_0 = +0.2$) results in five sequences of RCPs (circled in red), depending on the

friction coefficient, see color legend in Figure 9(a). For reasons of clarity, only one scanning trail is shown in Figure 9(b), while the others are simply pointed out giving the corresponding values of y_0 (gray).



(a)



(b)

Figure 9. Reconstructed sequences of contact points based on 25 simulated scanning sweeps: (a) original object surface superimposed with reconstructed contact points in dependence on the friction coefficient μ; (b) top view including information about the scanning trails.

Considering Figure 9, the following results must be highlighted:

- For $y = 0$ it appears that the sequences of RCPs coincide to a large extend regardless of the friction coefficient. However, this does not mean that the RCPs are not affected by friction. Instead, for increasing friction coefficients, the RCPs shift along the object's surface within the x-z-plane. For instance, it can be seen that for the largest friction coefficient $\mu = 0.4$, the sequence of RCPs exceeds all others. This reflects the fact that (for the plane special case) increasing friction coefficients increase the scanning range due to a delayed "snap-off".
- For $y \neq 0$, the sequences of RCPs do no longer coincide for different friction coefficients. While outermost sequences ($|y_0| = 0.4$) differ slightly for different friction coefficients, the inner ones ($|y_0| = 0.2$) seem to be stronger affected by friction. In this way, the sequences of RCPs based on simulations with $\mu = 0$, are completely symmetric w.r.t. y-z-plane. In contrast, for $\mu > 0$ the frictional force causes asymmetries. For the preset object distance $h = 0.4$, scanning sweeps with $|y_0| = 0.4$ do not include tangential contacts in contrast to those with $|y_0| = 0.2$. This statement remains true regardless of the friction coefficient. For $|y_0| = 0.2$, the transitions between tip and tangential contacts and vise versa are clearly identifiable by the discontinuities (kinks) of the sequences of RCPs. For instance, considering the dark blue sequence ($\mu = 0$, $y_0 = 0.2$), the first discontinuity (at $x \approx 0.5$) reflects the transition from tip to tangential contact and the second one (at $x \approx 0.5$) the inverse transition, see Figure 6(a). Increasing friction coefficients weaken the extent of the first discontinuity and reinforce the extent of the second one, see Figure 9(b).

Summarizing, increasing friction coefficients affect the sequences of RCPs. This may falsely give the impression, that increasing friction results in increased distortion of the reconstruction result. For instance, in Figure 9(b), the reconstruction result appears to be distorted in a way that the width of the object in y-direction is underrated at the beginning and overrated at the end of the scanning sweep. This seems to be the case especially with the yellow sequences ($\mu = 0.4$). However, it is important to understand that such distortion does not take place at all. Instead, the isometric view in Figure 9(a) suggests that all reconstructed points lie perfectly on the original object surface. To highlight and verify this suspicion, the reconstruction error of each RCP is calculated and shown by the color-bar in Figure 10. The reconstruction error is defined as the smallest (perpendicular) distance between a RCP P_{rek} and the original object surface C: Let $P(a, b, c) \in C$ be the point of C closest to P_{rek}. Then $\vec{e} = \overrightarrow{PP_{rek}}$ is collinear with $\vec{n}(a, b)$ and the error of P_{rek} is $|\vec{e}|$. Obviously, the maximum reconstruction error lies within the numerical boundaries regardless of the friction coefficient. This fact confirms an empirical observation and hypothesis originally made in [20] and leads to the final result of the present paper:

- For each clamping position P_0 during scanning, the frictional force affects the deformation of the rod and, thus, the contact point P_1 on C.
- Consequently, the support reactions at the base of the rod are affected by friction
- Evaluating the support reactions in order to reconstruct the contact position in space, different friction coefficients would result in different locations of the RCP. For instance, a point P_1^*, which is reconstructed in the presence of friction ($\mu > 0$) is different from the point P_1, which would have been reconstructed in the absence of friction

- However, both points P_1^* and P_1 would lie on the original object surface C.
- Thus, friction leads (of course) to a shift of the RCPs on the scanned 3D object surface itself, but does not affect the reconstruction error. This is what we refer to as friction invariant reconstruction.

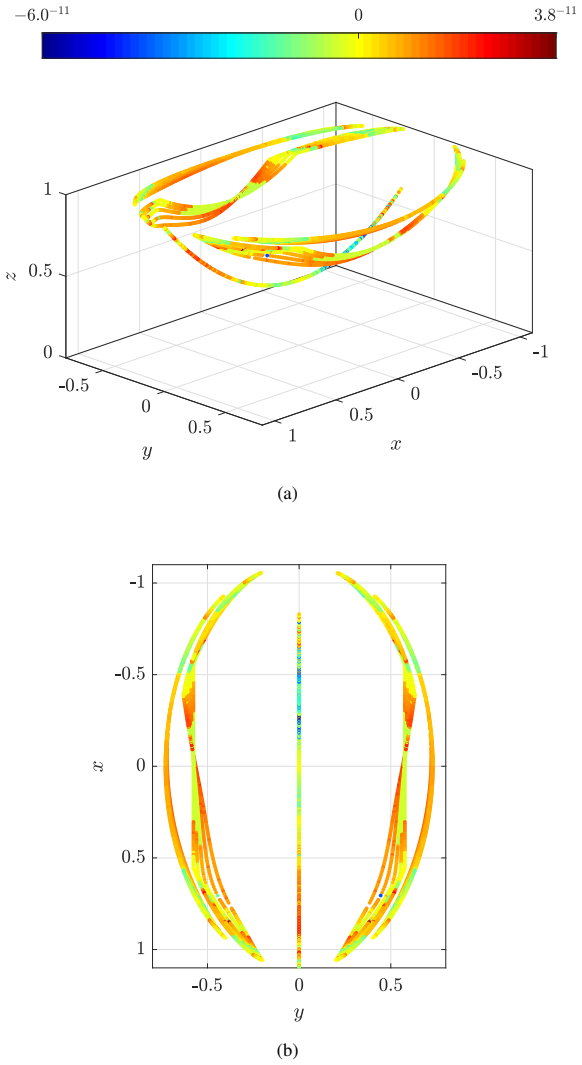


Figure 10. Reconstructed sequences of contact points based on 25 simulated scanning sweeps highlighting the reconstruction error using a color-bar: (a) isometric view; (b) top view.

IV. Conclusion

Within this paper, we presented a vibrissa-inspired tactile sensor concept for 3D object surface scanning and reconstruction. For that purpose, a one-sided clamped rod was swept along the object's surface by incremental displacements of its clamping position, relatively to the object. In contrast to other publications, we firstly approximated the micro-mechanics of the contact assuming Coulomb friction, instead of presuming an ideal contact (no frictional effects). Being restricted to a quasi-static model, the direction of the frictional force was approximated inferring from one contact point to the next one. Based on the novel modeling approach, two consecutive processes were analyzed separately: Firstly, scanning sweeps along a known object surface were simulated with a preset friction coefficient in Step 1. In this way, the support reactions (observables) at the base (clamping) of the rod were generated theoretically. Secondly, we demonstrated how to use the support reactions from Step 1 in order to reconstruct sequences of contact points, finally approximating the surface of the scanned object. Both steps were implemented in a Matlab algorithm and simulated to demonstrate the general applicability. The simulation results showed that friction (obviously) affects the support reactions during object scanning. However, using these support reactions to reconstruct contact points, it surprisingly turned out that the reconstruction result is friction invariant, i.e., friction does not affect the reconstruction error. This is a novel and not self-evident finding, revealing a major advantage of the vibrissa-inspired sensor concept. Thus, extending the mechanical model including Coulomb friction yields new insights compared to previous works. Instead of representing a self-contained investigation, the present paper should be seen as a preliminary concept study, indicating that the presented measuring principle is highly suited to complement optical sensors in the environmental exploration of robots. In doing so, we aim to implement the presented concept into an intelligent tactile sensor in the future.

Finally, it remains to mention that the results and hypotheses of the present paper are based on a quasi-static model. This means the scanning sweep is realized by incremental displacements of the clamping, resulting in a sequence of consecutive equilibrium states. Against this backdrop, only static friction but no dynamical effects, e.g., stick-slip effects can be discussed. However, instead of repeatedly stopping the clamping position in order to wait for the stationary state, in practice, the scanning sweep rather has to be thought of as a continuous movement of the clamping and, thus, a continuous sweep of the rod over the object surface. Therefore, dynamical effects like stick-slip effects as observed in [20] are likely to occur in reality. Therefore, it remains to verify the theoretical results on practical examples by using an experimental setup, which has already been attacked in the first steps.

References

[1] L. Merker, M. Scharff, K. Zimmermann, and C. Behn, "Surface Sensing of 3D Objects Using Vibrissa-like Intelligent Tactile Sensors," INTELLI 2020, The Ninth International Conference on Intelligent Systems and Applications, Porto, Portugal, October 18–22, 2020, pp. 18–23. ISBN: 978-1-61208-798-6.

[2] J. Minguez, F. Lamiraux, and J. Laumond, "Motion planning and Obstacle Avoidance," In: Springer Handbook of Robotics. Ed. by B. Siciliano; O. Khatib, Berlin, Heidelberg: Springer, pp. 1177–1202, 2016.

[3] M. Brecht, B. Preilowski, and M. M. Merzenich, "Functional architecture of the mystacial vibrissae," Behav. Brain Res., vol. 84, pp. 81–97, 1997.

[4] E. Guić-Robles, C. Valdivieso, and G. Guajardo, "Rats can learn a roughness discrimination using only their vibrissal system," Behav. Brain Res., vol. 31, pp. 285–289, 1989.

[5] G. E. Carvell and D. J. Simons, "Biometric Analyses of Vibrissal Tactile Discrimination in the Rat," J. Neurosci., vol. 10, pp. 2638–2648, 1990.

[6] T. J. Prescott, B. Mitchinson, and R. A. Grant, "Vibrissal behavior and function," Scholarpedia, vol. 6, p. 6642, 2011.

[7] H. M. Belli, A. E. T. Yang, C. S. Bresee, and M. J. Z. Hartmann, "Variations in vibrissal geometry across the rat mystacial pad: Base diameter, medulla, and taper," J. Neurophysiol., vol. 117, pp. 1807–1820, 2016.

[8] D. Voges et al., "Structural Characterization of the Whisker System of the Rat," IEEE Sens. J., vol. 12, pp. 332–229, 2012.

[9] D. Campagner, M. H. Evans, M. S. E. Loft, and R. S. Petersen, "What the whiskers tell the brain," Neurosci. vol. 368, pp. 95–108, 2018.

[10] S. Ebara, T. Furuta, and K. Kumamoto, "Vibrissal mechanoreceptors," Scholarpedia, vol. 12, p. 32372, 2017.

[11] J. H. Solomon and M. J. Z. Hartmann, "Radial distance determination in the rat vibrissal system and the effects of Weber's law," Phil. Trans. R. Soc. B., vol. 366, pp. 3049–3057, 2011.

[12] J. H. Solomon and M. J. Z. Hartmann, "Extracting object contours with the sweep of a robotic whisker using torque information," Int. J. Robot Res. vol. 29, pp. 1233–1245, 2010.

[13] J. H. Solomon and M. J. Z. Hartmann, "Artificial whiskers suitable for array implementation: accounting for lateral slip and surface friction," IEEE Trans. Robot., vol. 24, pp. 1157–1167, 2008.

[14] D. Kim and R. Möller, "Biomimetic whiskers for shape recognition," Proceedings of the 1995 IEEE International Conference on Robotics and Automation, Nagoya, Japan, 21-27 May 1995, pp. 1113–1119.

[15] M. Kaneko, N. Kanayama, and T. Tsuji, "Active antenna for contact sensing," IEEE Trans. Robot. Autom., vol. 14, pp. 278–291, 1998.

[16] A. E. Schultz, J. H. Solomon, M. A. Peshkin, and M. J. Z. Hartmann, "Multifunctional Whisker Arrays for Distance Detection, Terrain Mapping, and Object Feature Extraction," IEEE International Conference on Robotics and Automation, Barcelona, Spain, 18-22 April 2005, pp. 2588–2593.

[17] C. Will, C. Behn, and J. Steigenberger, "Object contour scanning using elastically supported technical vibrissae," ZAMM J. Appl. Math. Mec., vol. 98, pp. 289–305, 2018.

[18] T. N. Clements and C. D. Rahn, "Three-Dimensional Contact Imaging With an Actuated Whisker," IEEE Trans. Rob., vol. 22, pp. 844–848, 2006.

[19] L. A. Huet, J. W. Rudnicki, and M. J. Z. Hartmann, "Tactile sensing with whiskers of various shapes: determining the three-dimensional location of object contact Based on mechanical signals at the whisker base." Soft Robot., vol. 4, pp. 88–103, 2017.

[20] L. Merker, J. Steigenberger, R. Marangoni, and C. Behn, "A vibrissa-inspired highly flexible tactile sensor: scanning 3D object surfaces providing tactile images," Sensors, vol. 21, 2021. DOI: 10.3390/s21051572.

[21] L. Merker, S. J. Fischer Calderon, M. Scharff, J. H. Alencastre Miranda, and C. Behn, "Effects of Multi-Point Contacts during Object Contour Scanning Using a Biologically-Inspired Tactile Sensor," Sensors, vol. 20, 2020. DOI: 10.3390/s20072077.

[22] V. L. Popov, "Contact mechanics and friction," Springer, 2010.

[23] L. Merker, M. Scharff, and C. Behn, "Approach to the dynamical scanning of object contours using tactile sensors," IEEE International Conference on Mechatronics (ICM), pp. 364–369, 2019. DOI: 10.1109/ICMECH.2019.8722882.

Hybrid Knowledge-based and Data-driven Text Similarity Estimation based on Fuzzy Sets, Word Embeddings, and the OdeNet Ontology

Tim vor der Brück

School of Computer Science and
Information Technology
Lucerne University of Applied Sciences and Arts
Rotkreuz, Switzerland
e-mail: tim.vorderbrueck@hslu.ch

Michael Kaufmann

School of Computer Science and
Information Technology
Lucerne University of Applied Sciences and Arts
Rotkreuz, Switzerland
e-mail: m.kaufmann@hslu.ch

***Abstract*—Estimating the semantic similarity between texts is important for a wide range of application scenarios in natural language processing. With the increasing availability of large text corpora, data-driven approaches such as Word2Vec have become quite successful. In contrast, semantic methods, which employ manually designed knowledge bases such as ontologies, have lost some of their former popularity. However, manually designed expert knowledge can still be a valuable resource, since it can be leveraged to boost the performance of data-driven approaches. In this paper, we introduce a novel hybrid similarity estimate based on fuzzy sets that exploits both word embeddings and a lexical ontology. As ontology, we use OdeNet, a freely available resource developed by the Darmstadt University of Applied Sciences. Our application scenario is targeted marketing, in which we aim to match people to the best fitting marketing target group based on short German text snippets. The evaluation showed that the use of an ontology did indeed improve the overall result in comparison with a baseline data-driven estimate.**

Keywords–OdeNet; fuzzy sets; targeted marketing; histogram equalization.

I. Introduction

Note that this paper is an extended version of [1]. In comparison with the original conference paper, we updated some of the linguistic resources (stop word list, lemmatization, and OdeNet ontology), conducted additional experiments and gave a more detailed evaluation. In particular, we evaluated three additional coefficients for our ontology-based similarity estimate, namely the Sørensen-Dice coefficient (henceforth, the *Dice coefficient*), the overlap coefficient, and pointwise mutual information. Furthermore, we investigated the distribution of the gold standard annotations and determined milieu-wise precision, recall, and F1-scores for the most accurate similarity estimate.

The approach presented here was developed in cooperation with a marketing company with the goal of facilitating market segmentation, which is one of the key tasks of a marketer. Usually, market segmentation is accomplished by clustering demographic variables, geographic variables, psychographic variables, and behaviors [2]. In this paper, we will describe an alternative approach based on unsupervised natural language processing. In particular, our business partner operates a commercial youth platform for the Swiss market, where registered members receive access to third-party offers such as discounts and special events (e.g., concerts or castings). Several hundred online contests per year, which are sponsored by other firms, are launched over this platform, and an increasing number of them require members to write short free-text snippets (e.g., to elaborate on a perfect holiday at a destination of their choice in case of a contest sponsored by a travel agency). Based on the results of a broad survey, the platform provider's marketers assume six target groups (called *milieus*) exist among the platform members. For each milieu (with the exception of the default milieu *Special Groups*) a keyword list was manually created to describe its main characteristics. To trigger marketing campaigns, an algorithm has been developed that automatically assigns each contest answer to the most likely target group: we propose the youth milieu as the best match for a contest answer, for which the estimated semantic similarity between the associated keyword list and user text snippet is maximal. For the estimation of text relatedness, we devised a novel semantic similarity estimate based on a combination of word embeddings and OdeNet (*Offenes deutsches Wordnet - open German wordnet*), where the latter is a freely available lexical ontology recently developed by the Darmstadt University of Applied Sciences.

The remainder of this paper is organized as follows: In the next section, we survey some of the related work in the area of semantic text similarity estimation. Our proposed methodology is described in Section III. Section IV introduces the OdeNet ontology and compares it with GermaNet. In Section V, we investigate the way similarity estimates can be combined that exhibit very different probability distributions. The obtained evaluation results are given in Section VI and discussed in Section VII. Finally, we conclude the paper in Section VIII with an overview of the accomplished results and possible future work.

II. Related Work

There is a multitude of existing approaches to estimating text similarity by means of ontologies. Liu and Wang [3] match each word of a text to a concept in an ontology and derive a vector representation for it consisting of its weighted one-hot-encoded hypernyms, hyponyms, and the matched concept itself, where the weights are specified beforehand and they assume the maximum value of 1 for the latter. An entire document can then be represented by the centroid vector

of all words in the documents. As usual, the comparison with other documents can be accomplished by applying the cosine measure on the centroids. In contrast to Liu and Wang, Mabotuwana et al. [4] disregard the hyponyms for constructing the word vectors and set the weight of a hypernym to the reciprocal of the number of nodes on the shortest path in the ontology from the matched concept to the hypernym. A downside of this method is that simple path length count is quite unreliable in capturing semantic similarity, which is a finding of Resnik [5]. Therefore, the latter introduced information content (IC), which is the negative logarithm of the occurrence probability of a word and aims to compensate for differences of semantic similarities between nodes of taxonomy edges. The IC constitutes also the basis for several novel semantic similarity measures introduced by Lastra Díaz et al. [6], [7]. Mingxuan Liu and Xinghua Fan [8] propose enriching texts with semantically related words (hyponyms/hypernyms) to improve the categorization of short Chinese texts, which is the approach, we want to follow here. However, in contrast to Mingxuan Liu and Xinghua Fan, we will not represent the words occurring in the texts by ordinary sets but instead by fuzzy sets, that allow us to incorporate word vectors in our similarity score. The approach using fuzzy sets has the additional advantage that very general hypernyms or overly specific hyponyms, which are not really related to the input texts anymore and possibly introduce noise, can be downvoted.

All the state-of-the-art methods described so far return a single scalar value as a similarity estimator. however, Oleshshuk and Pedersen's approach derives a similarity vector, which represents the semantic similarities on different abstraction levels of the ontology as estimated by the Jaccard index [9].

An alternative approach to estimate semantic similarity is the use of word embeddings. These embeddings are determined beforehand on a very large corpus typically using either the skip-gram or the continuous bag-of-words variant of the Word2Vec model [10]. The skip-gram method aims to predict the textual surroundings of a given word by means of an artificial neural network. The influential weights of the one-hot-encoded input word to the nodes of the hidden layer constitute the embedding vector. For the so-called *continuous bag-of-words* method, it is just the opposite, i.e., the center word is predicted by the words in its surrounding. Alternatives to Word2Vec are GloVe [11], which is based on aggregated global word co-occurrence statistics and Explicit Semantic Analysis (ESA) [12], in which each word is represented by the column vector in the tf-idf matrix over Wikipedia. The idea of Word2Vec can be transferred to the level of sentences as well. In particular, the Skip-Thought vector model [13] derives a vector representation of the current sentence by predicting the surrounding sentences. An alternative to Skip-Thought vectors are Bert Sentence Embeddings that are based on a transformer architecture [14]. If vector space representations of the documents are established, a similarity estimate can then be obtained by applying the cosine measure on the embeddings centroids of the two documents to compare.

There is some prior work to devise similarity estimates combining ontologies and word embeddings. Faruqui et al.'s [15] approach aims to retrofit the embedding vectors in such a way that related words with respect to the employed ontology have preferably similar vector representations. Goikoetxea et al. [16] generate random walks on WordNet to extract sequences of concepts. These sequences are then fed into the ordinary Word2Vec to create (ontology) embeddings vectors. They evaluated several possibilities to combine such vectors with word embeddings, such as averaging or concatenating them. A downside of this approach in comparison with our proposed estimate is that at least 1 million of such random walks must be generated to obtain sufficiently reliable results. Therefore, the required format conversion, which needs to be repeated for every change in the ontology, is quite time-consuming.

III. Proposed Method

A straightforward and simple method to estimate the similarity between two texts is applying the Jaccard index on their bag-of-words representations [17, p. 299]. This coefficient is given as:

$$jacc(A,B) := \frac{|A \cap B|}{|A \cup B|} \tag{1}$$

where A (B) is the set of words of the first (second) text. In this scenario, the first text is the snippet entered by the user and the second text is the keyword description of the youth milieu.

An alternative to the Jaccard index is the Dice coefficient [17], which is defined as follows:

$$DSC(A,B) := \frac{2|A \cap B|}{|A| + |B|} \tag{2}$$

One can define distance measures for these two coefficients, which are called the Jaccard distance and the Dice distance, respectively, by subtracting them from 1. In contrast to the Jaccard distance, the Dice distance does not satisfy the triangle inequality [18, p. 29] and is therefore not a proper distance metric. Note that the Dice coefficient and Jaccard index can be transformed into each other by the following formulas:

$$\begin{aligned} jacc(A,B) &= DSC(A,B)/(2 - DSC(A,B)) \\ DSC(A,B) &= 2\,jacc(A,B)/(1 + jacc(A,B)) \end{aligned} \tag{3}$$

Furthermore, we consider the overlap coefficient, which is given by [17, p. 299]:

$$overlap(A,B) := \frac{A \cap B}{\min\{|A|, |B|\}} \tag{4}$$

It assumes the maximum value of 1 if either one of the input sets is a subset of the other.

While these coefficients work reasonably well for long texts, they usually fails for short text snippets since in this case, it is very likely that all overlaps are caused by very common words (typical stop words), which are actually irrelevant for estimating text similarity. One possibility to increase the number of overlaps is to extend the two texts by means of an ontology [8], i.e., adding to a text the words from the ontology that are semantically close (hence reachable by a short path) to the words of that text. In particular, we decided to add all synonyms, hypernyms, and direct hyponyms of all words appearing in the investigated text. Hereby we follow the hypothesis of Rada et al. [19], which states that taxonomic relations are sufficient to capture semantic similarity between ontology concepts. Note that hyponyms and hypernyms may not be uniquely defined since a single word can occur in

several synsets. In principle, two possibilities to deal with arise in this situation:

1) Use hyponyms / hypernyms of all possible synsets for the expansion
2) Employ Word Sense Disambiguation to select only the synset that corresponds to the indented meaning of the word. The drawback of this approach is that the Word Sense Disambiguation might choose the incorrect synset, especially with short text snippets, which can result in missing overlaps and therefore inexact similarity estimates.

Currently, we use possibility 1 but consider possibility 2 for a future version of our approach.

The two sets used in the coefficients stated above (Jaccard, Dice, and Overlap) are crisp, which means that all words are treated alike. However, the words that are newly induced by the ontology are probably less reliable for capturing the semantics of the text than the original words are. Furthermore, not all of the newly introduced words are equally relevant. However, our current model cannot capture those relationships. Therefore, we extend our set representation to allow for fuzziness (i.e., we employ fuzzy sets instead of conventional crisp sets).

For conventional sets, the decision of whether an element belongs to this set is always binary (i.e., it can uniquely be decided whether an element belongs to a set or not). This is different from a fuzzy set, where the membership of an element can be partial. In particular, each fuzzy set is assigned a real-valued function $\mu : X \rightarrow [0, 1]$ (X: all potential elements of our set) assuming values in the interval [0,1] and specifying the degree of membership for all elements. If this membership function only assumed the values 0 or 1, the fuzzy set would actually be equivalent to a conventional set.

Set union and intersection are also defined in terms of fuzzy sets, namely in the following way:

$$\begin{aligned} \mu_{A \cap B} &= \min\{\mu_A, \mu_B\} \\ \mu_{A \cup B} &= \max\{\mu_A, \mu_B\} \end{aligned} \tag{5}$$

The cardinality of a fuzzy set is defined as the total sum over all membership values:

$$|F| := \sum_{x \in X} \mu_F(x)$$

With intersection, union, and fuzzy set cardinality, all three coefficients described above (Jaccard, Dice, and Overlap) can be defined for fuzzy sets analogously to ordinary sets.

In addition to these coefficients, we also employ pointwise mutual information, which is defined as:

$$pmi(A, B) := lb\left(\frac{P(A \cap B)}{P(A)P(B)}\right) \tag{6}$$

where $A \cap B$ denotes the Fuzzy set intersection between A and B. The probability of a fuzzy event represented in the form of a fuzzy set E is given by $|E|/n$ [20], where n denotes the number of elements in the fuzzy set, in this case all lemmas of the German language possibly occurring in one of the texts. Note that the cardinality of E is defined as the sum of all fuzzy membership values and is therefore different from n.

To avoid dealing with negative infinity values of the pointwise mutual information, which occur if the sets A and B are disjoint, we follow the approach of [21] and clip all values less than -2.0. To combine the pointwise mutual information with the Word2Vec-based similarity estimate, we linearly scale all its values into the interval $[0, 1]$.

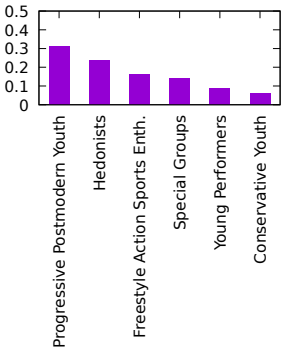


Figure 1. Distribution of youth milieus over the three contests.

What remains is to define the fuzzy membership function. Let $Cent(A)$ be the word embeddings centroid of our original words. We then define the membership function μ as follows:

$$\mu(w) := (\max\{0, \cos(\angle(Cent(A), Emb(w)))\})^i \tag{7}$$

where $Emb(w)$ is the embedding vector of a word w and the use of the maximum operator prevents the membership value from being complex. The exponent i allows us to adjust the influence of the word embeddings gradually. Full influence is obtained by setting i to 1. In contrast, the influence diminishes if i is set to 0.

Our similarity estimate is then used to assign user responds from several online contests in form of short text snippets to the best fitting youth milieu out of *Progressive Postmodern Youth* (people primarily interested in culture and arts), *Young Performers* (people striving for a high salary with a strong affinity to luxury goods), *Freestyle Action Sports Enthusiasts*, *Hedonists* (rather poorly educated people who enjoy partying and disco music), and *Conservative Youth* (traditional people with a strong concern for security). A sixth milieu called *Special Groups* comprises all those who cannot be assigned to one of the upper five milieus. The distribution of the 6 milieus over the three considered contests is given in Figure 1 as a histogram. This figure shows that the milieus are quite unevenly distributed with the most frequent milieu *Progressive Postmodern Youth* appearing around five times more often than the rarest one (*Conservative Youth*).

For each milieu (with the exception of *Special Groups*) a keyword list was manually created to describe its main characteristics (see Table I). To trigger marketing campaigns, an algorithm has been developed that automatically assigns each contest answer to the most likely target group: we propose the youth milieu as the best match for a contest answer, for which the estimated semantic similarity between the associated keyword list and user respond is maximal. In case the highest similarity estimate falls below the 10 percent quantile for the distribution of highest estimates, the *Special Groups* milieu is selected.

TABLE I. KEYWORD LISTS DESCRIBING THE YOUTH MILIEUS.

Youth milieu	Keywords
Progressive Postmodern Youth	clothing, music, art, freedom, culture, educated
Young Performers	rich, elite, luxury, luxurious
Freestyle Action Sports Enthusiasts	Sports, Fitness, Music
Hedonists	poor, communication, self-fulfilment, entertainment, party, music, disco
Conservative Youth	conservation of value, conservativism, citizenship, Switzerland

TABLE II. EXAMPLE USER ANSWER FOR THE TRAVEL DESTINATION CONTEST (TRANSLATED INTO ENGLISH).

Choice	Country	Snippet
1	Jordan	Ride through the desert and marvel at Petra during sunrise before the arrival of tourist buses
2	Cook Island	Snorkelling with whale sharks and relax
3	USA	Experiencing an awesome week at the Burning Man Festival

The ontology we employ for our similarity estimate is OdeNet, which is a freely available lexical resource recently developed by the Darmstadt University of Applied Sciences that will be explained in more detail in the next section.

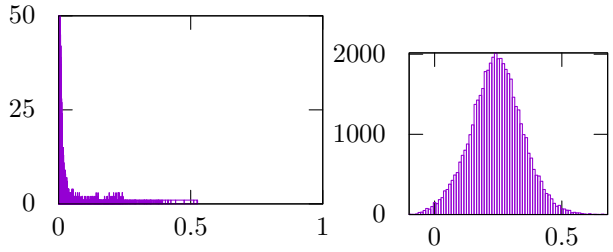
IV. ODENET ONTOLOGY

Freely available machine-readable lexical ontologies for German are rather sparse. On the one hand, there are websites such as Wiktionary and Open-Thesaurus, which are targeted at human users. Much effort would need to be spent to bring the associated resources to a form that can be exploited efficiently by a computer. On the other hand, there is GermaNet [22], which is suitable both for human users as well as for automated processing. However, GermaNet is not a free resource. While it may be freely used in purely academic projects, as soon as industry partners are involved, the academic license is no longer eligible and the project partners have to sign a commercial license agreement.

The lexical ontology of OdeNet [23] is devised to fill this gap. It has been compiled automatically from the Open-Thesaurus synonym lexicon (https://www.openthesaurus.de/), the Princeton WordNet of English [24], and the Open Multilingual WordNet English [25]. Afterwards, it was manually error-checked and applied to comprehensive revisions. Similar to WordNet, semantic concepts are represented by synsets, which are interconnected by linguistic and semantic relations such as hyponymy, hypernymy, meronymy, holonymy, and antonymy. In total, it currently contains around 120 000 lexical entries and 36 000 synsets. The entire resource is available as an XML file obtainable at Github [26]. We found OdeNet to be very easy to use and well-designed.

V. COMBINING SIMILARITY SCORES

Besides our ontology based measure, we implemented several other measures such as ESA, the cosine of word embedding centroids, Skip-Thought vectors, etc. Usually, a stronger and more reliable similarity estimate can be obtained by combining measures. One possibility for that is majority vote, i.e., suggesting the class that most of the measures



(a) Ontology-based estimate (Jaccard Index).

(b) Cosine of embeddings centroids.

Figure 2. Histograms of similarity estimates.

suggest. One drawback of majority vote is that the individual measures should be of comparable performance and that we need at least three of them. Furthermore, a majority vote only returns a decision for one of the classes but no (numerical) score. However, we actually need such a score to determine the 10 percent quantile (cf. previous section). An alternative to a majority vote is a weighted average. Albeit, there is again an obstacle. While all our semantic similarity estimates assume values between 0 and 1 (Note that the cosine of word embeddings centroids can assume (usually small) negative values as well.), their distributions can be quite different (see Figure 2). Considering this case, we would like to combine the cosine of word embedding centroids and our ontology based similarity measure by a weighted sum. The first type of estimate is normally distributed and covers almost the entire value range. However, although in principle our ontology based similarity estimate can reach the value of 1, most of its values are located inside the interval [0,0.1]. To make both estimates comparable with each other, we are conducting a histogram equalization for them prior to their combination. Such an equalization levels out the relative occurrence frequencies of estimate intervals, so that the resulting values are approximately uniformly distributed. This is accomplished by transforming the similarity estimates using the cumulative probability distribution function cdf. Formally, an estimate s is mapped to the value $cdf(s)$. One downside of our method is that the resulting similarity estimate is probably biased. However, in our scenario, we are not so much interested in the actual value of our estimate but instead focusing mainly on the correct ranking of target groups. Thus, the modification of the estimate's probability distribution is unproblematic. The combined estimate sim is formally given as:

$$sim := w \cdot cdf(sim_{odenet}) + (1 - w) \cdot cdf(sim_{w2v}) \quad (8)$$

where

- w: in the influencing weight of the OdeNet similarity based estimate, the default value is 0.5
- sim_{odenet}: the score obtained by the OdeNet similarity estimate (Jaccard, Dice, Overlap, or Pointwise Mutual Information over fuzzy sets)
- sim_{w2v}: cosine of the angle between the Word2Vec embeddings centroids of user text snippet and youth milieu keyword description

TABLE III. OBTAINED ACCURACY VALUES FOR SEVERAL SIMILARITY ESTIMATES. ODENET+EMB.: LINEAR COMBINATION OF OUR ONTOLOGY BASED MEASURE WITH COSINE OF WORD EMBEDDINGS CENTROIDS. RW=RANDOM WALK BASED METHOD PROPOSED BY GOIKOETXEA ET AL., STV=SKIP-THOUGHT VECTORS, JC=JACCARD INDEX, OL=OVERLAP COEFFICIENT, PMI=POINTWISE MUTUAL INFORMATION, HE=HISTOGRAM EQUALIZATION [16]

Method	Contest 1	2	3	Total
Random	0.172	0.149	0.197	0.172
Jaccard	0.150	0.194	0.045	0.142
W2V	0.348	**0.328**	0.227	0.330
ESA	0.357	0.254	**0.288**	0.335
RW	0.281	0.149	0.273	0.263
Bert	0.109	0.149	0.136	0.118
STV	0.162	0.284	0.273	0.191
Emb.+JC	0.266	0.313	0.227	0.267
Emb.+JC (HE)	0.347	**0.328**	0.227	0.330
OdeNet(JC,crisp)	0.367	0.194	0.273	0.333
OdeNet(JC)	0.309	0.224	0.212	0.286
OdeNet(JC)+Emb.	**0.380**	0.269	0.273	**0.352**
OdeNet+Emb.+Mero	0.372	0.254	0.273	0.345
OdeNet(OL)+Emb.	0.370	0.209	**0.288**	0.339
OdeNet(Dice)+Emb.	0.372	0.254	0.273	0.345
OdeNet(PMI)+Emb.	0.370	0.224	**0.288**	0.341

TABLE IV. MINIMUM AND MAXIMUM AVERAGE INTER-ANNOTATOR AGREEMENTS (COHEN'S KAPPA).

Method	Contest 1	2	3
Min kappa	0.123	0.295/0.030	0.110/0.101
Max. kappa	0.178	0.345/0.149	0.114/0.209
# Annotated entries	1543	100	100

VI. EVALUATION

For evaluation, we selected three online contests (language: German), where people elaborated on their favorite travel destination (contest 1, see Table II for an example), speculated about potential experiences with a pair of fancy sneakers (contest 2) and explained why they emotionally prefer a certain product out of four available candidates. In a bid to provide a gold standard, three professional marketers from different youth marketing companies annotated independently the best matching youth milieus for every contest answer. We determined for each annotator individually his/her average inter-annotator agreement with the others (Cohen's kappa). The minimum and maximum of these average agreement values are given in Table IV. Since for contests 2 and 3, some of the annotators considered only the first 50 entries (last 50 entries respectively), we specified min/max average kappa values for both parts.

Before automatically distributing the texts to the youth milieus, we applied on them a linguistic preprocessing consisting of tokenization, stop word filtering, lemmatization, and compound analysis. The latter was used to determine the base form of each word, which was added as an additional token. Next to our own similarity estimates, we evaluated several baseline methods, in particular random assignments, Jaccard, ESA, the ontology-based approach of Goikoetxea et al. [16], cosine of word embedding centroids, Skip-Thought vectors, and Bert embeddings. The accuracy values given in Table III are obtained by comparing automated assignments with the majority vote of the assignments conducted by our human annotators. Since the keyword lists used to describe the characteristics of the youth milieus typically consist of nouns (in the German language capitalized) and the user contest answers might contain many adjectives and verbs as well, which do not match nouns very well in the Word2Vec vector

TABLE V. CORPUS SIZES MEASURED BY NUMBER OF WORDS.

Corpus	# Words
German Wikipedia	651 880 623
Frankfurter Rundschau	34 325 073
News journal *20 Minutes*	8 629 955

TABLE VI. PRECISION, RECALL AND F1-SCORE OBTAINED FOR THE THE YOUTH MILIEUS USING THE ONTOLOGY-BASED ESTIMATE (JACCARD-INDEX).

Milieu	Precision	Recall	F1-score
Special Groups	0.548	0.453	0.496
Freestyle Action Sports Enthusiasts	0.374	0.506	0.430
Hedonists	0.287	0.565	0.380
Progressive Postmodern Youth	0.507	0.211	0.298
Young Performers	0.091	0.064	0.075
Conservative Youth	0.200	0.032	0.056
All	0.335	0.305	0.289

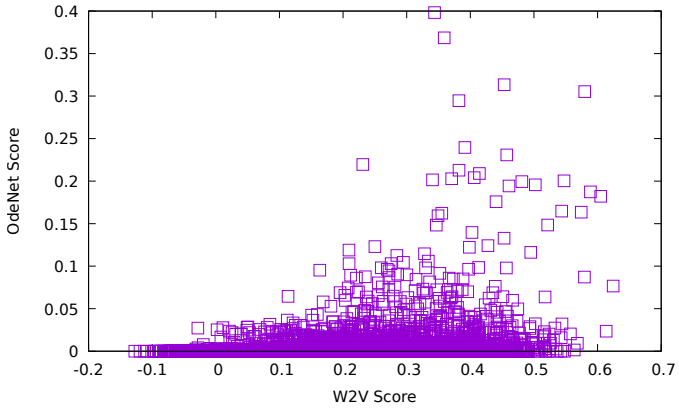


Figure 3. Scatterplot: Word2Vec (W2V) Embeddings Score vs OdeNet Score.

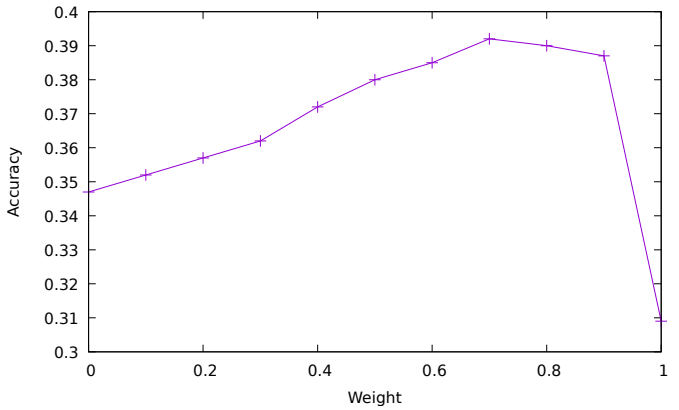


Figure 4. Accuracy of combined Word2Vec Embeddings and OdeNet score with respect to the influencing OdeNet score weight w.

representation, we actually conduct two comparisons for the Word2Vec centroids based on similarity estimate, one with the unchanged user contest answers and one by capitalizing every word beforehand. The final similarity estimate is then given as the maximum value of both individual estimates. For our proposed ontology based similarity estimate, we use the parameter settings $i = 0.5$ and weights of linear combination: 0.5, which performed best in several experiments with varying parameter values. Setting i to 0.5 seems to us to be a good compromise between considering only the ontology structure ($i = 0$) and fully weighting the word embedding vectors ($i = 1$).

In total, the following methods are evaluated:

1) Random: Just randomly assign one of the youth milieus to a text snippet
2) Jaccard: Estimate the semantic similarity of a text snippet and youth milieu keywords by applying the Jaccard index directly to their bag of words representations
3) Word2Vec: Estimate the semantic text similarity by applying the cosine measure to the word embedding centroids
4) ESA: Estimate the semantic text similarity by applying the cosine measure to the ESA embedding centroids
5) RW: Similarity estimate based on random walks over an ontology (here: OdeNet, in the original paper: WordNet) as proposed by Goikoetxea et al.
6) Bert: Estimate the semantic text similarity based on the centroids of Bert embeddings
7) STV: Estimate the semantic text similarity based on the centroids of Skip-Thought vectors
8) Emb.+JC: Averaging estimates of methods 2 and 3
9) Emb.+JC (HE): The same as above but additionally conducting a histogram equalization.
10) OdeNet (JC,crisp): OdeNet based similarity measure using the Jaccard with exponent i set to 0 which results in crisp (non-fuzzy) sets
11) OdeNet (JC): OdeNet based similarity estimate employing Jaccard index with exponent i set to 0.5
12) OdeNet (JC)+Emb.: Averaging estimates of methods 3 and 11
13) OdeNet (JC)+Emb+Mero: Averaging estimates of methods 3 and 11, where in method 11, the lemmas are expanded not only by hyponyms but also by meronyms
14) OdeNet (OL)+Emb: Similar to method 11 but instead of Jaccard the overlap index is used
15) OdeNet (Dice)+Emb: Similar to method 11 but instead of Jaccard the Dice index is used
16) OdeNet (PMI)+Emb: Similar to method 11 but instead of Jaccard Pointwise Mutual Information is used

The Word2Vec word embeddings were trained on the German Wikipedia (dump originating from 20 February 2017) merged with a Frankfurter Rundschau newspaper corpus and 34 249 articles of the news journal *20 minutes*, where the latter is targeted to the Swiss market and freely available at many Swiss train stations (see Table V for a comparison of corpus sizes). By employing articles from *20 minutes*, we want to ensure the reliability of word vectors for certain Switzerland specific expressions such as *Velo* or *Glace*, which are underrepresented in the German Wikipedia and the Frankfurter Rundschau corpus.

The accuracy of the combined OdeNet / Word2Vec embedding score with respect to the weight of the OdeNet score is given in Figure 4. This diagram shows that the maximum accuracy value is obtained at a weight of $w = 0.8$, which demonstrates that a rather large OdeNet weight is required for obtaining a high accuracy. This fact seems a bit surprising, since the standalone OdeNet similarity estimate performs much poorer than its Word2Vec embedding counterpart.

Furthermore, we give the F1-score of the combined OdeNet / W2V embedding similarity estimate for the individual youth milieus in Table VI. The highest F1-score is obtained for the *Special Groups* youth milieu, which insinuates that oftentimes the appropriate youth milieu is not expressed by the contest participants in the text snippets. The second-best detectable milieu is *Freestyle Action Sports Enthusiasts*, which is caused by the fact that in the first and largest contest containing elaborations of possible dream holidays, the participants frequently mention sports activities such as the surfing or snorkeling that they plan to conduct.

Finally, the scatter plot of the OdeNet (Jaccard) vs Word2Vec embedding similarity estimate is specified in Figure 3. This plot demonstrates that the relationship between both estimates is highly nonlinear and the Ontology-based estimate frequently scores text pairs rather low that assume a high Word2Vec Embeddings estimate value.

VII. Discussion

The evaluation shows that although our ontology based method lags behind the cosine of Word2Vec centroids in terms of accuracy, their linear combination performs considerably better than both of the methods alone. Furthermore, it outperforms both its crisp counterpart (exponent i:=0), the approach of Goikoetxea et al. if applied to OdeNet, used with 100 million random walk restarts, and combined with Word2Vec word embeddings by vector concatenation (RW in Table III) and also two deep learning based approaches (Skip-Thought vectors and Bert embeddings [14]). The rather low accuracy of both deep learning approaches (Skip-Thought and Bert) is caused by the fact that the words of the keyword lists describing the youth milieus are arbitrarily ordered and therefore these lists can not be captured sufficiently well by a language model trained on ordinary texts like Wikipedia. In further experiments, we could show that especially Bert embeddings are very vulnerable to ungrammatical input. For instance, a simple stop word filtering degrades its performance already considerably.

Remarkable is the low performance of our approach on contest 2. Further analysis revealed that in several cases the correct youth milieu in this contest was indicated by the only word that was either a town name (“Basel”) or a rather rare noun that is not contained in OdeNet, which demonstrates that the given ontology is indeed very useful for estimating semantic similarity.

Note that the OdeNet ontology is still under active development and contains several gaps in the semantic relations. For instance, it comprises no hyponyms of *sports*, which makes it difficult to correctly assign people to the *Freestyle Action Sports Enthusiasts* target group. Another downside is that OdeNet contains no inflected forms so far. Thus, we have to employ a lemmatizer in order to identify hyponyms and hypernyms for such word forms.

VIII. Conclusion and Future Work

We presented a similarity estimate based both on word embeddings and OdeNet ontology. In contrast to most state-of-the-art methods, it can directly employ the given ontology format. Time consuming format conversions into vectors or matrices are not necessary, which simplifies its usage significantly. Additionally, by using fuzzy sets, hypernyms/hyponyms introduced by the ontology that are too general/specific and therefore not really related to the input texts any more, can be downvoted. The application scenario is targeted marketing, in which we aim to match people to the best fitting marketing target group based on short German text snippets. The evaluation showed that the obtained accuracy of a baseline method considerably increases if combined by a linear combination with our ontology based estimate. In general, this estimate attains a good performance, if the ontology contains the key terms relvant for the application scenario. As future work we want to further investigate hybrid data-driven and knowledge-based semantic similarity estimates. In particular, we plan to employ additional semantic relations besides hypernyms, hyponyms, synonyms, and meronyms such as holonyms or antonyms. Furthermore, all the model parameters are currently manually specified. It would be preferable to determine them automatically through the use of grid search or more sophisticated Artificial Intelligence methods such as Bayesian search [27]. Finally, we want to experiment with other types of hierarchically ordered lexical resources, which are not necessarily ontologies, such as the Wikipedia category taxonomy.

Acknowledgment

We thank Jaywalker GmbH as well as Jaywalker Digital AG for their support regarding this publication and especially for annotating the contest data with the best-fitting youth milieus. This research has been funded by the FMSquare Stiftung, an international foundation for the promotion of fuzzy management methods.

References

[1] T. vor der Brück, "Estimating semantic similarity for targeted marketing based on fuzzy sets and the odenet ontology," in International Conference on Advances in Semantic Processing, 2018.

[2] M. Lynn, "Segmenting and targeting your market: Strategies and limitations," Cornell University, Tech. Rep., 2011, online: http://scholorship.sha.cornell.edu/articles/243 [retrieved: 09/2018].

[3] H. Liu and P. Wang, "Assessing text semantic similarity using ontology," Journal of Software, vol. 9, 2014.

[4] T. Mabotuwana, M. C. Lee, and E. V. Cohen-Solal, "An ontology-based similarity measure for biomedical data - application to radiology reports," Journal of Biomedial Informatics, vol. 46, 2013.

[5] P. Resnik, "Using information content to evaluate semantic similarity in a taxonomy," in Proceedings of the 14th International Joint Conference on Artificial Intelligence (IJCAI), 1995.

[6] J. J. Lastra-Díaz and A. García-Serrano, "A novel family of IC-based similarity measures with a detailed experimental survey on WordNet," Engineering Applications of Artificial Intelligence, vol. 46, 2015, pp. 140–153.

[7] J. J. Lastra-Díaz, A. García-Serrano, M. Batet, M. Fernández, and F. Chirigati, "HESML: A scalable ontology-based semantic similarity measures library with a set of reproducably experiments and a replication dataset," Information Systems, vol. 66, 2017, pp. 97–118.

[8] M. Liu and X. Fan, "A method for Chinese short text classification considering effective feature expansion," Internation Journal of Advanced Research in Artificial Intelligence, vol. 1, no. 1, 2012.

[9] V. Oleschchuk and A. Pedersen, "Ontology based semantic similarity comparison of documents," in Proceedings of the 14th International Workshop on Database and Expert Systems Applications (DEXA), 2003.

[10] T. Mikolov, I. Sutskever, C. Ilya, G. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," in Proceedings of the Conference on Neural Information Processing Systems (NIPS), Lake Tahoe, Nevada, 2013, pp. 3111–3119.

[11] J. Pennington, R. Socher, and C. D. Manning, "Glove: Global vectors for word representation," in Proceedings of the Conference on Empirical Methods on Natural Language Processing (EMNLP), Doha, Katar, 2014.

[12] E. Gabrilovic and S. Markovitch, "Wikipedia-based semantic interpretation for natural language processing," Journal of Artificial Intelligence Research, vol. 34, 2009.

[13] R. Kiros et al., "Skip-Thought vectors," in Proceedings of the Conference on Neural Information Processing Systems (NIPS), Montréal, Canada, 2015.

[14] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," in Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics (NAACL), 2019.

[15] M. Faruqui et al., "Retrofitting word vectors to semantic lexicons," in Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics (NAACL), 2015.

[16] J. Goikoetxea, E. Agirre, and A. Soroa, "Single or multiple? Combining word representations independently learned from text and WordNet," in Proceedings of the AAAI Conference on Artificial Intelligence, Phoenix, Arizona USA, 2016.

[17] C. D. Manning and H. Schütze, Foundations of Statistical Natural Language Processing. MIT Press, 1999.

[18] K. Latha, Experiment and Evaluation in Information Retrieval Models. Boca Raton, Florida: CRC Press, 2018.

[19] R. Rada, H. Mili, E. Bicknell, and M. Blettner, "Development and application of a metric on semantic nets," IEEE Transactions on Systems, Man, and Cybernetics, vol. 19, no. 1, 1989, pp. 17–30.

[20] O. Pavlačka1 and P. Rotterová, "Probability of fuzzy events," in 32nd International Conference on Mathematical Methods in Economics, 2014, pp. 760–765.

[21] A. Salle and A. Villavicencio, "Why so down? the role of negative (and positive) pointwise mutual information in distributional semantics," CoRR, vol. abs/1908.06941, 2019. [Online]. Available: http://arxiv.org/abs/1908.06941

[22] B. Hamp and H. Feldweg, "GermaNet - a lexical-semantic net for German," in Proceedings of the ACL workshop Automatic Information Extraction and Building of Lexical Semantic Resources for NLP Applications, 1997.

[23] M. Siegel and F. Bond, "OdeNet: Compiling a German WordNet from other resources," in Proceedings of the 12th Global WordNet Conference, 2021, pp. 192–198.

[24] C. Fellbaum, Ed., WordNet: An Electronic Lexical Database. Cambridge, MA: MIT Press. Cambridge, Massachusetts: MIT Press, 1998.

[25] F. Bond and R. Foster, "Linking and extending an open multilingual wordnet," in Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics, Sofia, Bulgaria, 2013, pp. 1352–1362.

[26] M. Siegel et al., "OdeNet," last access: 11/22/2021. [Online]. Available: https://github.com/hdaSprachtechnologie/odenet

[27] J. Snoek, H. Larochelle, and R. P. Adams, "Practical Bayesian optimization of machine learning algorithms," in Proceedings of the Conference on Neural Information Processing Systems (NIPS), 2012, pp. 2951–2959.

Building a Decision-Making System for Handling a Drone Operator's Emotional States Using a Brain-Computer Interface

Diana Ramos
Faculdade de Engenharia da Universidade do Porto
Capgemini Engineering
Porto, Portugal
email: dianacristina.teixeiraramos@altran.com

Gil Gonçalves
Faculdade de Engenharia da Universidade do Porto
Porto, Portugal
email: gil@fe.up.pt

Ricardo Faria
R&D Tech Leader
Capgemini Engineering
Lisboa, Portugal
email: ricardoandre.pintofaria@altran.com

***Abstract*—Drones enable humans to perform certain high-risk and attention operations and safety-critical tasks remotely, which are boosted by the use of Brain-Computer Interfaces. However, these technologies are correlated with the cognitive state of the operator, who is prone to stress and diversions, which brings instability to drone control. In this paper, we propose a decision making system aiming to decide, upon the operator's emotional state, whether the command should or should not be sent to the drone. By building a predictive operator's digital twin for cognitive emotional detection and by benefiting from a visual facial expression classifier, this system computes the coordinates and sends them to the drone through a Robot Operating System 2 client. Results show that both the digital twin and the facial expression classifier are capable of detecting emotions in a real-time setting and the system provides a reliable and secure way of commanding drones through the mind. Drone swarms could be integrated as this solution eases the addition of more ROS2 client nodes.**

***Keywords*—drone; Brain-Computer Interface; digital twin; Robot Operating System 2; drone swarms.**

I. Introduction

This paper showcases the implementation of a drone operator digital twin that relies on a brain-computer interface available data streams in order to evaluate whether the operator is in a suitable condition to send commands to the drone. This work is an extended version of [1] (that detail preliminary results regarding human emotion recognition); thus, details regarding the modelling of the cognitive digital are further analyzed and explained in this document as well as different validation tasks that comply with other requirements adjacent to the main purpose of the work (i.e., experiments with more than one drone as proof-of-concept of a ROS2 client-subscriber communication system for multi-drone control).

The drone sector has been growing with higher demand through the years. The common belief is that drones are singularly used for military affairs; however, they are functional and versatile systems. One major use is providing monitoring services, i.e., target searching, surveillance for security purposes and others.

Even thought they have attracted companies due to their visionary application, the most significant change is how civilians have been adopting this technology in their lives. Photography and cinematography are key activities that lead to potential customer interest because of the accessibility drones provide to reach high places, enough to capture a panoramic shot. Either way, drones are essentially useful for people with less motor skills that costly go through their everyday routines. In this case, a drone could serve as an assistive device [2].

Still, the most impactful usage of drones is their applicability to complete high-risk and safety-critical operations with success, often in locations unreachable and/or dangerous to humans. Although there is a risk of compromising the mission due to faulty hardware or control management, the drone operator is isolated from the target site, which ensures the safety of all the stakeholders involved.

A. Problem Overview

Drones are considered critical systems. By definition, a critical system is a technology that brings its inherent risks while executing, whose failures could be significantly damaging [3]. For instance, the failure of these systems could lead to financial loss where the hardware could be damaged by a collision with other objects or could compromise the mission itself by a poor performance from the operator. Nonetheless, controlling one drone is already a complex task. Operators are responsible for, not just to perform standard operations (i.e., takeoff and landing) with success, but also to safely execute them. When adding unsafe and critical operations to the task log, the control complexity increases significantly. The operator needs abilities at their peak, full attention and focus when performing these operations, to provide a reliable and stable control.

Hand control allows operators to remotely send commands to drones; however, as these are critical systems, operators need to be cautious with the commands they deliver. The Brain-Computer Interface (or BCI) is an alternative mechanism that aims to optimize this control. As humans are prone to

fatigue, increasing mental workload and emotions, the control will become uncertain and insecure, that is, the operator cannot be fully consistent with his performance for longer periods of time due to the organic degrading factors of its human condition. These situations can lead to events of disruption and/or disasters, such as the collision with objects that the system cannot autonomously detect.

All things considered, the problem of controlling a drone with a BCI rises on the addition of the human factor and his involuntary natural incapability of managing a critical system consistently. To address this issue, this work focuses, primarily, on how to reduce or avoid the operational impact on the drone when the operator sends a command under the influence of a negative emotion.

B. Proposed Solution

The hypothesis of this work is that, by adopting a digital twin [6] to virtually represent the operator and by using machine learning techniques, it is possible to process, filter and predict whether the human operator has high mental workload and/or impactful emotions and decide whether the commands produced should or should not be sent to the drones. With the goal of validating the formulated commands, the digital twin is complemented with a visual emotion recognizer that will classify the operator's visual facial expression into a set of emotional states. Additionally, a Robot Operating System 2 (or ROS2) client node can be used in order to send the commands to the drone.

C. Structure of the Document

This document is structured in seven sections: (I) the current section detailing the context and proposed solution to accommodate the urged necessities mentioned; (II) literature review of the state-of-the-art key technologies selected in the proposed solution; (III) that details the methodology adopted for the development of the digital twin; (IV) implementation insights for the development of the solution; (V) validation of the solution including an incremental test environment for measuring the accuracy and other suitable test scenarios for showcasing the value of the developed system; (VI) conclusion regarding the system and its performance and (VII) listing goals to achieve in future work.

II. State-of-Art

A BCI is defined as "a device that connects the brain to a computer and decodes in real-time a specific, predefined brain activity" [4]. This technology can use direct or indirect methods to do so, namely by evaluating the nerve cells activity or by assessing the levels of blood oxygen for these cells [4]. This technology has proved its relevance in many areas, for instance, there was a study aiming to deliver accurate real-time and precise command classification for drone reliable control. An Electroencephalography (or EEG) headset was used to record the brain activity, followed by a motor imagery acquisition. This mechanism involved four tasks, based on the subject visualizing physical movements instead of performing them. Then, a classification methodology was developed by combining the Common Spatial Paradigm (or CSP) and the Linear Discriminant Analysis algorithms (LDA) [5]. Using this method, the authors were able to improve classification precision in real-time. The solution was validated using a fixed-wing drone use case [5].

Another crucial component of this work is the digital twin. Nowadays, a virtual twin is described as a virtual representation that carries information to realistically behave and change as a physical hardware [6]. This technology is constantly evolving to serve each project needs. One variant that derives from it is the digital twin environment [6] with predicting capabilities. The main goal is to train the digital twin to gain predictive capabilities in order to anticipate the hardware's response or behaviour in situational events during run time. One example is a research work that proposes a framework to improve the estimates of certain measurements of physical systems, more specifically a drone, by implementing a virtual layer, i.e., a digital twin, that would represent the real device and predict its performance [7]. This approach implies that each piece of the drone has its own prediction models that should learn and be updated through time to, ultimately, accurately anticipate some metrics that are valuable to the end-user.

The goal of this work is to predict the emotional state of a drone operator in real-time. In this specific area, there are some articles that detail visionary approaches on how to analyse the mental workload of a subject or, more importantly, the emotional spectrum. One of these works [8] makes use of a dataset composed by sounds with the goal of triggering certain emotions on the subject. The *arousal* (or excitement) and *valence* (defines the whether the emotion is positive or negative) (Figure 1) are measurements that are computed according to a set of frequency channels of the brain-wave activity, for instance, the *alpha* and *beta* signals. By using machine learning techniques and by training with the Linear Discriminant Analysis (or LDA) and Support Vector Machine (or SVM) algorithms, the authors were able to classify the emotional states of the subjects in the following categories: *happiness, anger, sadness*, and *calm*.

Even though there is a wide range of projects that target these key technologies, this particular research area that crosses the digital twin framework with a BCI for drone control is undeveloped. While the BCI mechanism is being targeted for scientific purposes, it is possible to bring it as a control mechanism of drone systems (as validated in [9]). However, these studies suggest the remote control of drones whenever as far as the human operator can continuously formulate commands. In contrast, this work allows the evaluation of the human emotional state to permit or break this cycle of continuous commands, in the context of the execution of high-risk tasks (where the error margin is very limited). Ultimately, the purpose is to establish a mental control mechanism and assessment of the human condition, while incorporating an efficient communication channel with ROS2.

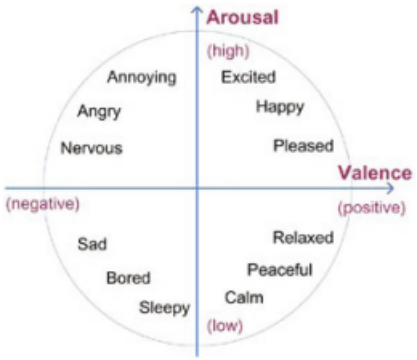


Figure 1. Emotional map based on *arousal* and *valence* (from [8]).

III. Research Methodology

A central piece of this work is the applicability of machine learning techniques to build a cognitive profile of a drone operator (the digital twin). In this work, the Cross Industry Standard Process for Data Mining (or CRISP-DM) methodology [10] was used. This dictates a systematic approach for building an intelligent software component by means of data analysis and application of machine learning algorithms for knowledge gathering. The main steps adopted are the following:

1) **Data understanding:** data acquisition from known and reliable sources that delivers important information for the problem, in addition to analysis of this data to acquire knowledge on its quality and required processing (i.e., finding missing values).
2) **Data Preparation:** prepared the acquire data for the modeling phase based on prior knowledge gathered from the data analysis. This task consumes the most resources, because it has a direct impact on the quality and performance of the final digital twin.
3) **Modeling:** manipulate the previous dataset to build a prediction model, i.e., the digital twin. In this phase, training multiple machine learning algorithms and validating in a set of performance metrics is crucial to assess which algorithm best fits the proposed problem.
4) **Evaluation:** the best fitted prediction model and its outcomes provides a way of gathering more knowledge and validate if all goals established are being met.

A. Brain-Computer Interface Headset

As a starting point, the *Emotiv Epoc+* [11] headset was chosen (Figure 2), developed by the *Emotiv* company, for data acquisition due to its portability and reliability as a commercial BCI. *Emotiv Epoc+* provides a source of data of interest since EEG signals to the motion of the head according to a set of 3-axis. It is composed by *arms* that fit on specific locations on the scalp and are arranged accordingly to ease the subject while placing it on the head. In addition, each sensor is embedded with a felt tip. Dumping this felt tip is crucial for the signal quality since it will connect the whole assembly to the scalp and allow brain activity to be detected (for more information check [11]).

Figure 2. *Emotiv Epoc+* hardware.

Between the multiple software that allow the exploration of data streams, it was used the *EmotivBCI*, developed by the *Emotiv* company. This provides a platform for direct command and facial expression's training and monitoring of multiple data streams. Fundamentally, in order to control a drone with a BCI, a first approach is to create a set of commands to be operational during experiments. This task implies the creation of strategies to be mentally reproducible whenever the operator desires. In this case, the operator could reproduce a certain command by visualizing the movement to the matching direction, followed by the natural eye movement. This procedure works around a method of training the command and testing in a live environment, both provided by the platform. The application is supported by a machine learning prediction model to build a profile and refine it each time the user has a training session. This allows the definition of a command through the identification of patterns the operator organically produces while training. For the purpose of this work, it was necessary that the operator was subjected to multiple sessions of command training to ensure its accuracy. After this stage, the operator was able to formulate *right* and *left* commands, and establish a *neutral* one (stationary state). For more information check [12].

B. Experimental Setup

This work falls within the context of a real-world use case and validated as such. In this environment, the drone used was the *crazyflie 2.1* quadcopter, developed by the *Bitcraze* company [13]. Measuring 92 mm of width, 92 mm of height and 29 mm of depth, this drone model is as lightweight as 27 g and holds on air for about 7 minutes. *Crazyflie 2.1.* is a suitable model for validating this work, because it is easier to control upon unexpected behaviors derived from hardware or communication failures, without having significant impact on their surroundings.

The second part of this experimental setup is the zone for flight demonstrations. As explained above, a drone is a

critical system and can result in unpredictable behaviors. To validate the proposed solution, a dedicated physical area was assembled, called the *arena* (Figure 3). The *arena* is a four-meter indoor area, gathered by net for safety precautions and a position system, similar to a GPS. The *loco positioning system* [14] was developed by the *Bitcraze* company and aims for locating the drone in the 3-dimensional space. For this purpose, *anchors* are placed on each vertex of the *arena*, serving as reference guides.

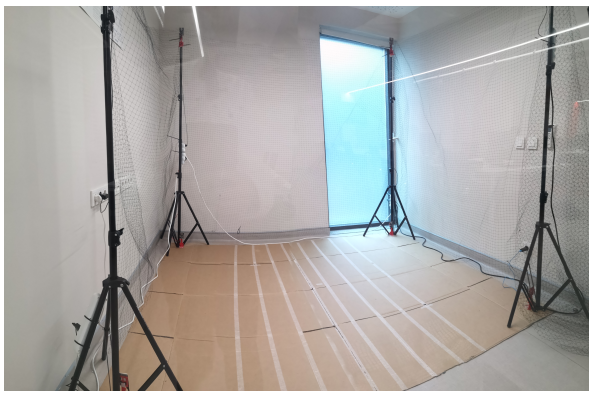


Figure 3. Drone Arena.

IV. Implementation

As mentioned in Section I-B, the core goal of this work is the development of a software platform capable of: (1) acquiring data from a BCI, (2) pre-process it, (3) identify the current emotional state of a drone operator and (4) decide whether he is in a suitable emotional condition to send commands. Even though brain-activity is the primary source of data for identifying the cognitive emotional level, a real-time visual input is also taken into account as a way of validating or invalidating this result. Nonetheless, as a decision is computed, it is equally crucial to establish a communication channel with the drone in order to forward all necessary information. This section will detail the implementation of such system and all necessary components that function together for these purposes.

As illustrated by Figure 4, this system is composed by four components: (1) the digital twin, (2) the visual classifier/component, (3) the decision making component and (4) the ROS2 component.

A. Digital Twin

The digital twin is a virtual representation of the operator and its goal is to classify, in real-time, the operator's emotional state. It relies on data collected by the BCI to build a cognitive profile, adapted to the operator. It is the core component of the proposed solution and provides decisive information to ascertain the destination of the command. The remaining components that follow are designated to support the digital twin and add complementary information for the decision.

According to Section III, this work comprehends that the digital twin, by means of machine learning techniques, will be built in an iterative manner following the tasks of: data acquisition, data preparation, modeling and evaluation. After data acquisition, procedures that follow are executed offline. The goal is to produce a digital twin, or a prediction model, that is going to be built under a set of training and testing tasks. Training requires data to be pre-processed and *cleaned* so that a machine learning algorithm can learn from it and find patterns on behaviors that correspond to certain emotions. This will result in an intelligent twin, which will be submitted to a testing session to validate whether the predictions match the correct emotion.

1) Data Acquisition: In a first approach, data acquisition was performed based on the analysis of available data streams from the BCI and fetched by a subscription procedure, which allows recording in real-time. For this purpose, three data streams, recorded by the *Emotiv Epoc+* headset, are collected: (1) the band power, i.e., power of the EEG data according to the sensor and frequency band; (2) the motion, based on the built-in gyroscope of the headset and (3) the facial expressions, recorded from facial muscle motions. The system communicates with the cortex API [15] to send requests for these data subscriptions and receive JSON responses with the resulting data streams as well as the classified commands, for the time period of the subscription.

2) Data Preparation: Since each request and response are unique to each stream, records are written to different files as separate datasets matching their type of data stream. In order to manipulate data on further tasks and, as needed by algorithms, these datasets need to be merged together. The newly collected data is integrated according to the nearest point in time of each observation (according with the *timestamp* feature), resulting into a single dataset. One particularity of this joint transaction is that it fills missing values of whatever feature with the closest value. This is useful due to different frequencies of the subscriptions and the asynchronous timestamps, which may not match exactly.

Columns with unique values are eliminated from this dataset, as well as features that do not add any value for the resolution of the problem (i.e., the *timestamp* feature).

Another strategy to improve the quality of the dataset is to perform feature engineering. This method transforms the current features and/or includes domain knowledge to increase their value and impact when solving the problem at hand. For this purpose, *one-hot-encoding* was performed on motion related-features (categorical data). This process consists on selecting each categorical column and transform it into a binary value. Motion related-features are transformed into binary columns, representing each type, through an one hot encoder. In addition, two features were added to the dataset: *arousal* and *valence* values, that are computed according to certain values of band power (according to [16] and equations (1) and (2)).

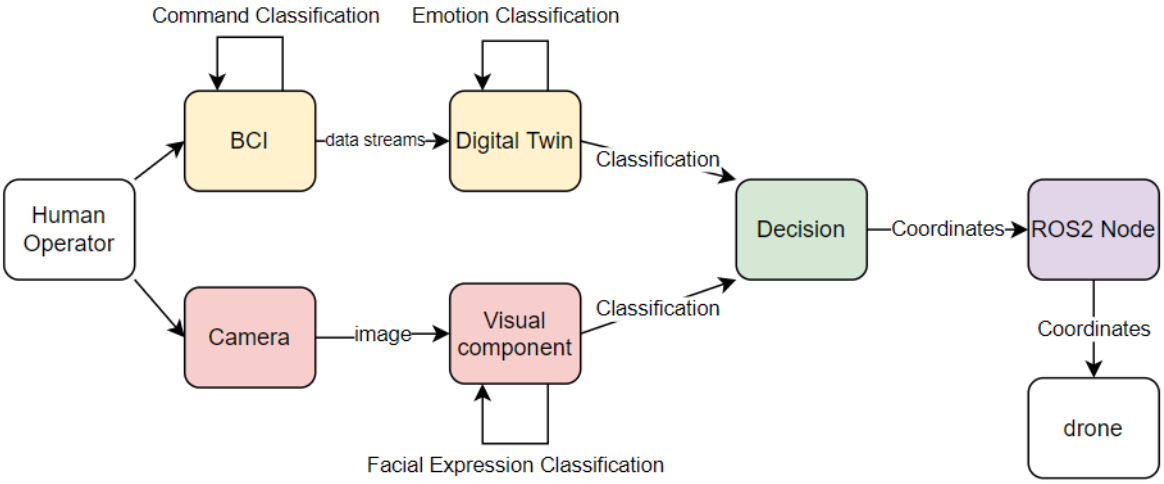


Figure 4. System architecture.

$$arousal = \frac{(F3/betaL + F4/betaL)}{(F3/alpha + F4/alpha)} \quad (1)$$

$$valence = \frac{F4/alpha}{F4/betaL} - \frac{F3/alpha}{F3/betaL} \quad (2)$$

3) Modeling and Evaluation: For the classification of the operator's emotional states, a set of classes were selected to represent positive and negative states. The positive classes are *calm* and *focused*, representing a stable cognitive state to send commands to the drone, as opposed to the negative classes (i.e., *stressed* and *distracted*) that detail an unstable cognitive state and, therefore, unacceptable state to send commands. In this work, the same operator simulated all the four emotions, at multiple days, in sessions of 8 seconds, reproducing a balanced dataset of about 19 600 observations per emotion (78 400 total). In this work, 70% of the data was split for training the algorithms and 30% for testing and evaluated 8 classifiers in four performance metrics (Table I).

TABLE I
EVALUATION OF ALGORITHMS

Algorithms	Performance Metrics			
	Accuracy	*Precision*	*Recall*	*f1-Score*
Decision Tree	0.995	0.995	0.995	0.995
k-Nearest Neighbors	0.997	0.997	0.997	0.997
LDA	0.911	0.916	0.911	0.912
Naive Bayes	0.614	0.645	0.614	0.617
Random Forest	0.999	0.999	0.999	0.999
SVM (linear kernel)	0.994	0.994	0.994	0.994
SVM (rbf kernel)	0.888	0.923	0.888	0.894
Neural Networks	0.948	0.949	0.948	0.948

As presented in Table I, Random Forest outperforms the remaining algorithms and is chosen for the training and modeling of the digital twin. In addition, the resulting confusion matrices are further analysed. These matrices display the true and false positives and negatives as way of assessing the level of confusion between classes from the machine learning algorithms. This work focuses primarily on a secure platform to send commands only when the operator is in a suitable condition to do so. Nevertheless, it also aims for minimizing the impact on the drone upon inaccurate classifications from the digital twin. Taking this into consideration, a confusion matrix is useful to assess the proportion of observations that were incorrectly classified as *calm* or when the operator was indeed *distracted* or *stressed*.

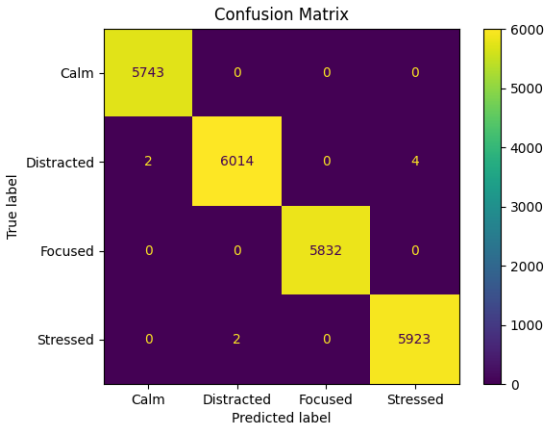


Figure 5. Random Forest confusion matrix (from [12]).

As illustrated by Figure 5, almost all observations were correctly identified by the Random Forest-based digital twin. Still, by analysing the true condition of the operator, in the negative spectrum, against the predicted labels, two observations that belonged to the *distracted* class were incorrectly classified as the *calm* class. From the testing set (30% of the whole dataset, about 23 520 observations), this error represents 0.009% of the sample. Although this demonstrates a flaw from the digital

twin, the probability of occurrence is minimum. For further details, check [12].

B. Visual Component

Regarding the visual component of the system, a camera captures the real-time image of the operator and uses a Convolutional Neural Network based-prediction model [17] from an open-source project [17], modeled and trained with the FER-2013 emotion dataset, to classify the visual expressions of the operator as a set of emotions. This component can output: *positive* emotions as *happy* and *neutral* and *negative* emotions as *angry, disgust, fear, sad and surprise*.

C. Decision Component

While the *EmotivBCI* application classifies the reproduced commands from the operator, the digital twin receives information from the headset and classifies the cognitive state of the operator. The visual component receives the image from the camera and classifies the facial expressions. This process results in three input variables for the decision component. This decision module will decide whether the operator is stable by mentally and visually evaluating his state. Only *positive* emotions detected on both components will allow the operator to send the command.

Considering the confidence percentage of the command and the digital twin upon the classification, the decision module, in case of an overall positive emotion detection, will compute the drone coordinates accordingly and send them through a ROS2 node. These coordinates are based on the distance to be travelled by the drone.

$$d = (c * i) * e \tag{3}$$

Equation (3) showcases the computation of the distance, where c is the confidence of the *EmotivBCI* classifier upon the identified command, e is the confidence of the digital twin upon the classified cognitive emotion and i is a safety increment with value 0.25 (meters). This means that classification errors from the prediction models might have an impact on the drone but is not decisive that it will be catastrophic due to filtering on the computation of coordinates.

D. ROS 2 Communication Node

For the connection between the system and the drone, a ROS2 client-server architecture is created between what is called the *base station*, meaning the server machine that manages the drone, and the client node (Figure 6). Nodes are software units that are responsible for the execution of some task. Here, there are two nodes, the client and the server. For these nodes to communicate, service schemes are created specifically to portray each task. The client instantiates the required scheme with the desired values and sends the request message to the other node. In this case, the operations available for the drone are: takeoff, relative motions and landing. These actions are translated into service schemes that are used by the solution, through the client node, with the desired values. In this context, the client node is implemented as a gateway of the decision module, sending a request with the coordinates. The server receives the coordinates and forwards them to the drone in real-time. Code listing 1 is an example of service that depicts the relative motion operation that matches the mental command formulated by the operator.

```
1 float32 x
2 float32 y
3 float32 z
4 float32 yaw
5 float32 duration
6 ---
7 int8 ret
```

Listing 1. *GoTo* service scheme

V. Results and Discussion

In this section, are presented the results and their discussion. This validation is divided into three parts: (V-A) where each emotion is experienced isolated, (V-B) a free environment with unexpected occurrence of emotions and (V-C) ROS2 communication validation with two drones to ensure a message can reach both. It is expected that, after the operator training session and digital twin training, the system is capable of detecting multiple emotional states of the operator in real-time and handle the drone accordingly.

A. Isolated Emotion Validation

To evaluate the different impacts of the solution, functionalities were split in a multi-level manner that go from the lowest experiment to the highest level (solution as a whole) to emphasize its value on securing a stable control environment for the drone. These experiments are:

- The ***baseline test***, defining the current state of drone control without the support of emotion recognition;
- The ***level 1 test***, representing the implementation of the core functionality which is the cognitive digital twin and the ROS2 client node;
- The ***level 2 test***, the cognitive digital twin with the addition of the computation of coordinates according to the prediction's confidence and the ROS2 client node;
- The ***full test***, having all the above functionalities and the support of the visual emotion recognition.

With the exception of the *baseline test*, which gives no importance to the mental state of the subject, each test covers the four mental states (*focused, calm, distracted and stressed*) individually, each one with sessions of 8 seconds. The subject had to be put under the same conditions in which he used to simulate the four emotions on the training phase.

One particularity is that levels 1 and 2 differentiate only in the computation of coordinates, which is a more visible advancement while operating the drone rather than an improvement of accuracy of the system. In this context, as a first validation session, the goal is to compute the accuracy

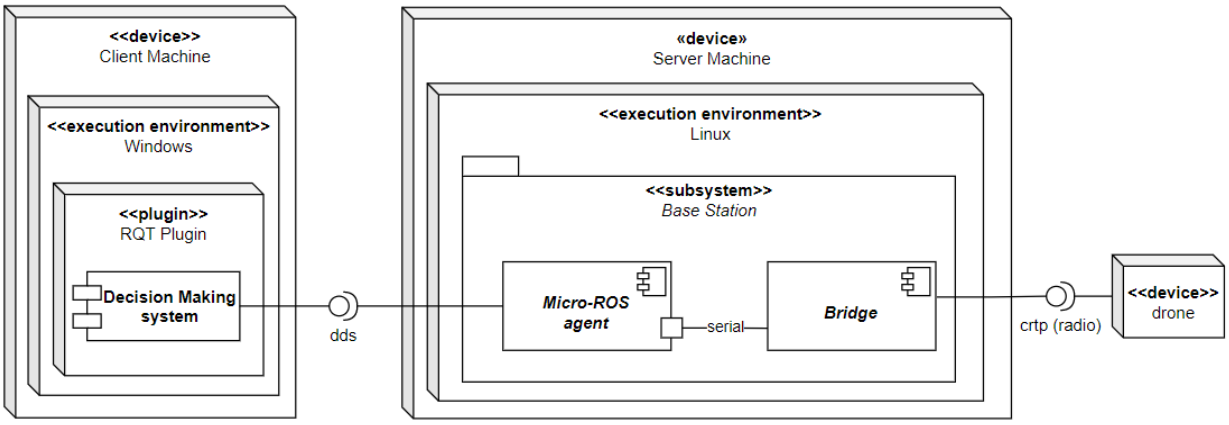


Figure 6. ROS 2 System architecture (from [12]).

of the digital twin when classifying the operator's mental emotions (experimental levels are seen as different temporal spaces). The second approach showcased in this section, is the analysis of observations of the negative spectrum. Here, the experimental levels are more significant to discuss the value of the functionalities of the system.

Given the environment set-up described in Section V, the number of observations per emotion and per experiment, for the same subject that trained the commands in the *EmotivBCI* application, are described in Table II.

TABLE II
NUMBER OF OBSERVATIONS PER EMOTION

Emotions	Group of Test		
	Level 1 Test	*Level 2 Test*	*Full Test*
Calm	142	120	85
Focused	94	101	90
Distracted	124	135	104
Stressed	134	120	112

From the number of observations, it was computed the success rate, or accuracy, for each emotion and per experiment (Figure 7). This metric is calculated by dividing the number of correctly classified observations by the total amount of observations. For the *calm* state, the highest accuracy of the digital twin was 87.5%, for the *focused* state a 98.8%, for the *distracted* state a 93.5% and for the *stressed* state a 100%, as described by the round values on Figure 7.

Even with a high average of success rate for detecting the subject's mental states, the most accurately classified emotion was the *stressed* state. The difference between them can be due to the distinct way the model is trained in this segment, which involves more physical movement to denote agitation, rather than a low on motion condition on the remaining ones.

Lower success rate depicted on *level 1* for the *focused* state can be explained by the different background noise and movement between the training and test phase. This caused the subject to deviate his attention, explaining the occurrences of *distracted* classifications during this period. In the next test

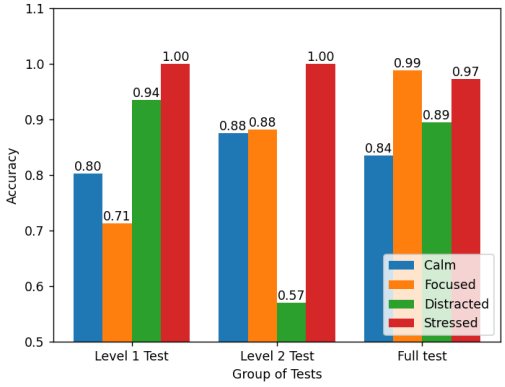


Figure 7. Accuracy bar chart (from [12]).

levels, this value is no lower that 80%, which is explained by the calmer environment. As opposed to this situation, the lower success rate on *level 2* for the *distracted* state classification can be explained by the lower amount interference or other diversions derived from background movement, which led to short occurrences of focus by the subject.

Regarding the classification of the negative emotional spectrum (*distracted* and *stressed* states), Tables III and IV give some insight about the number of sent commands under an incorrect classification.

TABLE III
DISTRACTED EMOTION RECOGNITION

Positive Detections	Group of Test		
	Level 1 Test	*Level 2 Test*	*Full Test*
Total number	6	11	10
Nº of neutral commands	4	7	6
Nº of sent commands	2	4	1
BCI positive, visual negative	N/A	N/A	3

As registered in Table III, at *level 1* were detected 6 positive states, 2 of them sent; at *level 2* were detected 11 positive emotions, 4 were sent and at the *full test*, 10 positive emotions

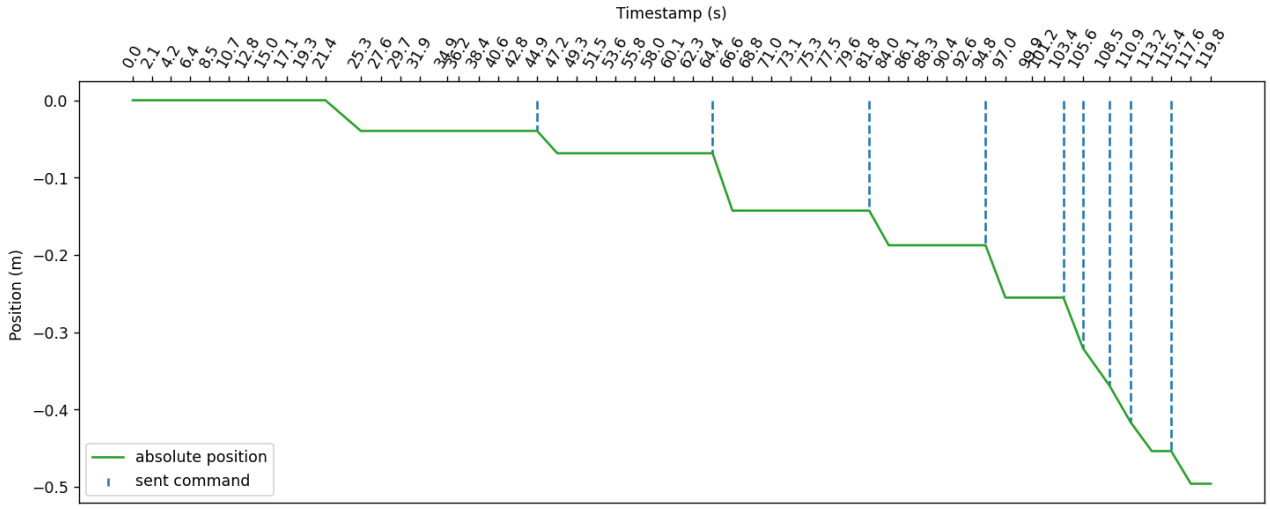


Figure 8. Real-time mission with distracting external events with a single drone (from [12]).

were detected, 1 command was sent to the drone and 3 were prevented due to the detection of a negative emotion by the visual component.

TABLE IV
STRESSED EMOTION RECOGNITION

Positive Detections	Group of Test		
	Level 1 Test	*Level 2 Test*	*Full Test*
Total number	0	0	2
Nº of neutral commands	0	0	1
Nº of sent commands	0	0	0
BCI positive, visual negative	N/A	N/A	1

As registered in Table IV, at the *full test* were detected 2 positive emotions and none were sent to the drone. One of them was a neutral command and the other was associated with a negative visual emotion, detected by the visual emotion component.

Since the training of mental commands is a task that requires time to practice and refine, it is challenging to reproduce a command at a live setting and in an equivalent environment the subject trained. Even with a digital twin inaccurate classification, most commands detected by the BCI are neutral ones, which have no impact on the trajectory of the drone. However, the command classifier can incorrectly output a *right* or *left* commands and these can potentially be sent to the drones. With the extra layer of the visual component, these unique situations are assessed by it and some of those errors are prevented. At a mission environment, where the operator needs to follow a sequence of commands, if there is a cancellation of a certain command, the operator will observe it and has enough time to reproduce the needed operation.

Considering that this is a 4-class classification problem, there is a probability of 25% that a baseline classifier correctly categorizes the subject emotion state and, in the *baseline test* characterized by the lack of machine learning, all commands are sent to the drones, regardless of the operator's emotional state, which could only be beneficial if the subject has perfect cognitive condition at all times.

B. Free Mission Flight

Even though previous section already validates the goal of this work, an isolated environment, where the operator simulates the four emotions, is a controlled scenario. In a free environment, realistically, is it common to occur unexpected events and different reactions from the operator. In this section, it was conducted a 2-minute validation session were the operator was able to send whatever command. For the purpose of validating the detection of mood swings, for instance between the *focused* and *stressed* states, two alarms were set to trigger at specific timestamps (27.6 and one minute and twenty eight seconds) after the experiment initiated.

Figure 8 illustrates the distance travelled by the drone, referencing the absolute position on the y axis (only *right* commands were sent). During this experiment, the operator was able to focus on the drone and send multiple commands, each one with a different distances. Even though the operator was consistently focused, at the trigger of the first alarm, the system detected a *distracted* mental state and a *surprised* visual state. The operator continued to be either *stressed* or *distracted* afterwards but was able to refocus on the drone and send new commands. At the sound of the second alarm, the system did not detect immediately a mental reaction but identified the physical reaction as *fear*. This validates the value and robustness of the solution when handling incorrect classifications by including a second classifier, the visual component, to break the command cycle. For more information regarding this experiment, check [12].

C. ROS2 Swarm Management Validation

Although a drone is a fundamental piece for the execution of complex tasks, a swarm of drones aims to optimize resource

allocation to more efficiently perform these high-risk tasks. This section details a third experiment composed to include two drones to receive and execute the same operations. The following code listings demonstrate the communication of two client nodes with the server node.

```
[INFO] [1623235179.701766700] [minimal_service]:
    Take off incoming request
 height: 0.500000
 duration: 2.000000
 response: 1

[INFO] [1623235179.704947100] [minimal_service]:
    Take off incoming request
 height: 0.500000
 duration: 2.000000
 response: 1
```

Listing 2. *Server node take-off* message.

```
[INFO] [1623235180.412227300] [drone1]: Sending
    information to server: height: 0.500000
 duration: 2.000000
 response: 1

[INFO] [1623235180.413246332] [drone2]: Sending
    information to server: height: 0.500000
 duration: 2.000000
 response: 1
```

Listing 3. *Client nodes take-off* messages.

```
[INFO] [1623235924.239647700] [minimal_service]: Go
    to incoming request
 x: 0.000000 y: 0.041117 z: 0.000000 yaw: 0.000000
     duration: 1.000000

[INFO] [1623235924.244461100] [minimal_service]: Go
    to incoming request
 x: 0.000000 y: 0.041117 z: 0.000000 yaw: 0.000000
     duration: 1.000000
```

Listing 4. *Server node go to* message.

```
[INFO] [1623235925.104762600] [drone1]: Sending
    information to server: x coordinate: 0.000000
 y coordinate: 0.041117
 z coordinate: 0.000000
 rotation: 0.000000
 duration: 1.000000
 response: 1

[INFO] [1623235925.106783200] [drone2]: Sending
    information to server: x coordinate: 0.000000
 y coordinate: 0.041117
 z coordinate: 0.000000
 rotation: 0.000000
 duration: 1.000000
 response: 1
```

Listing 5. *Client nodes go to* messages.

```
[INFO] [1623231818.296960000] [minimal_service]:
    Land incoming request
 height: 0.000000
 duration: 2.000000
 response: 1

[INFO] [1623231818.407826700] [minimal_service]:
    Land incoming request
 height: 0.000000
 duration: 2.000000
 response: 1
```

Listing 6. *Server node land* message.

```
[INFO] [1623235180.412227300] [drone1]: Sending
    information to server: height: 0.000000
 duration: 2.000000
 response: 1

[INFO] [1623235180.413246332] [drone2]: Sending
    information to server: height: 0.000000
 duration: 2.000000
 response: 1
```

Listing 7. *Client nodes land* messages.

As expected, the inclusion of an additional client node (one more drone) does not disrupt the functionality of the overall system and it is equally possible to send requests and receive messages from the same server client in parallel with the execution of other client nodes. This experiment demonstrates that is possible to work with a swarm of drones with ROS2 without significant structural efforts.

VI. Conclusion

In this work, it was analysed EEG data captured by the BCI *Emotiv Epoc+* of a drone operator and, using machine learning techniques, we were able to build a digital twin of the operator capable of predicting his emotional state and decide whether the commands should be sent to the *crazyflie* quadcopter. The classification of the emotional state not only is supported by EEG data but also by a visual component that analyses the facial expressions. In addition, the communication between the system and the drone is done through a ROS2 client node. Multiple machine learning algorithms were validated and Random Forest was the best fitted and therefore used for training the digital twin. Considering the research question pointed in Section I-A, results showed that the digital twin can accurately discriminate the operator's emotional states at a live setting and the combination of classification models improved the reliability of the system to decide upon the broadcasting of the reproduced commands.

While uniquely relying on the cognitive classifier, the digital twin, even with an overall satisfactory performance, allows inaccuracies to happen unexpectedly while the operator is not at his best state of mind. Including the visual component minimizes the impact of these situations. As part of our human condition, emotional reactions are often accomplished by mental and physical responses. While the digital twin could

produce an incorrect classification, the visual component will likely detect a negative emotion and the command will not be forwarded to the drone. A potential obstacle to the control of the drone could be the fact that the accumulative complexity of be mentally prepared to send commands, formulate a command with success and be physically stable could disrupt the execution of the command or sequence of command unnecessarily. From what could be a quick set of commands to perform a simple operation, it could demonstrate to be the opposite and even leading to the frustration of the operator. Regarding this issue, the visual component best detects the *happy* and *neutral* states. Without a proper physical reaction, this component will not detect other states and, therefore, these disruptions will most likely not happen.

Additionally, the implementation of a ROS2 framework in this work has proven to be crucial for the satisfactory functionality of the system as it allows the management and communication of information for one drone without explicit hardware/firmware constraints. This work also validates the ease of adding more drones (client nodes) for the execution of tasks of higher demand.

VII. Future Work

As future work, we aim for collecting more data to train the digital twin regarding the four emotional states and ensure it keeps improving its real-time detection.

Due to *Covid-19*, it was not possible to validate this solution with a wider range of subjects. Since a digital twin is adapted to each subject, we plan on creating more profiles of people with different demographics.

Emotiv Epoc+ was the target BCI for this work; thus, we would like to experiment this platform with other commercialized and, perhaps, open-source devices such as the *OpenBCI* headset.

References

[1] D. Ramos, G. Gonçalves, and R. Faria, "Digital Twin for Drone Control through a Brain-Machine Interface", *INTELLI 2021, The Tenth International Conference on Intelligent Systems and Applications*, 2021.

[2] M. A. Soto and M. Funk, "Look, a guidance drone! Assessing the Social Acceptability of Companion Drones for Blind Travelers in Public Spaces", *20th International ACM SIGACCESS Conference on Computers and Accessibility (ASSETS '18)*, Association for Computing Machinery, New York, NY, USA, pp. 417–419, 2018.

[3] M. Hinchey and L. Coyle, "Evolving Critical Systems: a Research Agenda for Computer-Based Systems", *2010 17th IEEE International Conference and Workshops on Engineering of Computer Based Systems*, pp. 430-435, 2010.

[4] A. Kubler, "The history of bci: From a vision for the future to real support for person hood in people with locked-in syndrome", *Neuroethics*, vol. 13, no. 2, pp. 163–180, 2020.

[5] R. M. Vishwanath, S. K. Saksena, and S. Omkar, "A real-time control approach for unmanned aerial vehicles using brain-computer interface", *CoRR*, vol. abs/1809.00346, 2018.

[6] M. Grieves, "Origins of the digital twin concept", Available online: https://www.researchgate.net/publication/307509727_Origins_of_the_Digital_Twin_Concept, 2016.

[7] H. Y. Jeon, C. Justin, and D. Mavris, "Improving prediction capability of quadcopter through digital twin", in *AIAA Scitech 2019 Forum*, pp. 1365, 2019.

[8] R. Ramirez and Z. Vamvakousis, "Detecting emotion from eeg signals using theemotive epoc device", *Brain Informatics*, pp. 175–184, Berlin, Heidelberg, 2012.

[9] R. Vishwanath, S. Saksena, and S. Omkar. "A real-time control approach for unmanned aerial vehicles using brain-computer interface", CoRR, abs/1809.00346, 2018.

[10] C. Manna, "An intro to the crisp-dm methodology". Available at https://medium.com/@chrismanna/an-intro-to-the-crisp-dm-methodology-c58cbe0371a3, Jun. 2019. Accessed on 21.03.2021.

[11] EMOTIV, "Emotiv epoc 14-channel wireless eeg headset". Available athttps://www.emotiv.com/epoc/, Sep. 2020. Accessed on 13.07.2021.

[12] D. Ramos, "Digital Twin for Drone Control using a Brain-Computer Interface" (Unpublished master's thesis). Faculty of Engineering of the University of Porto, 2021.

[13] Crazyflie 2.1, "Crazyflie 2.1". Available at https://www.bitcraze.io/products/crazyflie-2-1/, Sep. 2020. Accessed on 7.09.2021.

[14] Loco Positioning System, "Loco Positioning System". Available at https://www.bitcraze.io/documentation/system/positioning/loco-positioning-system/, Sep. 2020. Accessed on 7.09.2021.

[15] Cortex API, "Getting Started". Available at https://emotiv.gitbook.io/cortex-api/, Jun. 2021. Accessed on 7.09.2021.

[16] R. Ramirez and Z. Vamvakousis, "Detecting emotion from eeg signals using the emotive epoc device", in *Brain Informatics*, vol. 7670, pp. 175–184, 2012.

[17] U. Gogate, A. Parate, S. Sah, and S. Narayanan, "Real Time Emotion Recognition and Gender Classification", *2020 International Conference on Smart Innovations in Design, Environment, Management, Planning and Computing (ICSIDEMPC)*, pp. 138-143, doi: 10.1109/ICSIDEMPC49020.2020.9299633, 2020.

Advanced e-Learning by Inducing Shared Intentionality:

Foundation of Coherent Intelligence for Grounds of e-Curriculum

Igor Val Danilov
Academic Center for Coherent Intelligence
ACCI
Rome, Italy
e-mail: igor_val.danilov@acci.center

Sandra Mihailova
Rīga Stradiņš University
RSU
Riga, Latvia
e-mail: sandra.mihailova@rsu.lv

Abstract— **How is initial social knowledge acquired? The primary data entry (PDE) problem provides an understanding of the modality of social interaction between organisms disabled to communicate. The paper proposes the Model of Coherent Intelligence (MCI) and its neural foundation. The MCI shows how interpersonal dynamics shape shared intentionality in intimately related individuals. This hypothesis postulates two ideas: (1) cognition begins from a separation of sensory cues: Long Term Potentiation (LTP) can only be induced in neurons of particular Modality Specific (M-S) gateways (not all)–selective induction promotes selective sensitivity to the chaos of stimuli. (2) Neurons can learn Spike Timing Dependent Plasticity (STDP) by repeating the timing code of other organisms' mature neurons to modulate certain synaptic strength, which triggers either LTP or Long Term Depression. The paper suggests that shared intentionality in humans is the evolution outcome. Social animals demonstrate the quality of goal-directed coherence. The paper defines it as the ability of organisms to select only one stimulus for the entire group instantly. The manuscript shows the candidate for triggering the mechanism of goal-directed coherence and shared intentionality. The protein molecules contribute to animals' interaction ability, from essential motility organs in simple organisms (by presenting in receptors) to neural circuit assembly regulation and STDP in humans (by presenting in neurons). The study shows a direction for developing e-learning through stimulating learners' shared intentionality. An actual application of this approach is e-curriculum for children from 2 years of age.**

Keywords-coherent intelligence; goal-directed coherence; shared intentionality; e-learning; embodied cognition

I. INTRODUCTION

This article is an extension of the conference presentation "New findings in education: primary data entry in shaping intentionality and cognition" [1]. The academic knowledge on the study of mind historically and conceptually has settled three main approaches within cognitive science: cognitivism, connectionism, and embodied dynamicism [2]. Many theories of mind combine all three approaches, where they co-exist in various hybrid forms. The more interesting of them are the Embodied dynamic system [2], the theory of innate intersubjectivity and innate foundations of neonatal imitation [3], the theory of natural pedagogy [4], and the theory of sensitivities and expectations [5]. All these theories are plausible; the current paper observes different views to engage a gap in knowledge.

According to Thompson [2], cognitivism (the metaphor is the mind as digital computer) and connectionism (the mind as neuronal network), in different ways, appeal to the same computational principle of cognition. This principle based upon processing a signal within neuronal networks. This computational principle certainly requires the primary data entry as a necessary initial condition to launch processing. None algorithm and/or a sequence of instructions may perform the computation of any process without corresponding to the specific situation inputs, that should substitute variables and parameters of the formulas. The algorithm remains just a set of mathematical variables without this input. This argument may mean the necessity to input an initial set of social phenomena of the specific community to trigger this system – the Primary Data Entry (PDE) problem [6].

According to embodied cognitivists, the mind is an autonomous system by its self-organizing and self-controlling dynamics, which does not have inputs and outputs in the usual sense, and determines the cognitive domain in which it operates [2][7][8]. This approach is grounded on the dynamical hypothesis [9]. However, this interpretation of a dynamic system is not accurate [1]. Why does the dynamic system need PDE:

Argument A. According this approach embodied features of cognition deeply depend upon characteristics of the physical body. If the agent's beyond-the-brain body plays a significant causal role, then the primary data yet makes sense [1].

Argument B. In mathematics, a dynamic systems model is a set of evolution equations. It means that entering primary data is required. The dynamic system may not begin its life cycle without introducing initial conditions corresponding to specific situation inputs and parameters [1].

Argument C. The dynamical system hypothesis [9] has not claimed the lack of initial conditions. Dynamicists track primary data less than dynamic changes inside. However, it does not mean that primary data do not exist and do not necessary [1].

Given these above arguments, the PDE problem must be considered in the onset of cognition. The embodied dynamic

system approach tends to solve the above-noted gap by introducing the notion of dynamically embodied information [2]. Although, to introduce this concept, it is necessary to explain the categorization of reality through intentionality. According to embodied cognition approach, symbols encode the local topological properties of neuronal maps [2], a dynamic action pattern. The sensorimotor motor network yields pairing of the binary cue stimulus with the particular symbol saved in the structures and processes that embody meanings. 'Representational "vehicles" are temporally extended patterns of activity that can crisscross the brain-body-world boundaries, and the meanings or contents they embody are brought forth or enacted in the context of the system's structural coupling with its environment [2, p.36].' This idea requires introducing the nature of intentionality. In a multi-stimuli environment, the stimulus-consequence pair is unpredictable due to the many irrelevant stimuli claiming to be associated with the embodied dynamic information randomly. The bond of stimulus-consequence pair of a social phenomenon in the sensorimotor network requires categorizing reality by the nervous system before applying the innate reflex about this social phenomenon to a specific case. Therefore, dynamically embodied information can be useful if intentionality is already in place. However, the embodied dynamic system introduces intentionality without a biological and / or physical basis. The theory of natural pedagogy [4], and the theory of sensitivities and expectations [5], as well as many others may not solve the problem of PDE [9].

According to Trevarthen and Delafield-Butt [11], primary consciousness develops in embryogenesis and is the first operative in early fetal life. 'Consciousness as "acting with knowledge" requires a nervous system that regulates prospective perception in intentional engagement with the world [11, p. 22]'. In the first trimester, patterns of sensory regulation of movements of the fetus' body and limbs gain affective evaluation and sensitivity for sounds and rhythms of other human presence [11]. It means that the pure nervous system should already possess intentionality as well as initial knowledge about social reality: human sounds and rhythms also yield meanings. Even if fetuses can hear different sounds and feel rhythms from outside of the womb, this does not mean that they alone (independently) can process their meanings.

Searle et al. [12] argued that intentionality is the mental power of minds to represent or symbolize things, properties, and states of affairs. According to Crane [13], mental states or events or processes which have objects in this sense are traditionally called 'intentional,' and 'intentionality' is, for this reason, the general term for this defining characteristic of thought. The meaning of directed action implies the purpose of the action, which first requires the categorization of reality. It is a dichotomy of what happens first. Current knowledge does not solve it.

Tomasello [14], through the study on ontogenesis and phylogenesis, introduced the hypothesis of gradually increasing social bond development in children referred to time slices: (1) emotion sharing from the birth, (2) joint intentionality from the nine-month revolution, (3) collective intentionality at around three years of age, (4) reason and responsibility. Tomasello [14] introduced the beginning of cognition through the newborns' basic motive force of sharing intentionality. However, the mechanism of such emotion coordination is not clear because it is grounded on emotion sharing [14]. Whether or not protoconversations imply understanding emotional states. Many researchers, including the authors, believe that the hypothesis about the universality of emotional expressions is formed by limited experimental methods, since other research designs show the opposite outcome [15]-[19]. There is no evidence of a genetic mechanism that can link meaning in mind with certain social reality to apply an appropriate emotional pattern to a specific situation. Even if one assumes that the hypothesis of universal emotional expressions proves innate emotional patterns together with their meanings; even if newborns may alone recognize the basic facial expressions of caregivers and the specific situation to apply them; but in this case, newborns do not have time for such a "training course", because they demonstrate their achievements already in the first hours of life [20]. If there is no innate mechanism, then, apparently, emotional contagion can occur between individuals without their awareness [10][20][22]; it can happen even without awareness of the emotional stimuli existence [22]. Section II discusses the hyperscanning studies' outcomes, showing brain-to-brain synchronization. Section III presents the hypothesis of the neuronal foundation of shared intentionality. Section IV discusses the physical ground of goal-directed coherence–a forerunner of shared intentionality. Section V elaborates all findings.

II. Problem: How Does Social Interaction Encourage Cognition

Brain-to-brain relationships shape the mind during moment-to-moment interactions [23]. The dichotomy of newborns' succeed in beginning knowing and their communicative disability challenges our knowledge on social interaction modalities [20]. We believe that understanding the problem of the intentionality emergence in an organism at the beginning can explain the problem of PDE and the onset of consciousness. This knowledge can contribute to the study of cognition because obviously if and as soon as this implicit modality occurs it continues the whole rest of life. We believe that the caregivers' intentionality forms the intentionality in newborns. Fetuses and newborns are not able to behave intentionally on their own due to the lack of meaningful (informative) sensory interaction at the beginning [6][20][24]. We predict an implicit modality of social interaction that provides shared intentionality at the beginning. Cooperation in a group enhances intentionality, providing categorization.

According to Valencia and Froese [23], their review of studies based on EEG- and fNIRS hyperscanning methodologies shows evidence of inter-brain synchronization in the fastest frequency bands, supporting the possibility of extended consciousness. Among hyperscanning studies, we have chosen 4 studies conducted without explicit interaction between subjects. These studies compared differences of brain-to-brain synchronization in subjects when participants solved tasks together as confronting to the condition in which: (i) the subjects solved them individually [25][26]; (ii) the same task when

interacting with a machine [27]; (iii) the individuals from another team solved the same problem [28]. These studies declared an exclusion of sensory interaction between subjects. However, it should be noted that the subjects of all these studies knew about social encounters during the experiments. Therefore, instead of mental collaboration their results may simply mean an increase of brain activity due to similar emotional arousal in participants stimulated by the social encounter.

The near-infrared spectroscopy study (non-hyperscanning) on asleep newborns shows an increase of the neural response to a familiar (English language) versus unfamiliar language (Tagalog, a Filipino language) spoken by strangers in both conditions [29]. The language stimuli (the identical low-pass filtered sentences) were played through two speakers approximately 1.5 m from the infants' head. According to May et al. [29], these findings show that the newborn's neural processing of language is influenced by early language experience due to neonate brain responds to familiar versus unfamiliar language. To our mind, this outcome may lead to evidence of another inference. This experiment was not a hyperscanning technique. However, subjects were in pairs with their caregivers. Neonates classified these sound stimuli without the ability to perceive them. Sleeping newborns' brains reacted to sound stimuli that their sensing could not provide due to their brains' sensory isolation to meaningless and unfamiliar sounds. Sleepers seem to enter a standby mode, allowing them to balance the monitoring of their surroundings with sensory isolation [30]. Sleepers are sensitive to the semantic content of an auditory stream [30]-[32] and amplify relevant, meaningful stimuli [30][32]. The sleeping brain retains some residual information processing capacity, which, however, does not form enduring memories [33]. Neonates are not able to understand even Mother's speech although her sound is familiar. Given all these, any speech for neonates is meaningless, and asleep newborns may not be sensitive to the sounds even their native tongue (the language spoken by the mother during pregnancy) in experiments when these sounds were pronounced by outsiders. However, they were sensitive to them. Sleeping newborns' brains reacted to sound stimuli that their sensing could not provide due to their brains' sensory isolation to meaningless and unfamiliar sounds. We believe that this outcome may mean the implicit modality of newborns' interaction with caregivers since any other explanation of this outcome is excluded.

Recent hyperscanning shows an increase of coordinated neuronal activities in subjects during collective efforts without communication via sensory cues [34]. What are the neurobiological grounds of coordinated neuronal activities?

III. Foundation of Coherent Intelligence

A. Experiments on Problem-Solving in Groups

Recent research of 24 online experiments presented that unprimed participants show a more significant accuracy level when they complete the thought task simultaneously with confederates who are primed with the correct answer; if they were emotionally stimulated and completed the tasks without communication [10]. Primary groups [35] show empirical evidence of a more significant accuracy in problem-solving in the coherent intelligence state. In specific, we conducted 13 experiments in dyads (116 subjects) with P-value < 0,001 (probability-value in null hypothesis significance testing), and 7 experiments in primary group adults (41 subjects) with the P-value < 0,002. Experiments with 43 secondary group subjects (unfamiliar adults, M=20) show the effect only with the task of unfamiliar language translation. Non-semantic tasks–with synthetic language and two-color round symbols–did not stimulate the effect in 2 experiments with 207 secondary group subjects (unfamiliar). These results are consistent with research Danilov et al. [36] [37].

B. The Model of Coherent Intelligence

According to Danilov and Mihailova [24], a supranormal environmental case–e.g., first hours after birth–stimulates supranormal sensation in dyads. This can push the inherited mechanism of social entrainment of infants to the rhythm of the mother. Both the supranormal sensation and social entrainment may stimulate the common emotional arousal. The latter is increased by the ongoing supranormal sensation and the occurring rhythm of arbitrary movements of the infant. The continuing supranormal sensation and ever-increasing arousal of the infant and the mother along with the rhythm of the infant's unintentional movements stimulate early imitation and emotional contagion. The problem is how the infant capture and reproduce the kinematic of movements.

The MCI proposes that common emotional arousal together with the identical rhythm create coherent mental processes in dyads–Coherent Intelligence (Figure 1). At Sensorimotor Stage (by Piaget, or Stage 3 of the Model of Hierarchical Complexity MHC [38]), organisms do not maintain bilateral communication. According to Danilov and Mihailova [24], individuals are able to interact by distinguishing perceptual signals of identical modality by their significance. This ability can contribute to ostensive cues. After all, this meaningless interaction modifies into communication when individuals imbue perceptual impulses with mutually implied meanings, cascading their signals in response to the history of relations between them [24].

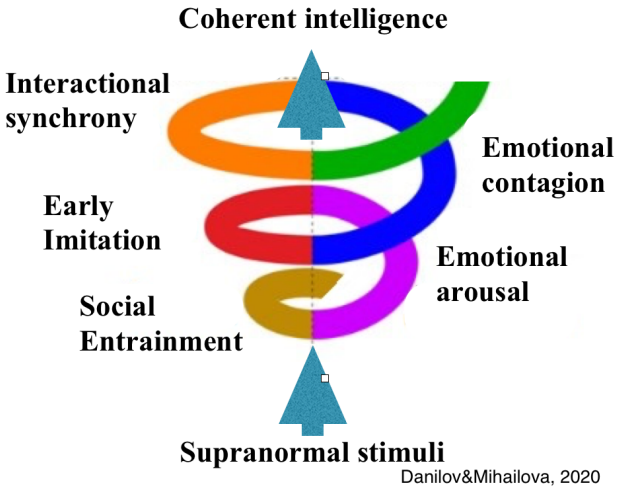


Figure 1. Interpersonal dynamics in Model of Coherent Intelligence[24]

C. Neuronal Foundation of the MCI

It seems consistent to say that intention arises from conscious intentionality, or intentionality shapes intention if intentionality becomes conscious. Specific brain regions may be engaged in shared sensory/cognitive processes irrespective of the feedback's valence and in encoding the subjective relevance of the feedback [39][40].

Outside areas involved in this processing, additional brain areas are specifically engaged according to the particular communicative modality [41]. According to Tettamanti et al. [42], Intention Processing Network (IPN) involves the medial prefrontal cortex, precuneus, bilateral posterior superior temporal sulcus, and temporoparietal junctions. Depending on different social interaction modalities, the IPN is complemented by activation of additional brain areas, reflecting different Modality-Specific (M-S) input gateways to the IPN [42]. The M-S gateways mediate the structural and semantic decoding of stimuli and provide M-S information [42]. Sensory inputs of a specific modality can activate the precise association of certain sensorimotor networks with specific brain emotion circuits [42].

We believe that this emotion-motion dynamics can involve the particular cognitive process of a high order. When two or more organisms are in common emotional arousal and simultaneously in the interactional synchrony, then these two different experiences may meet each other in high-order cognitive processing. Emotional arousal can trigger evolutionarily old brain circuits, which interact with high-order cognitive and linguistic processing [43]. It seems uncontroversial to say that infants' pure nervous system may experience emotions, but only primitive ones related to survival, such as those associated with hunger and pain. However, newborns cannot express emotions themselves appropriately to a specific social case on their own, even though they possess inherited neuronal patterns of primitive emotional impressions. They also cannot understand the expression of others' emotions (as is discussed above). They are only capable of experiencing primitive emotions, not correctly expressing them independently. Research on insects–organisms in stage 3 of MHC [38] like human newborns–assumes that they also experience emotions [44]. Researchers argued that agitated honeybees exhibit pessimistic cognitive biases: 'Whether animals experience human-like emotions is controversial and of immense societal concern. The next reason is that animals cannot provide subjective reports of how they feel, emotional state can only be inferred using physiological, cognitive, and behavioral measures. In humans, negative feelings are reliably correlated with pessimistic cognitive biases, defined as the increased expectation of bad outcomes. Recently, mammals and birds with poor welfare have also been found to display pessimistic-like decision making, but cognitive biases have not thus far been explored in invertebrates [44].'

In parallel, interactional synchrony stimulates a sensorimotor network engaging neural networks responsible for communicative intention processing (including high-order cognitive and linguistic processing)[41]. Neural networks of emotional excitation and the sensorimotor networks are separately connected to many different M-S gateways. Meanwhile their coherence intersects in certain M-S gateways of each organism depending on (i) pattern of neural circuit engaged through emotional excitation and (ii) pattern of the sensorimotor network [41].

We propose a rough hypothesis of how Long-Term Potentiation (LTP) can be induced only in particular M-S gateways, retaining information about the certain received stimulus [1]. Different areas of the brain exhibit different forms of LTP, their types depend on a number of factors, such as age and the neuron's anatomic location. However, the common processes are the same for all. The simple nature of Hebbian learning, based only on the coincidence of pre- and post-synaptic activity, LTP is persistent, lasting from several minutes to many months, and it is this persistence that separates LTP from other forms of synaptic plasticity [45]. Spike-timing-dependent plasticity (STDP)–that involves the pairing of pre- and postsynaptic action potentials (APs)–causes a variation of LTP or Long-Term Depression (LTD) [46]. The timing between pre- and postsynaptic APs modulates synaptic strength, triggering LTP or LTD [46]. The sign and magnitude of the change in synaptic strength depend on the relative timing between spikes of two connected neurons (the pre- and postsynaptic neuron) [46]. The structural organization of excitatory inputs supporting STDP remains unknown [46]. Even though the ensemble of emotion-motion integrated networks weakly stimulates the intersected neurons in their junction with M-S gateways. If all M-S gateways also simultaneously receive weak stimulation from the receptors (due to the chaos of stimuli received by the pure nervous system), then this multi-signal contributes to LTP in the neurons of particular M-S gateway at the junction of this emotion-motion ensemble due to the effect of the synaptic cooperativity, because of the following. LTP can be induced either by strong tetanic stimulation of a single pathway to a synapse, or cooperatively via the weaker stimulation of many. Neurons from the gateways in the connections of these networks receive cooperative stimulation [1]. Induction of cooperativity can ensure LTP.

According to Tazerart et al. [46], the synaptic cooperativity of only two neighboring synaptic inputs onto spines in the basal dendrites of L5 pyramidal neurons extends the pre–post timing window that can trigger potentiation. The engaged M-S gateways retain a certain stimulus, while other M-S gateways (also of the same sensory modality) remain depressed without keeping information of other stimuli. Therefore, specific M-S gateways are sensitive, and all these organisms respond to specific sensory modalities. Figure 2 shows a very rough schematic picture of this process. The induced emotion and sensorimotor networks (they are red in the picture) together activate certain M-S gateway even with weak stimulation of sensory input. The different colors of M-S gateways refer to different sensory modalities. At this point, the analysis encounters the ground of the PDE problem of how immature neurons learn the timing code to modulate certain synaptic strength, which triggers either LTP or LTD. Because the structural organization of excitatory inputs supporting STDP remains unknown [46].

The study of the PDE problem leads to the analysis of the axiomatic foundations of Psychology, Sociology, and Neuroscience–the basic notions that form these sciences–from the perspectives of the actual scientific paradigm.

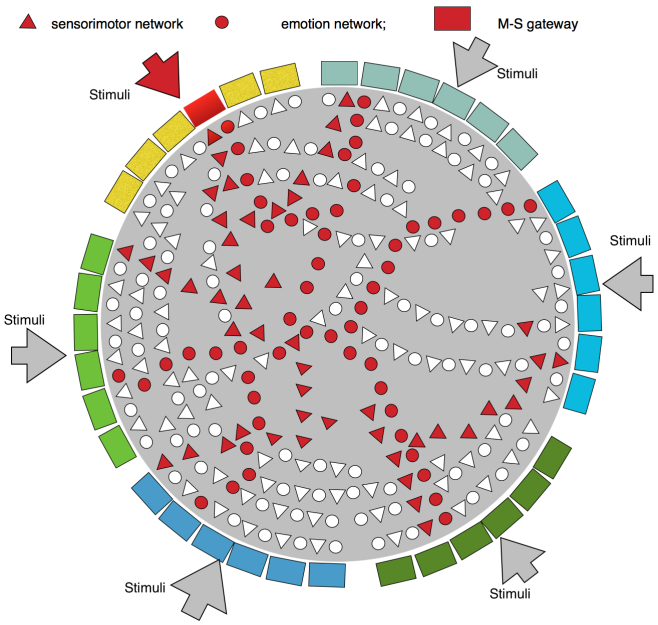


Figure 2. The very schematic picture of M-S Gateway Activation [1].

The question of "how can the blank mind begin to learn from social interaction" is reduced to "how immature neurons learn the timing code to modulate certain synaptic strength, that triggers either LTP or LTD"[1]. The sign and magnitude of the change in synaptic strength depend on the relative timing between spikes of two connected neurons (the pre- and postsynaptic neuron) [1]. How can neurons of an immature organism (even a newborn) learn the structural organization of excitatory inputs that support STDP? The further arguments show why we believe that the entanglement state of neurons can contribute to simultaneous LTP in neurons.

The daily routine develops neuronal patterns of primitive emotions and sensorimotor neuronal patterns in infants. Their everyday coherency with the social world forms various integrated neuronal patterns of different emotions from the existing ensemble of emotion scripts in their community. We believe that caregivers contribute to the formation of emotion scripts and, consequently, shaping of specific neuronal patterns in infants [1]. Obviously, adults experienced intentionality before their coherent mental process began with newborns. Life experience taught them particular emotion scripts and defined their precise motion kinematics, that formed more elaborated sensorimotor patterns. In routine cooperation with newborns, a caregiver enters in interactional synchrony with a newborn, under the influence of supranormal stimuli, being in social entrainment. Therefore, the similar M-S gateways are excited in the dyad. Meanwhile, the adult's current intentionality has already triggered a particular network that includes current emotion patterns and sensorimotor patterns. Part of it corresponds to a primitive complex emotion-sensorimotor network in the newborn with similar M-S gateways. This newborn's primitive network is less developed, although it is similar to the part of the adult's integrated complex network. It can be assumed that the neurons in the connections of different excited emotion patterns and sensorimotor patterns of both neural systems receive similar stimulation due to interactional synchrony and emotional arousal of organisms. The neurons of mature organism receive LTP, being induced cooperatively via many stimulations. If simultaneously, neurons of mature and immature organisms are also induced by a single harmonic oscillator, these neurons of specific M-S gateways go into the coordinated state [47]. According to Danilov [46], due to STDP, the precise order and timing of pre- and postsynaptic action potentials trigger LTP or LTD regulating the connection strengths between neurons. These M-S gateways of the neonate begin to react on the high-frequency sequence of stimulation in the same way as those M-S gateways in the caregiver and receive LTP [47]. The relationships of these neurons teach the specific M-S gateway of the newborn to react to the specific stimulation, supporting STDP in responding to a particular emotional and sensorimotor neural pattern [47]. In such a manner neurons of mature organisms train newborns' neurons, being in coherence; because the adult and infant neurons behave as a single unit [1].

Therefore, specific M-S gateways are sensitive in dyads, and these organisms equally respond to specific sensory modalities [1]. The induction of t-LTP and t-LTD in single spines follows a bidirectional Hebbian STDP learning rule [46]. Hebbian theory claims that an increase in synaptic efficacy arises from the learning process. The PDE problem in the chaos of irrelevant stimuli requires a teaching mechanism from the beginning. The coordinated state of neurons is a possible option of their cooperative activity, how infants' neurons learn spike-timing-dependent plasticity [1].

Emotion sharing indicates implicit modality of social interaction. The coordinated state of neurons in the certain M-S gateways is a possible option of how infants' neurons learn STDP [1]. This involvement of similar networks and the sensibility of the certain M-S gateways lasts as long as is necessary to teach the immature nervous system. The coordinated state of these neurons ensures their immediate response to the specific stimulus, regardless of the spatial division of organisms. Therefore, specific M-S gateways are sensitive, and these organisms equally respond to specific sensory modalities. This is an old evolutionary mechanism because interaction without sensory cues should be the primary and archetypal modality in biological systems beginning from bacteria. Section IV shows why we believe so.

IV. GOAL-DIRECTED COHERENCE

Knowledge about a coordinated state of neurons from different organisms can complement the set of social interaction modalities. The manuscript shows two possible options for involving cells into a coordinated state: entanglement of entire cells or their coordinated activity due to an agent (chemical element or compound). In the latter option, the entangled state of the agent leads cells to coordinated cooperation. Three candidates can pretend to become such an agent: they are the atom of hydrogen [48], the Posner molecule [49], and protein. According to Danilov and Mihailova [50], the idea of the protein as the agent

seems to be more plausible from the two other above mentioned.

A. Protein as the Agent

Proteins become biologically active only when folded into a three-dimensional structure of amino acids formed into particular, highly complex configurations. A relatively small protein of only 100 amino acids folds into its functional shape in nanoseconds [51]. This high rate of choosing through a vast number of different possible configurations (at least 10 to the power of 100 options!)[50] in forming the precise symmetric configuration corresponds only to quantum mechanisms [50]-[55]. Consequently, it is possible to assume that quantum mechanisms can have a widespread connection mode between protein molecules in nature.

In bacteria, protein takes part in photoreceptors [56]. The review on the photoreceptors in plant-associated bacteria identified common traits such as protein-protein interaction during signal transduction [57]. Even more, these light-sensitive proteins seem to control infectivity and virulence to a level that generates not too much harm to the plant host [57], interacting with plants cells. These facts are relevant to quantum relationships between amino acids of bacteria proteins and between amino acids of bacteria and those from the plant [50]. In humans, protein Reelin is essential for hippocampal integrity and synaptic plasticity. According to Faini et al. [58], this molecule contributes to neural circuit assembly, refinement, and function, as well as axonal guidance, synaptogenesis, and dendritic spine formation. Thus, the entanglement between protein molecules of neurons from different organisms could become a candidate for their connection that leads to neurons' coordinated activity in different organisms [50].

B. Goal-Directed Coherence in Biological Systems

Social animals demonstrate the quality of goal-directed coherence. This quality is defined by the ability of organisms to select only one stimulus for the entire group instantly. It seems that its main features can be defined as follows: a) bypassing sensing (insensitivity to sensory perception), b) independence from a distance, c) instantaneousness in time. There are a few arguments why proposes this definition:

a) Bypassing sensing. Bacteria are the smallest free-living (self-replicating) organisms. They were among the first life forms to appear on Earth. The phenomenon of community phototaxis in bacteria is the concerted movement of an entire colony of cells towards or away from the light source which mechanism is still undefined [59]. According to Danilov [60], by themselves, photoreceptors of bacteria cannot measure the field gradient and show the direction of movement. However, community phototaxis involves direct sensing of the position of a light source [60]. The ability of the single cell to independently determine the direction of movement contradicts the simplicity of its internal structure-organization [60].

According to Zirbes et al. [61], earthworms demonstrate the cooperative ability to choose the same direction of movement as their conspecifics. According to Danilov [60], this earthworms' ability shows the incongruence of the complexity required communication and a set of sensory modalities in earthworms because their simple nervous system and sensory receptors make communication impossible. Therefore, the only possible explanation for earthworms' cooperative achievements with a communicative disability is that these organisms can together separate sensory stimuli according to their significance [60]. For this reason, they need to share the significance of the specific cue.

b) Independence from a distance. Individual ants perform large distance foraging excursions up to 1200 m, from which they return on a direct, shortcut way to their nests and can infer this ground distance when walking over hills [62]. Individual ants successfully perform this task without direct visibility of the goal and changes in the environment (wind, light, etc.). They seem to choose the certain path strategy from different options through interaction with their nest-mates on a case-by-case basis [60]. There is a dichotomy between the perceptual ability of organisms and environmental conditions–such as inappropriate distance, landscape, and weather–that are needed for successful interaction through sensory cues [60].

c) Instantaneousness in time. According to Danilov [60], flocks of birds, schools of fish, and hordes of insects also show the phenotype of the synchronization, performing the cooperative movements. Moreover, these social organisms can instantly change the direction of movement and shape fantastic collective forms in motion at a high rate. These collective movements intend the joint ability of organisms to choose the same direction of movement that required simultaneous information exchange. The high-speed rate of changing movement direction can mean the interaction modality that proceeds instantaneously, bypassing sensory receptors. Furthermore, all biological systems demonstrate instant interaction if they successfully perform the two previous features of bypassing sensing and interacting, overcoming insuperable distance. Indeed, in a multi-stimuli environment, when many organisms are required to instantly choose one stimulus from a dozen irrelevant ones, only simultaneous information exchange (instant) provides the correct solution for choosing the one correct stimulus for the whole group.

Many other biological systems also show two or three features of the quality of goal-directed coherence. Such a quality is presented in the mother-fetus dyads in humans. The intriguing facts of fetal facial expressions, voice recognition, and twin fetuses co-movement highlight the vital role of interaction in mother-fetus dyads in cognitive development [60]. Common sense assumes that this is the way it should be, while biology emphasizes separating these organisms. Fetuses own their autonomous nervous system. There is no communication between these organisms–the mother can not explain to her fetus social meanings using sensory cues. Even the mother's voice is a social cue, unintelligible for her fetus. Indeed, the meaning "mother" begins from self-awareness, from understanding the meanings of "self" and "other" and then understanding many other essential needs–just hearing a sound every minute does not lead to understanding its meaning. Even an undeveloped nervous system of fetuses casts doubt on the possibility of communication, and even more so the absence of abstract thinking at this stage of development. Nevertheless, the above facts show that, during gestation, some social learning succeeds, despite the absence of communication.

According to Darwinism, if inherited valuable qualities appear in every generation, the useful variations become so noticeable that the organism evolves into a new species over several generations. If the quality has been preserved over many generations of a phylogenetic ancestor, it manifests itself in one form or another in different species of its offspring. These arguments mean that quality preserved in simple organisms through many generations should manifest itself in one form or another in more developed animals.

Even in simple organisms, the sustainability of an organism's development in colonies (first of all, increased protection against predators and foraging efficiency) contributes to the propagation of the corresponding phenotypic features. If the quality of goal-directed coherence propagates in different species through an evolution development, this quality's single primary physical mechanism could exist. The entanglement between protein molecules from different organisms could become a candidate for triggering this physical mechanism. Because this molecule contributes to animals' interaction ability, beginning from essential motility organs in simple organisms (by presenting in receptors) to neural circuit assembly regulation and spike-timing-dependent plasticity in neurons of organisms with the nervous system.

Nevertheless, there are three candidates for this coherence mechanism of the organisms' cooperativity. Further research is needed to understand if this physical mechanism exists in all animals and because of what agent.

C. Findings in Physics for Goal-Directed Coherence

In physics, all matter with a temperature greater than absolute zero emits thermal radiation, consisting of electromagnetic fields. Coherence means a fixed relationship between the phase of waves of a single frequency and identical waveforms of two or more waves. Therefore, two neurons can become coherent in the case of features correspondence of their radiation. Quantum coherence appears from the interference of particles' quantum waves with each other.

According to the received view in physics, in short, coherence is converted into quantum entanglement. Streltsov et al. [63] argued that coherence in a system is converted into entanglement with another separate system. Any nonzero coherence in a system can be converted into an equal amount of entanglement between that system and another initially incoherent one [63]. From this perspective, a single harmonic oscillator can induce quantum entanglement in two or more particles of different systems if the properties of this electromagnetic field are such as it induces coherence of these particles.

Marletto et al. [64] argued that they found empirical evidence of entanglement between bacteria and the light (modeled as a single quantum harmonic oscillator). If so, this empirical data is probably the first evidence of quantum entanglement within a colony of one of the most ancient living organisms in nature. However, this result can also be regarded as a coordinated activity of the bacteria colony due to a single harmonic oscillator. That is, even though the experiments by Marletto et al. [64] have shown the entanglement of objects close to quantum scale size, according to the received view in physics, these cells are not the objects of quantum physics. Therefore, the conclusion can be threefold. First, as these authors argued, they induced the entanglement between bacteria. The generation of entanglement between increasingly macroscopic and disparate systems is an ongoing effort in quantum science [65]. Recent studies have shown that the behavior of objects 15 micrometers in size is consistent with the quantum world's laws, such as the phenomenon of quantum entanglement [66]. In comparison, a neuron's nucleus has a diameter of 3 to 18 micrometers, and a neuron has a size of 4 to 100 micrometers. Second, these objects can also be considered systems of atoms. While this is also entanglement however from this point of view, it can be defined as an entangled state between two or more quantum systems. For instance, recent studies showed that it is also possible. An entangled state was generated between a millimeter-sized dielectric membrane and an ensemble of 109 atoms [65]. Third, the experimenters observed the coordinated activity of bacteria due to the entangled state of an agent in these cells. The current article proposes that the amino acids from the protein molecules or the protein molecules themselves can become such an agent of the entanglement.

The article considers protein molecules in photoreceptors of bacteria (and neurons as well) as the agent of this coherence. From this perspective, the single harmonic oscillator can entangle the protein molecules (or amino acids from the protein molecules) from different bacteria receptors. The coordinated states of photoreceptors of different bacteria lead to coordination in their motility. From this point of view, during the experiment by Marletto et al. [64], a single harmonic oscillator with identical frequency and waveform electromagnetic field as in protein amino acids (or the whole protein molecules) induced entanglement of different cells' proteins promoting similar activity of bacteria. Considering the essential role of protein in simple organisms (in the activity of the photoreceptors) and organisms with the nervous system (in neural circuit assembly regulation and spike-timing-dependent plasticity in neurons), the quality of Goal-Directed Coherence might be the ground of biological systems and widely distributed in nature.

Moreover, the properties of quantum entangled systems promote coordinated activity in biological systems overcoming the long distance between organisms. Entanglement is an essential property of multipartite quantum systems, characterized by the inseparability of quantum states of objects regardless of their spatial separation [65]. A recent study tested quantum entanglement over great distances, sending entangled pairs of photons to three ground stations across China—each separated by more than 1200 kilometers. Yin et al. used the Micius satellite, which was launched last year and is equipped with a specialized quantum optical payload. They successfully demonstrated the satellite-based entanglement distribution to receiver stations separated by more than 1200 km [67]. Therefore, goal-directed coherence in biological systems (shared intentionality in humans) should also be possible in a long distance between organisms.

The study shows a possible direction for progress in e-learning by designing an advanced e-curriculum that can stimulate shared intentionality in students. We believe that the ideas mentioned above contribute to an advanced e-

learning curriculum for young children. Recent case studies (conducted online) with the educational task to children aged 12 to 33 months show coordinating actions in the absence of communication through sensory cues in the mother-child dyads that promoted numerosity in infants and toddlers in a short course at an age younger than others (before then peers do) [68] [69].

V. CONCLUSIONS

The current study discussed a possible foundation of Shared intentionality for stimulating Coherent Intelligence that can form grounds of advanced e-curriculum. The analysis of recent empirical data yields a hypothesis of beginning cognition–the Model of Coherent Intelligence and its neuronal foundation of how the pure nervous system distinguishes sensory stimuli. This hypothesis postulates two new ideas of the PDE basis: (1) cognition begins from a separation of sensory stimuli: LTP can only be induced in neurons of particular M-S gateways (not all)–selective induction promotes selective sensitivity to the chaos of irrelevant stimuli. (2) Neurons can learn STDP in social interaction by repeating the timing code of other organisms' mature neurons to modulate certain synaptic strength, which triggers either LTP or LTD [1]. We believe that the MCI shapes intentionality in intimately related individuals. Coherent Intelligence is the integration of M-S gateways of particular brain areas, which contributes to different organisms' sensibility to similar sensory inputs [1].

Recent hyperscanning shows an increase in coordinated neuronal activities in subjects during collective efforts in the absence of communication through sensory cues [34]. This finding may mean indirect evidence of the hypothesis about the Model of Coherent Intelligence and its neuronal foundation. In addition, a growing body of literature on increasing the efficiency of cooperative decision-making in groups without sensory cues between subjects [10][24][36] [37] also shows empirical indirect evidence of the MCI.

The paper suggested that this social quality of humans is the outcome of evolution development. Social animals demonstrate the quality of Goal-Directed Coherence. The paper defined this quality as the ability of organisms to instantly select only one stimulus for the entire group. It also argued its main features: a) bypassing sensing (insensitivity to sensory perception), b) independence from a distance, c) instantaneousness in time.

The manuscript showed the candidate for triggering this physical mechanism–entanglement between protein molecules from different organisms (or amino acids from the protein molecules). This molecule contributes to animals' interaction ability, from essential motility organs in simple organisms (by presenting in receptors) to neural circuit assembly regulation and spike-timing-dependent plasticity in organisms with a nervous system (by presenting in neurons). Nevertheless, the paper proposed three candidates for this coherence mechanism of the organisms' cooperativity. Further research is needed to understand if this physical mechanism exists in all animals and what kind of agent it is.

We believe that this approach may contribute to studying the mind and, specifically, understanding the appearance of intentionality. In addition, we believe that these findings may contribute to an advanced e-curriculum, specifically in teaching children from 2 years of age with communication disabilities. Further research can also examine whether the MCI can provide a contactless interaction of the computer with neuronal circuits, in which the computer would become a part of the extended mind. This approach provides a wide range of possibilities for developing advanced intelligence systems, in specific a human-computer interface.

AUTHORS CONTRIBUTION

Igor Val Danilov formulated the hypothesis and wrote the first draft of the manuscript. Igor Val Danilov and Sandra Mihailova improved the text over several iterations.

ACKNOWLEDGMENT

The authors express their gratitude to Iegor Reznikoff, Emeritus Professor of University of Paris X-Nanterre France, for commenting on this paper.

REFERENCES

1. I.Val Danilov and S. Mihailova, "New Findings in Education: Primary Data Entry in Shaping Intentionality and Cognition." The Thirteenth International Conference on Advanced Cognitive Technologies, and Applications, COGNITIVE 2021, April 18 – 22, 2021 in Porto, Portugal.
2. E. Thompson, Mind in Life: biology, phenomenology, and the sciences of mind. The Belknap press of Harvard University press. Cambridge, Massachusetts London, England. 1st Harvard University Press paperback edition. 2010.
3. J. Delafield-Butt and C. Trevarthen, Theories of the development of human communication. 2012. [Online]. Available from: https://strathprints.strath.ac.uk/39831/1/Delafield_Butt_Trevarthen_2012_Theories_of_the_development_of_human_communication_final_edit_060312.pdf [retrieved: March, 2021].
4. G. Csibra and G. Gergely, "Natural Pedagogy," Trends Cogn. Sci., vol. 13, pp. 148–153, 2009.
5. S. R. Waxman and E. M. Leddon, "Early Word-Learning and Conceptual Development". The Wiley-Blackwell Handbook of Childhood Cognitive Development. 2010. [Online] Available from: https://www.academia.edu/12821552/Early_Word-Learning_and_Conceptual_Development [retrieved: March, 2021].
6. I. Val Danilov, "Imitation or Early Imitation: Towards the Problem of Primary Data Entry." The Journal of Higher Education Theory and Practice (JHETP). 21(4), 2021. https://doi.org/10.33423/jhetp.v21i4.4222.
7. F. J. Varela, Principles of Biological Autonomy. ISBN-10:0135009502, ISBN-13:978-0135009505. 1979.
8. F. J. Varela and P. Bourgine, "Towards a practice of autonomous systems. In Towards a Practice of Autonomous Systems." The first European conference on Artificial Life, ed. F. J. Varela and P. Bourgine, pp. xi–xviii. Cambridge: MIT Press. 1992.
9. T. Van Gelder, "The dynamical hypothesis in cognitive science." Behavioral and Brain Sciences, vol. 21 (5), pp. 615-628. 1998. DOI: 10.1017/s0140525x98001733.
10. I. Val Danilov and S. Mihailova, "Knowledge Sharing through Social Interaction: Towards the Problem of Primary Data Entry." The 11th Eurasian Conference on Language & Social Sciences which is held in Gjakova University, Kosovo. p.226. 2021. [Online] Available from http://eclss.org/

publicationsfordoi/ abst11act8boo8k2021a.pdf. [retrieved: March, 2021].

11. J. Delafield-Butt and C. Trevarthen, Development of human consciousness. 2016. [Online]. Available from: https://strathprints.strath.ac.uk/id/eprint/57845. [retrieved: March, 2021].
12. J. R. Searle, W. S. Slusser, and M. Slusser, Intentionality: An Essay in the Philosophy of Mind. Cambridge University Press. 1983.
13. T. Crane, "Intentionalism." The Oxford Handbook to the Philosophy of Mind. Oxford: Oxford University Press. pp. 474–493. 2009.
14. M. Tomasello, Becoming human: A theory of ontogeny. Belknap Press of Harvard University Press. 2019. https://doi.org/10.4159/9780674988651.
15. R. E. Jack, "Culture and facial expressions of emotion". Visual Cognition vol. 21, pp. 1248–1286, 2013, https://doi.org/10.1080/13506285.2013.8353.
16. C. Crivelli and M. Gendron, In e Science of Facial Expression (eds José-Miguel Fernández-Dols & James A. Russell) Oxford University Press, 2017.
17. L. F. Barrett, R. Adolphs, S. Marsella, A. Martinez, and S. Pollak, "Emotional Expressions Reconsidered: Challenges to Inferring Emotion in Human Facial Movements." Psychological Science in the Public Interest, vol. 20, pp. 1–68. 2019. https:// doi.org/10.1177/1529100619832930.
18. K. Hoemann et al., "Context facilitates performance on a classic cross-cultural emotion perception task". Emotion. (Washington, DC). vol. 19, pp. 1292–1313. 2019.
19. M. Gendron, K. Hoemann, A. N. Crittenden, S. M. Mangola, G. A. Ruark, and L. F. Barret, "Emotion Perception in Hadza Hunter-Gatherers". Scientific reports, vol. 10, pp. 3867 , 2020. https://doi.org/10.1038/s41598-020-60257-2.
20. I. Val. Danilov, "Social Interaction in Knowledge Acquisition: Advanced Curriculum. Critical Review of Studies Relevant to Social Behavior of Infants". The Twelfth International Conference on Advanced Cognitive Technologies and Applications COGNITIVE2020. 2020.
21. J. Decety and P. L. Jackson, "The functional architecture of human empathy". Behavioral and Cognitive Neuroscience Reviews, vol. 3, pp. 71-100. 2004.
22. M. Tamietto et al., "Unseen facial and bodily expressions trigger fast emotional reactions". PNAS, vol. 106: pp. 17661-17666, 2009.
23. A. L. Valencia and T. Froese, "What binds us? Inter-brain neural synchronization and its implications for theories of human consciousness". Neuroscience of Consciousness, vol. 6(1), 2020. doi: 10.1093/nc/niaa010.
24. I. Val Danilov and S. Mihailova, "Emotions in e-Learning: The Review Promotes Advanced Curriculum by Studying Social Interaction". The International Conference on Lifelong Education and Leadership for All (ICLEL). pp.8-17. ERIC Number: ED606507. 2020. [Online] Available from: https://faf348ef-5904-4b29-9cf9-98b675786628.filesusr.com/ugd/d546b1_2d77ecc9e07f4b0fb2cccce23af201f7.pdf. [retrieved: March, 2021].
25. C. Szymanski et al., "Teams on the same wavelength perform better: Inter-brain phase synchronization constitutes a neural substrate for social facilitation". Neuroimage, vol. 15, pp. 425–436, 2017.
26. F. A. Fishburn et al., "Putting our heads together: interpersonal neural synchronization as a biological mechanism for shared intentionality". Soc Cogn Affect Neurosci, vol. 13, pp. 841–849, 2018.
27. Y. Hu et al., "Inter-brain synchrony and cooperation context in interactive decision making". Biol Psychol, vol. 133, pp. 54–62, 2018.
28. L. Astolfi et al., "Neuroelectrical hyperscanning measures simultaneous brain activity in humans". Brain Topogr, vol. 23, pp. 243–256, 2010.
29. L. May, K. Byers-Heinlein, J. Gervain, and J. F. Werker, "Language and the newborn brain: does prenatal language experience shape the neonate neural response to speech?" Front. Psychology, vol. 2, pp. 222, 2011, doi: 10.3389/fpsyg.2011.00222.
30. T. Andrillon and S. Kouider, "The vigilant sleeper: neural mechanisms of sensory (de)coupling during sleep." Current Opinion in Physiology, vol. 15, pp. 47–59, 2020.
31. A. Ibanez, V. Lopez, and C. Cornejo, "ERPs and contextual semantic discrimination: degrees of congruence in wakefulness and sleep." Brain Lang, vol. 98, pp. 264-275, 2006.
32. G. Legendre, T. Andrillon, M. Koroma, and S. Kouider, "Sleepers track informative speech in a multitalker environment." Nat Hum Behav, vol. 3, pp. 274-283, 2019.
33. R. Cox, I. Korjoukov, M. de Boer, and L. M. Talamini, "Sound Asleep: Processing and Retention of Slow Oscillation Phase-Targeted Stimuli." PLoS ONE, vol. 9(7), pp. e101567. 2014, https://doi.org/10.1371/journal.pone.0101567.
34. D.R Painter, J.J. Kim, A.I. Renton et al. 2021. Joint control of visually guided actions involves concordant increases in behavioural and neural coupling. Commun Biol 4: 816. https://doi.org/10.1038/s42003-021-02319-3
35. C. H. Cooley, Social organization: A study of the larger mind , (pp. 23-31). New York, NY: Charles Scribner's Sons, xvii, pp. 426, 1909.
36. I. Val Danilov, S. Mihailova, and V. Perepjolkina, "Unconscious social interaction, coherent intelligence in learning." The 12th annual conference. ICERI 2019. doi: 10.21125/iceri.2019.0606.
37. I. Val Danilov and S. Mihailova, "Intentionality vs Chaos: Brain Connectivity through Emotions and Cooperation Levels beyond Sensory Modalities." The Thirteenth International Conference on Advanced Cognitive Technologies, and Applications, COGNITIVE 2021, April 18 – 22, 2021 in Porto, Portugal.
38. M. L. Commons, "The fundamental issues with behavioral development." Behavioral Development Bulletin, vol. 21(1), pp. 1-12, 2016, http://dx.doi.org/10.1037/bdb0000022.
39. S. Oldham, C. Murawski, A. Fornito, G. Youssef, M. Yucel, and V. Lorenzetti, "The anticipation and outcome phases of reward and loss processing: A neuroimaging meta-analysis of the monetary incentive delay task." Human Brain Mapping, vol. 39, pp. 3398-3418, 2018.
40. D. Martinsa et al., "Mapping social reward and punishment processing in the human brain: A voxel-based meta-analysis of neuroimaging findings using the Social Incentive Delay task." 2021, bioRxiv preprint doi: https://doi.org/10.1101/2020.05.28.121475.
41. I. Enrici, M. Adenzato, S. Cappa, B. G. Bara, and M. Tettamanti, "Intention Processing in Communication: A Common Brain Network for Language and Gestures." Journal of Cognitive Neuroscience, volume 23, issue 9, pp. 2415-2431, 2011, https://doi.org/10.1162/jocn.2010.21594.
42. M. Tettamanti, M. M. Vaghi, B. G. Bara, S. F. Cappa, I. Enrici, and M. Adenzato, "Effective connectivity gateways to the Theory of Mind network in processing communicative intention." NeuroImage, vol. 155, pp. 169-176, 2017, doi: 10.1016/ j.neuroimage.2017.04.050.
43. L. Nummenmaa et al., "Emotional speech synchronizes brains across listeners and engages large-scale dynamic brain networks." NeuroImage, vol. 102, pp. 498-509, 2014.

44. M. Bateson et al., "Agitated honeybees exhibit pessimistic cognitive biases." Curr Biol., vol. 21;21(12), pp. 1070-1073, 2011, doi:10.1016/ j.cub.2011.05.017.
45. W. C. Abraham, "How long will long-term potentiation last?" Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences. vol. 358 (1432), pp. 735–744, 2003, doi:10.1098/rstb.2002.1222.
46. S. Tazerart et al., "A spike-timing-dependent plasticity rule for dendritic spines." Nat Commun vol. 11, pp. 4276, 2020, https://doi.org/10.1038/ s41467-020-17861-7.
47. I. Val Danilov, "Human-Centered Computing Based on Shared Intentionality in Human Cognition Model: A New Approach To Contactless Human-Computer System." 23rd International Conference on Artificial Intelligence, along with CSCE congress, 2021.
48. D. K. F. Meijer, I. Jerman, A. V. Melkikh, and V. I. Sbitnev, "Biophysics of Consciousness: A Scale-Invariant Acoustic Information Code of a Superfluid Quantum Space Guides the Mental Attribute of the Universe", along with "Rhythmic Oscillations in Proteins to Human Cognition", edited by Anirban Bandyopadhyay and Kanad Ray, Springer Nature Singapore Pte Ltd. 2021, https://doi.org/ 10.1007/978-981-15-7253-1.
49. M. P. A. Fisher, "Quantum cognition: The possibility of processing with nuclear spins in the brain." Annals of Physics, Volume 362, pp. 593-602, 2015. ISSN 0003-4916, https:// doi.org/10.1016/j.aop.2015.08.020.
50. I. Val Danilov and S. Mihailova, "Neuronal Coherence Agent for Shared Intentionality: Hypothesis of Neurobiological Processes during Social Interaction." OBM Neurobiology. In press, 2021.
51. L. Luo and J. Lu, "Temperature Dependence of Protein Folding Deduced from Quantum Transition." 2011 https:// arxiv.org/abs/1102.3748
52. M. Baiesi, E. Orlandini, F. Seno et al. "Sequence and structural patterns detected in entangled proteins reveal the importance of co-translational folding." Sci Rep 9, 8426, 2019. https:// doi.org/10.1038/s41598-019-44928-3
53. D. Banerjee, "Energy transfer and entanglement in an optically active solution of amino acids." Quantum Physics, arXiv: 1909.07795 .2019.
54. J. I. Sulkowska, "On folding of entangled proteins: knots, lassos, links and θ-curves." Current Opinion in Structural Biology, Volume 60, pp. 131-141, 2020. ISSN 0959-440X, https://doi.org/10.1016/j.sbi.2020.01.007.
55. S. H. Zangbar, T. Ghadiri, M.S. Vafaee et al. "A potential entanglement between the spinal cord and hippocampus: Theta rhythm correlates with neurogenesis deficiency following spinal cord injury in male rats." J Neurosci Res. 98: pp. 2451–2467, 2020. https://doi.org/10.1002/jnr.24719.
56. C. Storey, C. Y. Allan, P. R. Fisher, "Phototaxis: Microbial." Wiley Online Library, 2020, https://doi.org/ 10.1002/9780470015902.a0000399.pub3
57. A. Losi and W. Gärtner, "A light life together: photosensing in the plant microbiota" , Photochemical & Photobiological Sciences, 20: pp. 451–473, 2021. https://doi.org/10.1007/ s43630-021-00029-7.
58. G. Faini, F. Del Bene, S. Albadri, "Reelin functions beyond neuronal migration: from synaptogenesis to network activity modulation," Current Opinion in Neurobiology, Volume 66, pp. 135-143, 2021. ISSN 0959-4388, https://doi.org/10.1016/ j.conb.2020.10.009.
59. A. Wilde and C. W. Mullineaux, "Light-controlled motility in prokaryotes and the problem of directional light perception." FEMS Microbiology Reviews, fux045(41), pp. 900–922, 2017. doi: 10.1093/femsre/fux045.
60. I. Val Danilov, "Contactless Human-Computer Systems via Shared Intentionality: A Concept Design for the Next Generation of Smart Prosthetic Limbs", in: Arai K. (eds) Proceedings of the Future Technologies Conference (FTC) 2021, Volume 3. FTC 2021. Lecture Notes in Networks and Systems, vol 360. Springer, Cham. https://doi.org/ 10.1007/978-3-030-89912-7_59.
61. L. Zirbes, J.-L. Deneubourg, Y. Brostaux, E. Haubruge, "A New Case of Consensual Decision: Collective Movement in Earthworms." Wiley Online Library, 116(6), 2010. https:// doi.org/10.1111/j.1439-0310.2010.01768.x.
62. B. Ronacher, "Path integration in a three – dimensional world: the case of desert ants." Journal of Comparative Physiology A., 2020. https://doi.org/10.1007/s00359-020-01401-1.
63. A. Streltsov, U. Singh, H.S. Dhar et al. "Measuring Quantum Coherence with Entanglement." Phys. Rev. Lett. 115, 020403. 2015. doi:https://doi.org/10.1103/PhysRevLett. 115.020403 .
64. C. Marletto, D. M. Coles, T. Farrow, and V. Vedral, "Entanglement between living bacteria and quantized light witnessed by Rabi splitting." Journal of Physics Communications, vol. 2(10), 2018. http://iopscience.iop.org/ article/10.1088/2399-6528/aae224/meta.
65. R. A. Thomas et al., "Entanglement between distant macroscopic mechanical and spin systems." Nat. Phys., 2020. https://doi.org/10.1038/ s41567-020-1031-5.
66. M. Sillanpää and S. Hong, "Spooky quantum entanglement goes big in new experiments." [Online] Science News 25 April 2018. Available from: https:// www.sciencenews.org/article/ spooky-quantum-entanglement-goes-big-new-experiments. [retrieved: March, 2021].
67. J. Yin et al., "Satellite-based entanglement distribution over 1200 kilometers" [Online] Science: vol. 356, issue 6343, pp. 1140-1144, doi:10.1126/science.aan3211, 2017. [retrieved: March, 2021].
68. I. Val Danilov, S. Mihailova, and I. Reznikoff, "Frontiers in Cognition for Education: Coherent Intelligence in e-Learning for Beginners Aged 1 to 3 years." The 20th Int'l Conf on e-Learning, e-Business, Enterprise Information Systems, and e-Government, along with CSCE congress, 2021.
69. I. Val Danilov, S. Mihailova, and I. Reznikoff, "Shared Intentionality in Advanced Problem Based Learning: Deep Levels of Thinking in Coherent Intelligence", The 17th Int'l Conf on Frontiers in Education: Computer Science and Computer Engineering, Along with CSCE congress, 2021.

Linked Open Data in the GIOCOnDa LOD Platform

Lorenzo Sommaruga, Nadia Catenazzi, Davide Bertacco, Riccardo Mazza
Department of Innovative Technologies
University of Applied Sciences and Arts of Southern Switzerland (SUPSI)
CH-6962 Lugano, Switzerland
e-mail: lorenzo.sommaruga@supsi.ch
nadia.catenazzi@supsi.ch
davide.bertacco@supsi.ch
riccardo.mazza@supsi.ch

Abstract – **The GIOCOnDa LOD platform, developed in the context of a project funded by the EU programme Interreg, aims to make Linked Open Data (LOD) available in an easy and convenient way, without the need of any software programming. This paper illustrates the implemented platform that, starting from a common Open Data repository, can automate the production and publishing of Linked Open Data. The platform is configurable and extensible as it enables to define mapping configurations for new datasets hiding the complexity of the ontology in the mapping process. The paper also presents some examples of data consuming applications, using some GIOCOnDa LOD datasets.**

Keywords - Linked Open Data (LOD); GIOCOnDa; LOD publishing; LOD consuming; OntoPia.

I. INTRODUCTION

The work described in this paper, firstly introduced in [1], was developed in the context of the EU Interreg GIOCOnDa project ("Integrated and holistic management of the open data life cycle", March 2019 – April 2021), funded by the Interreg V-A Italy-Switzerland Programme, which aims to create value by developing information products based on the re-use of public Open Data [2]. The project focuses on data that are relevant for the touristic sector, coming from the Insubric area, a cross-border territory across Italy and Switzerland. They include data about museums, accommodation facilities, and environment. The main data sources currently used are: Regione Lombardia open data portal [3], and ARPA (Regional Agency for the Protection of the Environment) [4] for Italian data; Wikidata, Ticino Turismo [5], and OASI [6] for Swiss data.

One of the main results of the project is the creation of a platform for the publication of LOD (Linked Open Data) by public administrations. In the GIOCOnDa LOD platform, open data, coming from various sources and in different formats, are converted into homogeneous Linked Open Data, based on standard ontologies, and published together with their metadata. The platform enables conversion of existing 3* Open Data to 5* Open Data, according to the well-known 5-star deployment scheme [7]: data are formalized in RDF, identified by URI and linked to other datasets. For this purpose, a specific interlinking module was developed and integrated into the platform.

This work answers an emerging need for generic and usable tools to produce linked open data. While the value of linked open data is widely recognized in the literature [8], their publication is still challenging.

Different works and platforms have been developed to support this process. One of the first significant projects is Lucero and the resulting Tabloid toolkit, which aims to help institutions and developers to publish and consume linked data [9]. Another interesting work, which supports US open government data production and consumption, is the TWC LOGD portal [10]. Here a workflow for linked open data deployment is defined, consisting of different stages, where the conversion process is automated by using the csv2rdf4lod tool. A more recent initiative is represented by the Italian cultural heritage platform "dati.beniculturali.it", promoted by the Italian Ministry of Culture, which collects and publishes standardized and interoperable LOD heterogeneous datasets [11].

An extensive survey of methods, tools, and techniques for generating and publishing linked open data reveals that the proposed approaches to produce linked data are often specific to a use case and usually concern a specific domain, such as, media, library, finance, education, and healthcare. These approaches can hardly be adapted to other use-cases and domains [12].

The main innovative aspect of the approach adopted in the GIOCOnDa platform is that it is generic and supports different source types and formats. In addition, it does not require programming skills or a deep knowledge of the RDF and OWL formalisms.

This paper describes the publishing process from Open Data to Linked Open Data in the GIOCOnDa platform and presents some examples of applications that consume LOD produced using the platform. It is structured as follows: Section II presents the methodology adopted to publish LOD data; Section III focuses on the process of conversion of Open Data to LOD, one of the main steps of this methodology; Section IV presents the main functionalities of the GIOCOnDa LOD platform, in particular the LOD catalogue, and the input and output mapping, two processes that enable to convert data from different input data sources to the final LOD format; Section V explains how the interlinking module works and is integrated into the platform; Section VI presents possible use cases of

publishing a new dataset in the GIOCOnDa LOD platform; finally, Section VII proposes some examples of applications that exploit Linked Open Data to support users in their needs.

II. Methodology to Publish Linked Open Data

From a methodological point of view, a number of best practices, recommendations and guidelines have been defined. For example, Bauer and Kaltenböck [13] provide a step-by-step model, highlighting the most important issues that need to be considered in LOD publishing; W3C [14] presents best practices designed to facilitate LOD development and delivery; the "Agenzia per l'Italia Digitale" (AGID) [15] proposes a general methodological approach for the interoperable opening of public data through the LODs. This methodology basically consists of the following steps: selection of dataset, data cleaning, analysis and RDF modelling, enrichment, interlinking, validation, and publication.

The approach adopted in GIOCOnDa is in line with the above best practices and guidelines and, in particular, with those proposed by AGID. The selection of datasets was made on the basis of the results of a previous need analysis phase carried out with a number of stakeholders during the project. As a starting point, data about museums, accommodation facilities, and environment of the Insubric region are selected.

Concerning data cleaning, it is assumed that the selected datasets are already published as *clean* and *accurate* open data, where a quality check is already accomplished.

Once selected, datasets are deeply analysed to understand their structure; then appropriate ontologies and vocabularies are identified to model them.

In particular, the adopted ontologies are taken from the OntoPia network [16], also presented in [17]. They include, for instance, the Cultural-ON ontology for museums and the ACCO ontology for accommodations. In the GIOCOnDa LOD platform, data are imported from different sources and converted into the RDF format, according to these standard ontologies. The conversion process is detailed in the next section.

As additional steps, datasets are enriched with metadata and interlinked to other datasets. Metadata are added to the single datasets following the DCAT-AP standard. Interlinks to other datasets are created by identifying alignments and similarities between different datasets. For instance, a museum from the "Regione Lombardia" dataset can be declared "the same as" a museum described in Wikidata. The identification of interlinks is mainly carried out using the Silk software libraries [18], as explained in Section 5.

Finally, datasets are published using Openlink Virtuoso Universal Server [19], where they can be queried through a SPARQL endpoint.

III. The Conversion Process from Open Data to LOD in the GIOCOnDa Platform

The core of the system lies in the mapping functionality of heterogeneous data into Linked Open Data, according to standard ontologies.

This conversion is a complex process that depends on the initial format and on the final standard RDF format. From a literature study it emerges that the most frequently adopted approach is the implementation of ad-hoc middleware. For example, to convert a relational database to LOD, a typical solution is to use declarative languages, such as D2R [20] or R2RML [21] that require ontological and programming skills.

In the GIOCOnDa LOD platform, the complexity of the conversion process is simplified by defining a converter, facilitated by a graphical user interface that an expert can use to configure the conversion. This process can be explained through a simple example: we would like to convert two different datasets about museums into a common interoperable format. The first dataset concerns *Lombard museums* retrieved from the Regione Lombardia portal in JSON format by means of REST APIs [21]. The second is represented by *Tessin Canton museums* retrieved from Wikidata through SPARQL queries.

Figure 1 shows an excerpt from the Lombard museums visualized on the Regione Lombardia portal [22], while Figure 2 shows an example of a Swiss museum in Wikidata [23].

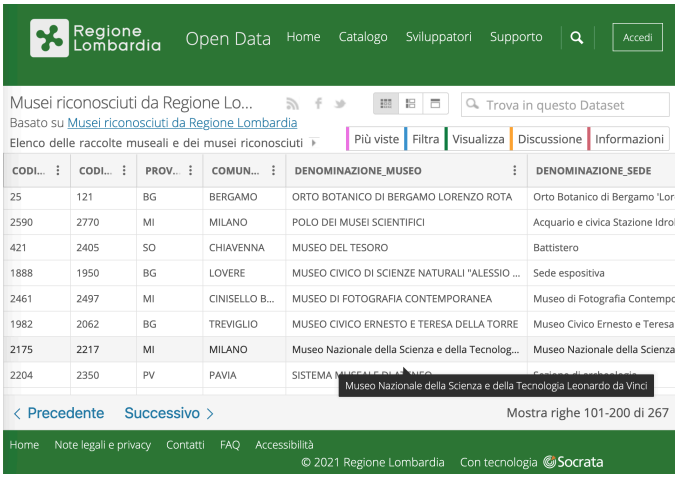


Figure 1. Lombard museums from the Regione Lombardia Open Data portal.

Figure 2. The Swiss *Vela* Museum in Wikidata.

To be able to configure the mapping from the original to LOD format, the structure of the two museum data sources has to be analysed by an expert and an appropriate ontology selected. In this phase it is important to find the most appropriate ontology to model the domain. Cultural-ON [24]

and its connected ontologies were chosen because they are representative of the museum domain and can be exploited to support transnational interoperability.

The next step consists of analysing the different descriptive fields of the museum datasets: for instance, each Lombard museum is described in terms of 79 fields, such as *denominazione museo* (name), *telefono* (telephone), *codice sede* (site code) as shown in Figure 1.

For each field, the objective is to find a match with the ontology classes and properties. For example, a museum could be represented as an instance of the *cis:Museum* class of the Cultural-ON ontology, where *cis* is the prefix of the ontology namespace; the *telephone* field can be mapped into a property of a *smapit:OnlineContactPoint* instance of the Social Media / Contact and Internet ontology [25].

Figures 3 and 4 present some details of the conversion result of a Lombard and a Tessin canton museum, respectively, into RDF Turtle, according to Cultural-ON and its connected ontologies. In particular, in the excerpts, the light blue border highlights the *hasSite* relation to the *Site* instance, and the orange one highlights the *hasOnlineContactPoint* relation to the *ContactPoint* instance with their respective properties.

```
museum:Museo_2175_sede_2217_Museo_…_Leonardo_da_Vinci
     a cis:CulturalInstituteOrSite, cis:Museum ;
     rdfs:label "Museo Nazionale … Leonardo da Vinci" ;
     cis:institutionalName "Museo … Leonardo da Vinci" ;
     cis:hasSite site:Sede_2217;
     smapit:hasOnlineContactPoint
                  contactPoint:Contatti_Museo_Leonardo_da_Vinci ;
...
site:Sede_2217
     a cis:Site, poiapit:PointOfInterest;
     rdfs:label "Museo Nazionale … Leonardo da Vinci";
     cis:siteAddress
       address:Indirizzo_della_Sede_Museo_scienza_Leonardo_da_Vinci;
     clvapit:hasGeometry geometry:geometry_Museo_Leonardo_da_Vinci .
...
contactPoint:Contatti_Museo_Leonardo_da_Vinci
      a smapit:OnlineContactPoint ;
      smapit:hasEmail email:email_museo_Leonardo_da_Vinci;
      smapit:hasTelephoneNumber phone:phone_museo_Leonardo_da_Vinci;
      smapit:hasTelephoneNumber fax:fax_museo_Leonardo_da_Vinci ;
      smapit:hasWebSite website:web_museo_Leonardo_da_Vinci .
```

Figure 3. Excerpt from the Lombard Museum of Science and Technology converted in RDF Turtle.

```
museum:Museo_Q3867651_Museo_Vincenzo_Vela
        a cis:CulturalInstituteOrSite, cis:Museum ;
        rdfs:label "Museo Vincenzo Vela" ;
        cis:institutionalName "Museo Vincenzo Vela" ;
        cis:hasSite site:Sede_Q3867651 ;
        smapit:hasOnlineContactPoint
                     contactPoint:Contatti_Museo_Vincenzo_Vela .
site:Sede_Q3867651
        a cis:Site, poiapit:PointOfInterest ;
        rdfs:label "Museo Vincenzo Vela ";
        cis:siteAddress address:Indirizzo_della_Sede_Museo_Vincenzo_Vela ;
        clvapit:hasGeometry geometry:geometry_Museo_Vincenzo_Vela .
...
contactPoint:Contatti_Museo_Vincenzo_Vela
      a smapit:OnlineContactPoint ;
      smapit:hasEmail email:email_museo_Vincenzo_Vela;
      smapit:hasTelephoneNumber phone:phone_museo_Vincenzo_Vela ;
      smapit:hasWebSite website:web_museo_Vincenzo_Vela .
```

Figure 4. Excerpt from The Swiss Vela museum converted in RFD Turtle.

It should be noted that the two museums, initially formalised in different ways on their original portal, are finally described in a common interoperable RDF format. This translation process leads, in this case, to information loss because there is not a full match between the initial format and the ontological one. The ontology is not expressive enough to represent all fields of the original data sources, although it contains more classes and properties than the original file format. For instance, the *number of visitors* is not included in the Cultural-ON ontology.

In the conversion process, the mapping from the initial input data format to the final RDF format would need to be configured for each data source. This requires knowing the OWL syntax and understanding the classes and properties of the selected ontologies.

To simplify the conversion process, an internal vocabulary was created to describe in a homogeneous and simple way data coming from different sources, without knowing the details of the ontology and further separate the input from the output. The main advantage of having this vocabulary is to hide the complexity of the ontology in the mapping process. The internal vocabulary is organized in categories, that represent contexts or ontologies; each category contains classes; each class has a number of fields. For instance, to describe museums we have defined the *museum Cultural-ON* category; this category contains classes, such as *museum* and *discipline*, and fields, such as *geographical coordinates*.

Thanks to the internal vocabulary, the conversion process is divided in two steps (see Figure 5):

- the conversion from the input data format to the internal vocabulary (*input mapping*)
- the conversion from the internal vocabulary to the ontological LOD format (*output mapping*).

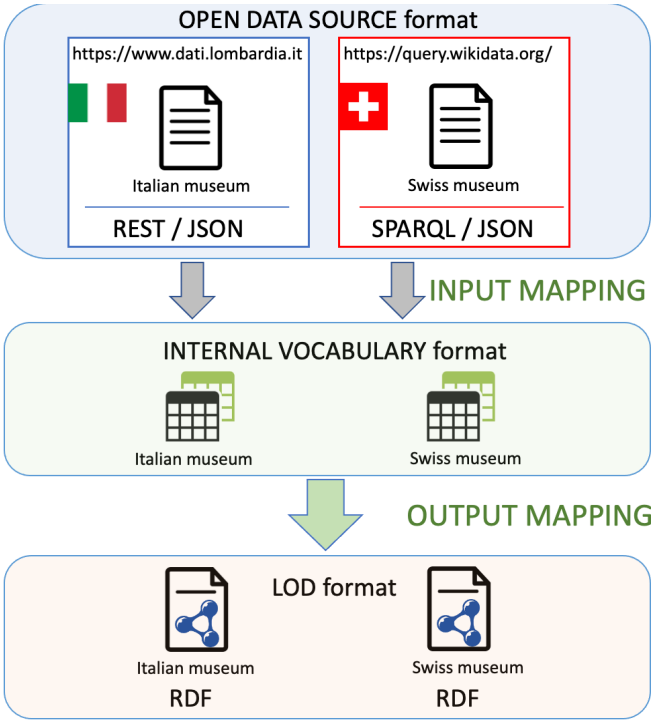


Figure 5. Two step conversion process: the museum data example.

Going back to the museum example, the two datasets, originally described in different formats and with different descriptive fields, are translated by means of the input mapping specifications into a common format, which is described by the internal vocabulary (defined in the output mapping). The resulting datasets are then converted to the LOD format, according to standard ontologies, by means of the output mapping specification. This guarantees standardization and semantic interoperability.

While it is necessary to configure the input mapping of each imported dataset towards the internal vocabulary, the output mapping of a specific category (e.g., museum) to the corresponding LOD format has to be configured only once. The first step can be accomplished by a user who knows the input format, the domain, and the internal vocabulary; the second step requires a wide knowledge of the ontologies and of the OWL language.

This mechanism that converts Open Data into Linked Open Data based on two independent and configurable steps is the peculiar feature of GIOCOnDa. With respect to other LOD frameworks, we introduced a novel flexible and dynamically configurable system that simplifies the conversion and production of LOD.

It is worth noting that the conversion process may require a certain amount of time to produce LOD, which may vary from seconds to minutes according to the complexity of the data sources and the mapping specifications. This makes the LOD dataset not appropriate for real-time applications, although dataset updates can take place at regular intervals.

IV. GIOCONDA LOD PLATFORM

The GIOCOnDa LOD platform [26] is mainly oriented to domain and ontology experts, who operate using their own user accounts to create and modify datasets. Public administrations can submit new datasets for conversion into LOD. The LOD datasets are also made accessible through a public portal without registration.

The platform, implemented as a Java-based web application, provides different functionalities that enable the publication of LOD datasets starting from open datasets, and their visualization in a catalogue or in a map.

The web app presents a menu consisting of different items: dataset catalogue, input mapping, and output mapping.

A. LOD Datasets

The catalogue shows the list of the existing datasets, as shown in Figure 6, and enables the creation of a new LOD dataset by converting an existing open dataset on the basis of the input and output mapping configuration. The system supports dataset updates at regular intervals (e.g., for air quality measurements) and propagates the changes to the RDF representation.

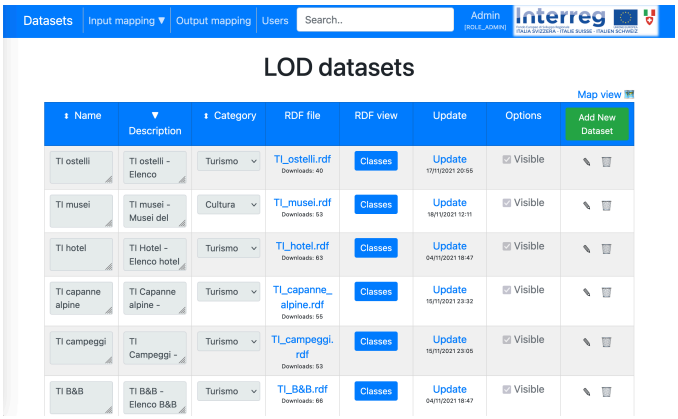


Figure 6. LOD datasets catalogue.

By clicking on the "map view" button, it is possible to visualize data on the map, whenever they have geographical coordinates, as shown in Figure 7.

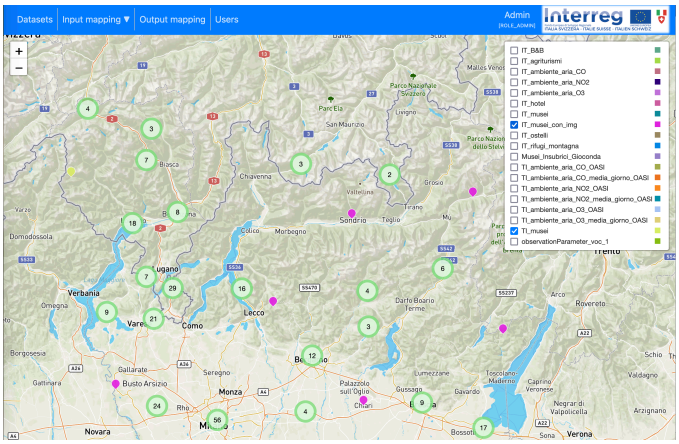


Figure 7. LOD Datasets visualized in the map.

Approximately 25 datasets about accommodation, museums, and air quality were boosted to LOD through the GIOCOnDa platform. Concerning validation, the output mapping process guarantees by design and implementation that the produced datasets are accurately serialized in RDF format conforming to the selected ontologies. A further manual checking was accomplished on some resources of each typology.

From the LOD datasets page (shown in Figure 6), it is possible to add a new dataset by clicking on the corresponding button. This action activates the conversion process, driven by the input and output mapping on the specific data sources. The "Add new dataset" function was used to generate the single datasets shown in Figure 6, but it could also be used to combine more sources to produce a unified dataset (see Figure 8). For instance, to create the "Insubric museum" dataset, it is possible to select the Italian and Swiss museums as data sources, "IT musei" and "TI musei" respectively.

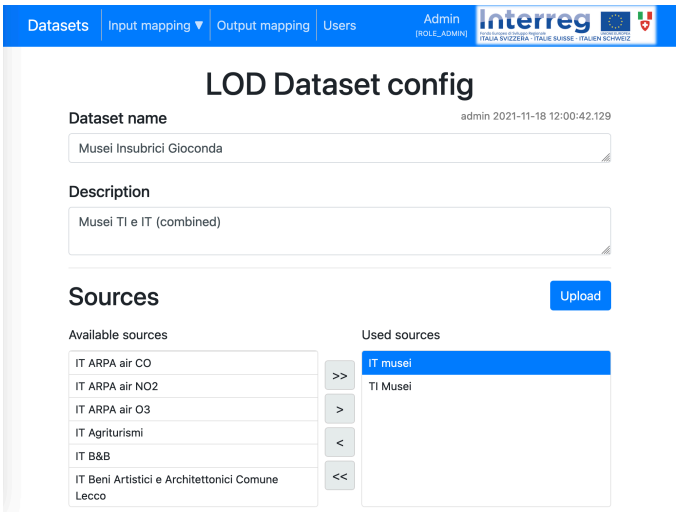


Figure 8. New dataset creation.

B. *Input Mapping*

The Input mapping concerns the configuration of the conversion from the input format to the internal vocabulary. Together with the output mapping, it enables to configure the conversion from different input data sources to the final LOD format.

The system accepts input data retrieved from sources in different formats, such as JSON, CSV, and XML, and using different import modes, such as Rest APIs, SOAP APIs, and SPARQL Queries. For each input format and import mode the conversion towards the internal vocabulary is configured through the input mapping.

Figure 9 clarifies how the mapping mechanism works: the first column shows the fields of the original data source; the second and third columns concern the internal vocabulary, where in particular the second identifies the category, and the third the field.

Categories and fields of the internal vocabulary are predetermined and selected from a drop-down menu, while the source fields must be entered by hand, according to the data structure in use, as explained later.

In the example shown in Figure 9, fields of the "Cultural-ON museums" category are used; the "email_sede" field of the initial source, representing the email of the museum site, for example, is mapped into the "email" field of the "museum Cultural-ON" category of the internal vocabulary. In some cases, it is possible to group multiple fields of the data source into a single field of the internal vocabulary; for example, to compose the field "full_address", more fields of the input source are used.

In this mapping, particular attention is dedicated to how geospatial data are represented [27]. This is essential to guarantee interoperability and efficient sharing of information across different regions and national standards. The Cultural-ON ontology assumes, by default, that spatial data are represented in the geocentric Datum WGS84 and that the coordinates are expressed in terms of latitude and longitude. Therefore, data are transformed into this system when they are imported into the GIOCOnDA platform and appropriate metadata are added to make the Coordinate Reference System (CRS) explicit.

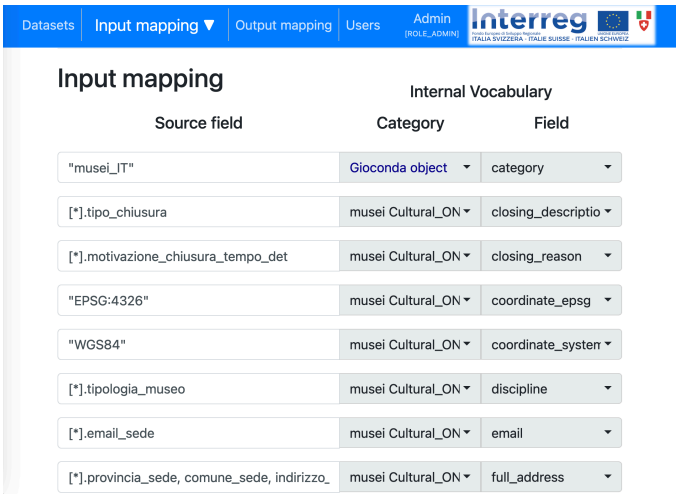


Figure 9. Input mapping.

The input mapping mechanism is the same for the different import modes, which only differ for the way the original data structure is imported.

For instance, Figure 10 shows an example of configuration of a JSON data source imported through Rest APIs.

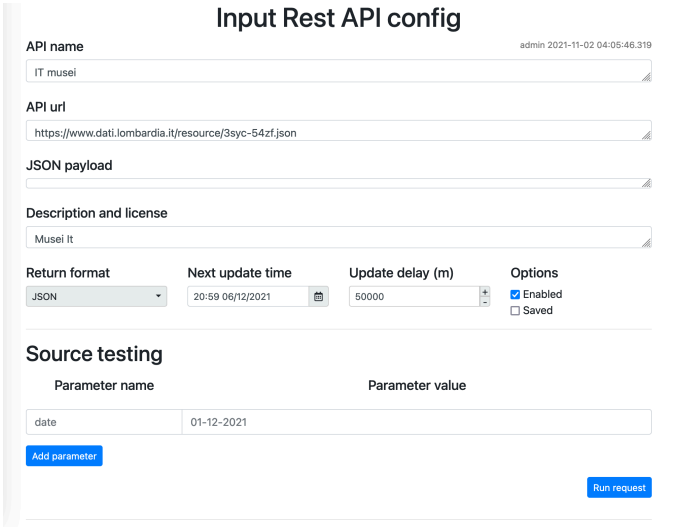


Figure 10. Rest API configuration.

On this page the data source is first defined through a name, a URL, and the licence, followed by a testing section and the input mapping section as presented in Figure 9. In the "Source testing" section there is a "run request" button, which is used to verify the proper functioning of the specified REST call; this button visualizes all the returned fields of the input source together with their values. This call is also useful to examine the field names to be used as input sources in the configuration of the input mapping.

C. Output Mapping

From the output mapping page, it is possible to create, modify and extend the internal vocabulary, and define its mapping to the ontology. This process requires a deep knowledge of ontological concepts and existing reference ontologies. Nevertheless, this mapping has to be done only once for each category by an expert.

As already said, the internal vocabulary consists of several categories, similar to contexts or ontologies; each category contains classes, with a number of associated fields. Examples of categories include museums, addresses, accommodations, etc.

As shown in Figure 11, the output mapping defines the match between internal vocabulary classes and ontology classes, and between fields of the internal vocabulary and object and datatype properties of the ontology; this is visible by activating the "Show fields" button.

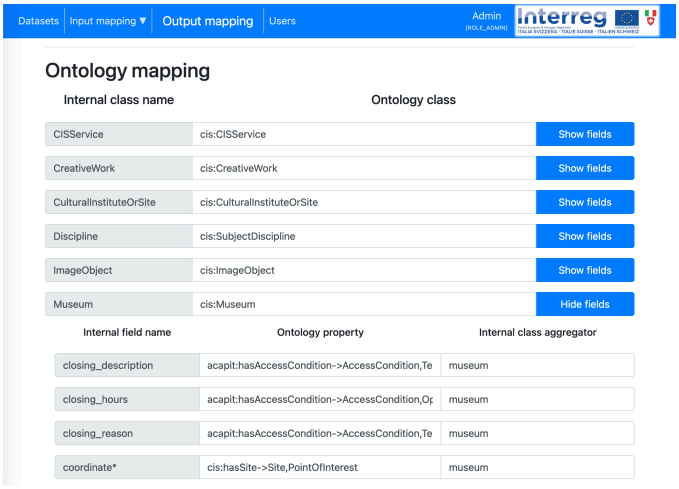


Figure 11. Output mapping: class and field match.

It is worth noting that only categories and fields of the internal vocabulary defined in the output mapping can be used in the input mapping (but not classes), providing in this way a simplified version of the data structure for non-expert users.

Another action that takes place in the output mapping is the interlinking configuration. For this purpose, a specific interlinking module was developed and integrated into the GIOCOnDa LOD platform. The next section describes how the interlinking process works.

V. The Interlinking Module

In order to boost a dataset to 5* level, it is necessary to enrich the RDF file with connections towards external sources. The rules to create cross-reference links towards external datasets, such as Wikidata, are defined in specific files, generated using the Silk Link Specification Language, serialized as .xml using the Linkage Rule Editor.

Once the Silk files are generated, it is necessary to configure the interlinking and then activate it.

The configuration is accomplished in the output mapping page, where one or more interlinking files can be associated to each category of the internal vocabulary (e.g., museum). This association is called "interlinking configuration".

The activation takes place on the datasets page, where it is possible to enable or disable interlinking on a specific dataset.

The rest of this section describes the Linkage Rule Editor, the interlinking configuration, and the interlinking activation in more detail.

A. The Linkage Rule Editor

The Linkage Rule Editor is part of the Silk Linked Data Integration Framework [18]. This open-source framework for integrating heterogeneous data sources was selected for the easiness to be adapted and integrated into the GIOCOnDa LOD platform. In general, the primary use cases of Silk include: generating links between different data sources and applying transformations to data from structured data sources; Linked Data editors can use Silk to set up RDF links from their data sources to other data sources on the web. Silk is powerfully based on the Linked Data paradigm, where, on the one hand, RDF provides an expressive data model for the representation of structured information; on the other hand, RDF links are set up between entities in different data sources.

The rules to create cross-reference links towards external datasets are defined in the Linkage Rule Editor page (Figure 12).

Link specifications can be created using the graphical interface or manually in XML. Using the Silk Link Specification Language (Silk-LSL) declarative language, developers can specify which types of RDF links must be detected between data sources, as well as the conditions that data elements must meet to be interconnected. These link conditions can combine various similarity metrics and can take into account the graph around a data element (entity), which is identified using the RDF path language.

Silk accesses the data sources via the SPARQL query language, supporting the use of local and remote SPARQL endpoints.

Figure 12 shows the creation of the owl:sameAs link between different entities based on the comparison of their geo-location coordinates. In particular, the first sourcePath and targetPath take the latitudes, while the latter two take the longitudes both from instances of a museum of the GIOCOnDa LOD dataset and of Wikidata. We use the following criteria to match entities from the two data sources: if two entities of the same type (e.g., one museum from the GIOCOnDa LOD dataset, another from Wikidata) are located within 10 meters each other, then they are considered as the same entity.

The process corresponding to the rules defined in the example of Figure 12 is the following. First, the coordinates are obtained via the corresponding SPARQL queries; then a numerical comparison (*numericEquality*) is performed both between the latitudes and between the longitudes, with a precision up to the 4th decimal place and a maximum threshold of 0.0001. This is equivalent to about ten meters in the physical world. If both pass the equality comparison (*min* represents a logical "and" aggregation), the interconnecting link is created by declaring an *owl:sameAs* relationship. In simple words, two museums are considered the same if their distance is below ten meters.

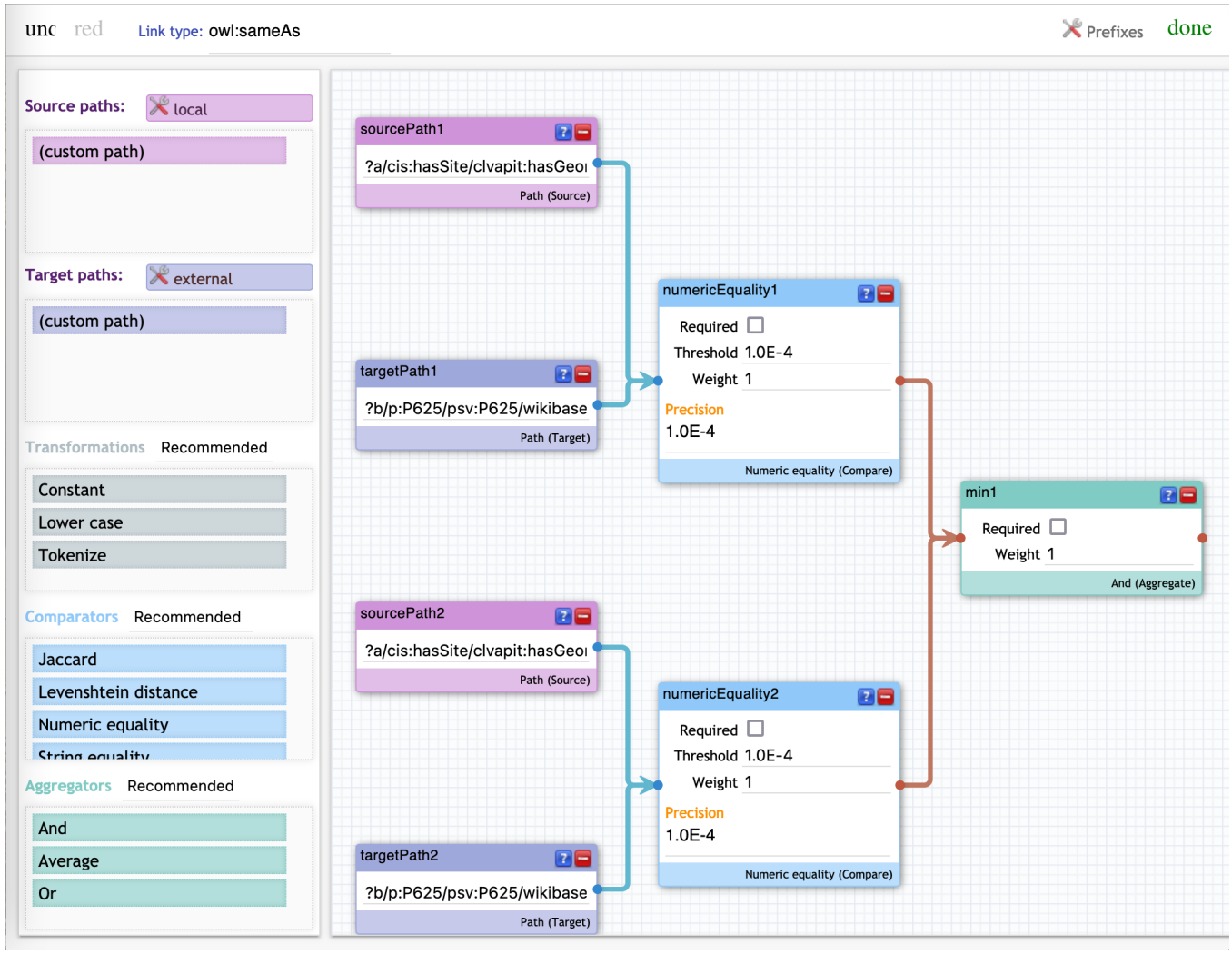


Figure 12. Example of the Linkage Rule Editor use.

B. *Interlinking configuration*

Figure 13 shows an example of the interlinking configuration within the Output mapping page for a specific category. On this page it is possible to upload a rule set from an external file, remove an existing set, create a new rule set, or modify an existing one, by opening the Linkage Rule Editor page (see also Figure 12).

It is worth noting that the Linkage Rule editor is a page fully integrated into the GIOCOnDa LOD platform, after adaptation of the Silk original editor page.

C. *Interlinking activation*

The interlinking activation takes place on the configuration page of a dataset, where, through selection, it is possible to activate or deactivate one or more configuration files on the current dataset (Figure 14).

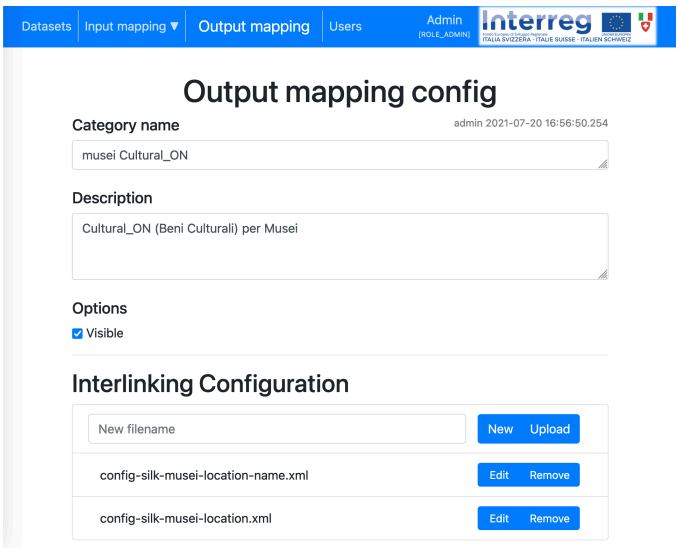


Figure 13. Example of interlinking configuration for a category in the output mapping configuration page.

The ability to simultaneously use multiple configurations allows interlinking between the generated RDF dataset and multiple data sources to be performed; for example, interlinking with the Wikidata endpoint with one configuration file, while interlinking with Open Street Map with a second one.

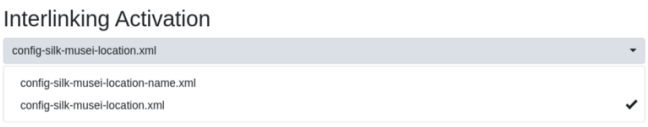


Figure 14. Example of interlinking activation.

VI. LOD PUBLISHING PROCESS: USE CASES

To better understand how the process of publishing a new dataset works, three different use cases can be distinguished:

- request for publication of a dataset with the **same structure** of an existing one (for example, a new dataset structured as the Lombard museums). Since the mapping from the input format to the internal vocabulary of that category is already configured, the conversion will simply take place by duplicating the existing input mapping configuration; the output mapping is already configured;

- request for publication of a dataset with a **new initial structure** that **can be mapped** into the existing internal vocabulary (for example, a new dataset about museums, structured in a different way compared to the existing datasets). In this case it will be necessary to configure the input mapping from the input format to the internal vocabulary of that category; the output mapping is already configured;

- request for publication of a dataset with a **new initial structure** that **cannot be mapped** into the existing internal vocabulary (for example, a new dataset relating to bike sharing, for which a vocabulary has not been defined yet). In this case it will be necessary: to look for an ontology that models the domain; to extend the internal vocabulary by adding the category "bike sharing"; to define the correspondence between the internal vocabulary and the domain ontology (output mapping); to configure the matching between the initial format and internal vocabulary (input mapping).

In summary, in the first case, there is no need to configure input and output mapping, since they already exist; in the second case, a new input mapping configuration is required; in the third one both the input and the output mapping need to be configured.

In all cases, once the two configurations of input and output mapping have been defined, it is possible to proceed with the creation of the new dataset.

VII. CONSUMING GIOCONDA OPEN DATA

Linked Open Data is useless if it cannot be extracted in a format convenient for processing and exploiting the enrichment and interlinking of LOD.

To this end, the GIOCOnDa platform provides a SPARQL endpoint that can be queried to extract LOD data from the RDF data sources stored in the Virtuoso server, providing ontology-based data access. The endpoint is available at the URL: https://gioconda.supsi.ch:8890/sparql. Thanks to this endpoint, anyone can fetch homogeneous data to implement new applications that can exploit the enrichment and interlinking of LOD, with a limited knowledge of the complex notions of ontologies and RDF.

As case studies, two applications that provide new visualizations of the data are illustrated. The first provides an overview of the museums in Tessin and Lombardy, whose data have been already described in the previous section. The second application provides a list of holiday accommodations in Tessin and Lombardy.

A. Overview of museums in Tessin and Lombardy

As first example, we implemented an application that provides an overview of museums in Tessin and Lombardy. The application is interactive and allows zooming and focusing on a particular museum to access its main information: name of the museum, type of museum, geographical location, address, and a picture. These attributes are collected from public sources based on Linked Open Data (namely: Regione Lombardia portal and Wikidata), converted and interlinked in the GIOCOnDa platform.

The purpose of the application is to allow people to explore and have an overview of the distribution of various types of museums in the two regions. People who have to take important decisions related to the presence of museums in a certain area, such as opening a new museum or establishing a new partnership among different museums covering the same subjects, may use this application.

To implement the application, the visual analytics tool Tableau [28] was used. Tableau allows the integration of a large number of data sources, but unfortunately it does not allow the direct integration of SPARQL endpoints. To overcome this problem, the data.world platform [29] was used as a bridge to connect the GIOCOnDa SPARQL endpoint with the Tableau platform. Data.world is a cloud-based service that allows the creation of a repository of data that can be integrated with dozens of applications for analytics and visualization. In that way, we can produce visual representations of data accessing data directly from the GIOCOnDa platform (see Figure 15).

Figure 15. Structure of integration of GIOCOnDa, Data.word, and Tableau.

With this architecture, the dashboard illustrated in Figure 16 was implemented in Tableau. This dashboard consists of 4 coordinated views. On the top left there is a bar chart that shows the quantity of the different types of museums in the two regions.

The user can filter out specific elements to see only data related to a particular region (Lombardy/Tessin) or location (city). A detailed view on the right provides a picture of the museum (if existing) and its details (name, location, type, and number of visitors per year). On the bottom left, the largest area is dedicated to a dot representation of museums on a map, and a tabular list of them on the right. Hovering the cursor of the mouse over a particular dot opens a new window with the details of the museum. Using this dashboard, the viewer can graphically represent LOD data extracted from GIOCOnDa datasets to make sense of the various types of museums and their spatial distribution in the two regions.

B. *Overview of accommodations in Tessin and Lombardy*

The second case study provides a list of holiday accommodations in Tessin and Lombardy. Like in the previous application, the purpose is to explore the different types of accommodations across the two regions in a uniform way. For this new application the source data are taken from Wikidata (that provides a list of Italian accommodations) and Ticino Turismo (for the Swiss data). GIOCOnDa regularly fetches data from these data sources, in order to keep data regularly updated, and store these open data in its internal data structures.

To extract the data from the GIOCOnDa platform, the same infrastructure described in the previous case study was used. In this case, new SPARQL queries had to be written and a new visual interface to extract and visualize data relevant to this new goal was implemented.

Figure 17 shows the dashboard implemented for this purpose. As in the previous application, we kept the approach of using a dot map to represent locations of accommodations. The circle colours encode the different types of accommodations (hotel, hostel, B&B, ...). A bar chart at the top shows the percentages of each type of accommodation in the two regions. A possible scenario for usage of this visualization is the following. An entrepreneur wants to invest in opening a new hospitality structure in Ticino. He/she may check this visualization and see that Agriturismo (farmhouses) and B&B are largely underrepresented in Tessin with respect to Lombardy, so he/she decides to open one or more structures of this typology. This could be an example of a decision driven by data and supported by this application.

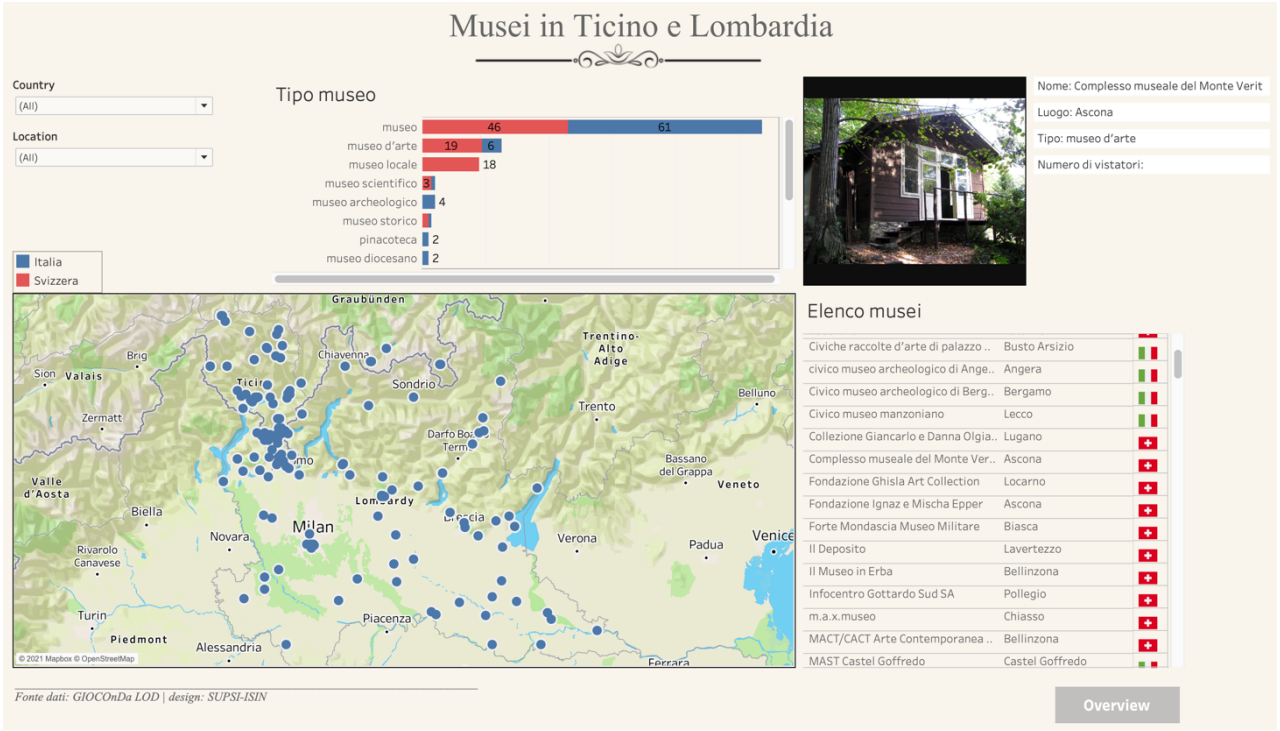


Figure 16. Overview of museums in Ticino and Lombardy.

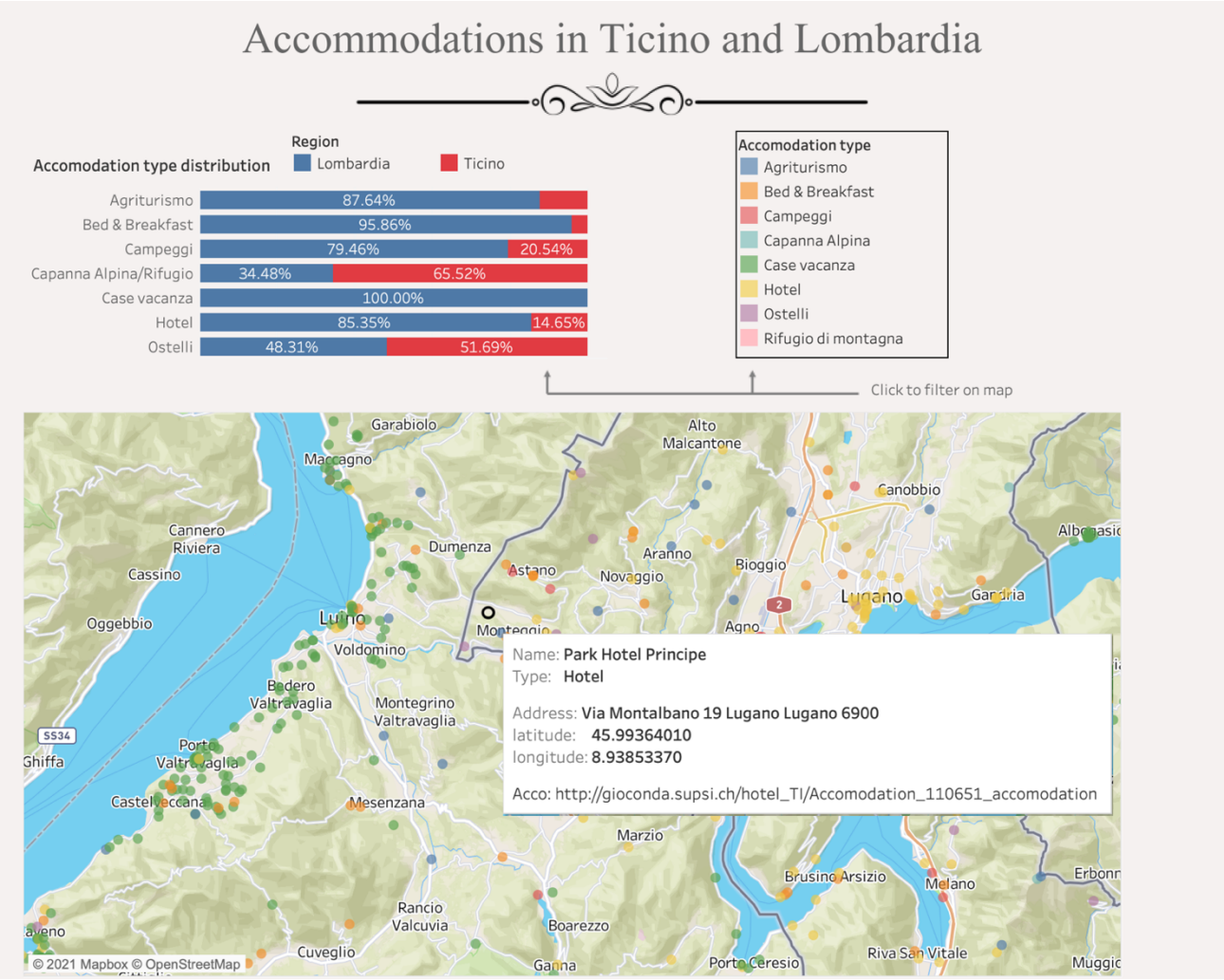


Figure 17. Overview of holiday accommodations in Ticino and Lombardy.

VIII. Conclusion and Future Work

This paper has presented a platform that facilitates the process of conversion of open data to Linked Open Data and some examples of applications that consume LOD to reveal information that cannot be captured in a different way.

The GIOCOnDa LOD platform contains a number of conversion configurations that are already available to translate different data sources to LOD in various domains.

The complexity of mapping existing data to standard ontologies is one of the major issues preventing a larger diffusion of LOD. The GIOCOnDa platform reduces the complexity of this process that would require deep knowledge of ontologies and programming skills.

If we analyse the conversion process in further detail, it is possible to distinguish different use cases and complexity levels, according to the initial structure of the dataset to be converted.

If the dataset has the same structure of an existing one (for example, a new dataset structured as the Lombard museums), the conversion is very simple, since the input mapping is similar to an existing one (it just needs to be duplicated) and the output mapping is already defined. However, the platform is also flexible and extensible, and enables to import and convert other datasets: for example, the conversion to LOD of a new dataset about museums, with a structure that can be mappable into the existing internal vocabulary, only requires the configuration of the input mapping from the initial format to the internal vocabulary, because the output mapping is already configured. More labour-intensive but still possible is to convert a dataset with a new structure, not mappable into the existing internal vocabulary; for example, a new dataset about bike sharing.

In addition to the conversion of structured open data to standard RDF open data, another important step of the adopted methodology to produce LOD is the identification and creation of interlinks between datasets. A specific interlinking module was developed to configure and activate the process of identification of cross-reference links towards external datasets. The integration of the interlinking module

in the GIOCOnDa LOD platform enables lifting datasets to 5* level, creating added value through an Extract-Transform-Load (ETL) pipeline. This is demonstrated, for instance, in a showcase that presents data about museums of the Insubric region taken both from the GIOCOnDa LOD datasets and from Wikidata.

Despite the benefits offered by the platform to publish LOD datasets (visual interface, configurability, extensibility), it also presents some limitations: the main one is the possibility of information loss during the conversion to LOD if there are fields not represented in the selected ontology. A possible solution would be the extension of the selected ontologies with additional fields and the publication of the new version with appropriate documentation.

By using the platform functionalities, a consistent number of datasets were created and will be produced in the future. However, it can be difficult to interpret data in a tabular way; data gain more value when they are visualized. By means of visualization tools it is possible to communicate data findings and identify critical information to pull insights.

To demonstrate the potential of Linked Open Data some applications were developed concerning museums and accommodations. In both cases the applications provide an overview of the distributions of these entities in the Insubric area, supporting decision making in these domains. Once the number of datasets increase, new opportunities will be available to develop consuming applications.

In conclusion, it is possible to state that the GIOCOnDa project has reached its objectives thanks to the LOD platform and the connected applications: a consistent number of LOD datasets were published and examples of data consuming applications were implemented to show the LOD potential.

In the future the platform will be further developed, completing and improving some parts, such as increasing the internal vocabulary to publish new LOD datasets, simplifying the user interface and refining the interlinking rule definition.

ACKNOWLEDGMENT

We acknowledge the European Regional Development Fund, the Tessin Canton and the Swiss Confederation for the financial support provided to the project. We also acknowledge all the partners involved in the project for their contribution to the different project activities.

REFERENCES

[1] L. Sommaruga, N. Catenazzi, D. Bertacco, and R. Mazza, “From Open Data to Linked Open Data - The GIOCOnDa LOD platform”, ALLDATA 2021 (The Seventh International Conference on Big Data, Small Data, Linked Data and Open Data), April 18, 2021 to April 22, 2021 - Porto, Portugal.

[2] GIOCOnDa, Integrated and holistic management of the Open Data life cycle (Gestione integrata e olistica del ciclo di vita degli Open Data) https://progetti.interreg-italiasvizzera.eu/it/b/78/gestioneintegrataeolisticadelciclodivitadegliopendata [retrieved: 2021.11.30].

[3] Open Data, Regione Lombardia, https://dati.lombardia.it [retrieved: 2021.11.30].

[4] Arpa Lombardia, https://www.arpalombardia.it/ [retrieved: 2021.11.30].

[5] Ticino Turismo, https://www.ticino.ch [retrieved: 2021.11.30].

[6] OASI, Osservatorio Ambientale della Svizzera italiana, https://www.oasi.ti.ch [retrieved: 2021.11.30].

[7] T. Berners-Lee, “Linked Data”, http://www.w3.org/DesignIssues/LinkedData.html, 2009 [retrieved: 2021.11.30].

[8] T. Heath and C. Bizer, Linked Data, “Evolving the Web into a Global Data Space”, Synthesis Lectures on the Semantic Web: Theory and Technology, Morgan & Claypool Publishers, 2011.

[9] The Lucero Project, https://lucero-project.kmi.open.ac.uk/ [retrieved: 2021.11.30].

[10] L. Ding, T. Lebo, J. S. Erickson, D. DiFranzo, G. T. Williams, X. Li, J. Michaelis, A. Graves, J. G. Zheng, Z. Shangguan, J. Flores, D. L. McGuinness, and J. A. Hendler, “TWC LOGD: A portal for linked open government data ecosystems”, Journal of Web Semantics, vol. 9(3), pp. 325-333, Sep. 2011, doi: 10.1016/j.websem.2011.06.002.

[11] Open Data e Linked Data, https://www.beniculturali.it/open-data-e-linked-data [retrieved: 2021.11.30].

[12] A. Meherhera, I. Mekideche, L. Zemmouchi-Ghomari, and A. Réda Ghomari, “A Survey of Current Approaches for Transforming Open Data to Linked Data”. 4th Edition of the National Study Days on Research on Computer Sciences, JERI’2020, Jun 2020, Saida, Algeria. hal-03211592, https://hal.archives-ouvertes.fr/hal-03211592/document [retrieved: 2021.11.30].

[13] F. Bauer and M. Kaltenböck, Linked Open Data: The Essentials - A Quick Start Guide for Decision Makers. Edition mono/monochrom, Vienna, Austria, 2012, ISBN: 978-3-902796-05-9, https://www.reeep.org/LOD-the-Essentials.pdf [retrieved: 2021.11.30].

[14] W3C, “Best Practices for Publishing Linked Data”, W3C Working Group Note 09 January 2014, https://www.w3.org/TR/ld-bp/ [retrieved: 2021.11.30].

[15] Agenzia per Italia Digitale, Guidelines for semantic interoperability through Linked Open Data (“Linee Guida per Interoperabilità semantica attraverso i Linked Open Data”), http://www.agid.gov.it/sites/default/files/documentazione_trasparenza/cdc-spc-gdl6-interoperabilitasemopendata_v2.0_0.pdf 2012 [retrieved: 2021.11.30].

[16] Ontologie e Vocabolari Controllati, https://github.com/italia/daf-ontologie-vocabolari-controllati [retrieved: 2021.11.30].

[17] G. Lodi, “OntoPiA –The network of ontologies and controlled vocabularies for public admistration (OntoPIA - La rete di ontologie e vocabolari controllati per la pubblica amministrazione)”, Open Data Sicilia – raduno annuale, 9/10 novembre 2018, http://ods2018.opendatasicilia.it/presentazioni/Lodi-OntoPiA.pdf, 2018 [retrieved: 2021.11.30].

[18] Silk, The Linked Data Integration Framework. http://silkframework.org/ [retrieved: 2021.11.30].

[19] Open Link Virtuoso, https://virtuoso.openlinksw.com [retrieved: 2021.11.30].

[20] C. Bizer and R. Cyganiak, “D2R Server - Publishing Relational Databases on the Semantic Web”, 5th International Semantic Web Conference (ISWC 2006), Athens, USA,

November 2006, http://richard.cyganiak.de/2008/papers/d2r-server-iswc2006.pdf, 2006 [retrieved: 2021.11.30].

[21] R2ML: RDB to RDF Mapping Language https://www.w3.org/TR/r2rml [retrieved: 2021.11.30].

[22] Musei riconosciuti da Regione Lombardia https://www.dati.lombardia.it/Cultura/Musei-riconosciuti-da-Regione-Lombardia/3syc-54zf [retrieved: 2021.11.30].

[23] Museo Vela https://www.wikidata.org/wiki/Q3867651 [retrieved: 2021.11.30].

[24] Cultural-ON (Cultural ONtology): Cultural Institute/Site and Cultural Event Ontology, https://w3id.org/italia/onto/Cultural-ON [retrieved: 2021.11.30].

[25] Social Media / Contact and Internet ontology - Italian Application Profile https://w3id.org/italia/onto/SM [retrieved: 2021.11.30].

[26] GIOCOnDa LOD platform https://gioconda.supsi.ch/ [retrieved: 2021.11.30].

[27] J. F. Toro Herrera, D. Carrion, M. A. Brovelli, N. Catenazzi, L. Sommaruga, and D. Bertacco "Geospatial Data dissemination in the GIOCOnDA project", extended abstract, ASITA (Federazione delle Associazioni Scientifiche per le Informazioni Territoriali e Ambientali, http://atti.asita.it/ASITA2021/Pdf/039.pdf) July 2021, [retrieved: 2021.11.30].

[28] Tableau: Business Intelligence and Analytics Software https://www.tableau.com/, [retrieved: 2021.11.30].

[29] "The catalog for metadata and data management", https://data.world, [retrieved: 2021.11.30].

Implementing Ethical Issues into the Recommender Systems Design Using the Data Processing Pipeline

Olga Levina
Brandenburg University of Applied Sciences
Brandenburg an der Havel, Germany
e-mail: levina@th-brandenburg.de

Abstract—**Applying information systems within a business process requires a good understanding of the expected benefits, system requirements as well as of the effects that the process change will have on its actors and stakeholders. Integrating machine learning based systems (MLS) into a business process requires an even broader focus on potentially affected users and stakeholders. Leading to changes in the process, but also in the user and stakeholder behavior, ethical values are directly influenced by the decisions taken during the data processing stages within system development. In this paper, a scenario of an MLS, a fictional recommender system for food delivery, is used to identify potential ethical issues that occur during the composition and usage of the artifact. Data centered analysis of the system development is applied to identify, which ethical values are mostly affected in each data processing stage. It is argued that even when the used data for MLS is not originated from an individual, and thus is not necessarily subject to privacy regulations, ethical analysis and socially-aware engineering of the information system are still required. Suggestions what ethical aspects can be implemented into the design of the MLS are derived here based on the presented scenario. The effects of MLS application in a business process are furthermore briefly outlined for every stage of data processing. Using this scenario-based approach allows identification of social and technical aspects that can be affected by the application of MLS in business context.**

Keywords- socio-technical systems; machine learning based systems design; ethical values; ethical analysis; business process.

I. INTRODUCTION

The challenge of the integration of ethical issues in the information systems design has been specifically laid out by Levina in [1].

The pervasiveness of algorithmic systems in our daily lives is stimulating public and research debate about their potential effects on the individual behavior and also on the society as a whole. Several companies and governmental initiatives react to this development by publishing ethical principles on how their Information Technology (IT) artifacts that involve Machine Learning (ML) components are created, leading to the so called "principle proliferation" [2]. Evidently, Information Systems Research (ISR) should manifest its leading role in pursuing practices for the creation of IT artifacts that are not only technically innovative but also socially acceptable.

This paper provides a contribution by presenting and discussing the outcomes of the ethical analysis of a paradigmatic case of a Machine Learning-based System (MLS) application. Here, an MLS is an Information and Communication Technology (ICT) that is composed of one or more algorithms working together and capsuled into one or more executable software components [3]. The ethical analysis demonstrates what ethical values are affected the most in which data processing stage. These insights allow software developers and system architects to focus the introduction of socio-technical activities accordingly. The results of the applied analysis approach lay the ground for theoretical development of a mixed methods approach that is focused on ethical reasoning in ISR and engineering of socio-technical systems [4].

The presumption of this mixed methods approach is, that ethical compliance of an IT-system is an integral part of the design process, as well as the product use. The linkage to ethical questions and the design of an IT artifact can be historically established in several ways. First of all, the core of the engineering activities, such as software engineering and IT systems design, is the solution to the design problem [5]. Since there are multiple possibilities to solve a problem, (software) engineers weight one alternative against another. The decision criteria for the design alternatives can be financial restrains, user requirements and functional fit of the alternatives. Once the chosen alternative is realized as an artifact, it will have good and bad effects. Hence, one obvious moral obligation of a (software) engineer in the role of the solution creator is to pick a design alternative that does not induce harm [5]. Thus, to create an IT system that takes into account the effects of its application on the business processes, users, as well as the effected parties, theses potential effects need to be taken into account in its design [4][5]. It is e.g., the case, when digital systems such as a recommender or a digital assistant system provide a service for its user.

A service, in the physical world, as well as digital, comes with costs that are not only monetary. It entails partial loss of autonomy in the realm it is being offered. User accepts the service if the assessed amount of the autonomy loss is acceptable and thus the user provides consent to this loss by agreeing to use the service instead of performing the offered function him- or herself. The engagement with the service can furthermore be associated by the user with loss of autonomy due to opaque processes of result generation. Social reluctance of these practices is evident. Only 19% of surveyed users of digital services believe that tech companies design their services with people's best interests in mind and

47% feel they have no choice but to sign up to services despite having concerns [8].

Identifying and complying to ethical issues in the MLS design can thus enable autonomous decisions for the user within the interaction with the service. In addition, this quality can provide a distinctive feature on the market of IT products.

Following this reasoning, the goal of this research is to expand the present literature on potential ethical issues of MLS. The structure of the paper follows this reasoning. An example scenario is presented in Section V using the data process centered ethical analysis [7] described in Section IV and requirements of socio-technical system design in Sections II and III. Suggestions about how the identified ethical issues in Section VI can be integrated into the IT system are provided in Section VII. Using the offered scenario, the process and supplementary effort to include these aspects into the artifact design can be assessed by the system designer or business engineer, providing an actionable radius to create socially acceptable IT products, as well as to lay the ground for future research questions. Conclusion and outlook on the future work finish the paper.

II. State of the art in Socio-technical aspects of recommender systems

Socio-technical systems are described via Baxter and Sommerville [4] as systems that involve a complex interaction between humans, machines and the environmental aspects of the work system. Machine learning-based systems incorporate this interaction already in their input, i.e., the data from which patterns are derived and test data sets for the mathematical models are the result of an interaction between a human and a business information system. Thus, their implementation into the organizational processes has an intermediate effect on the actors on the outside and inside of the organization. Specifically, in this context, socio-technical considerations are not just a factor within the systems development process, but they have to be considered at all stages of the development life-cycle.

For MLS the system development life-cycle includes data processing. Data processing is furthermore divided into phases of data collection, data processing, model definition, model training and calculation of the results. The socio-technical factors are triggered when the MLS results are implemented into a business process, requiring a human decision or a decision that concerns human actors. To catch these challenges the ALTAI principles were established by the European Commission [9] to help evaluate a socially aware MLS design. These are: Participation, Transparency, Human Autonomy and Auditability. These principles are considered here as facilitators for the software design approach that focuses on the person affected by the software result rather than the direct user of the software.

Identifying ethical issues that might occur during the system design allows conclusions on the ethical values that are affected in different stages of system design. This activity is considered here the first step of the incorporation of these values into the design of a socially-aware information system. Hence, a scenario for an MLS, a recommendation application, is described Section V and used to demonstrate an approach to identifying ethical issues.

III. Overview of the state of the art of Ethical analysis approaches for recommender systems

Ethical issues in the context of IT-artifacts have gained increasing attention in research over the last decade. Paraschakis [10][11] explores e-commerce recommender applications and suggests five ethically problematic areas: user profiling, data publishing, algorithms design, user interface design and online experimentations, i.e., exposing selected groups of users to specific features before making them available for everybody. Milano et al. [12] conduct an exhaustive literature review of the research on recommender systems and their ethical aspects and identify six areas of ethical concern: ethical content, i.e., content that is or can be filtered according to societal norms; privacy as one of the primary challenges of a recommender system; autonomy and personal identity, opacity, i.e., lack of explaining how the recommendations are generated; fairness, i.e., the ability to not reflect social biases; polarization and social manipulability by insulating users from different viewpoints or specifically promoting one-sided content. Milano et al. [12] also show that the recommender systems are designed with the user in mind, neglecting the interests of the variety of other stakeholders, i.e., interest groups that are being directly or indirectly affected by the recommendation. Polonioli [13] presents an analysis of the most pressing ethical challenges posed by recommender systems in the context of scientific research. He identifies the potential of these systems to isolate and insulate scholars in information bubbles. Also, popularity biases are identified as an ethical challenge potentially leading to a winner-takes-all scenario and reinforcing discrepancies in recognition. Karpati et al. [14] analyse food recommendation systems and identify several ethically questionable practices. They name the commitment to already given preferences and thus to the values of the designers as a contradiction to the potential for ethical content. Privacy, autonomy and personal identity that the authors identify as potentially vulnerable and hence suggest need to be realized via an informed concern and a disclosure about the business model used. Opacity about the origin of the recommendations as well as of the criteria and algorithms used to generate the recommendations. Fairness, polarization and social manipulability as well as robustness of the system complete the list of identified ethical issues for a food recommender.

These approaches discuss ethical impacts of recommender systems from the perspective of the receivers of the recommendations. Milano et al. [12] argue that the social effects such as manipulability and personal autonomy of the user are hard to address, as their definitions are qualitative and require the implementation of the recommender system in the context they operate, while Karpati et al. [14] offer a multi-stakeholder approach to address these issues. The data process-centered approach to analyzing ethical issues suggested by Levina [15] identifies the decision points during the MLS development, while advocating the inclusion of a laboratory phase into the

system design to assess the potential consequences (see also [16]).

This research applies a combination of data processing and ethical analysis in the attempt to identify how or whether the identified threads to ethical values that can be realized and mitigated in an MLS design.

IV. ANALYZING ETHICAL ISSUES WITHIN THE DATA PROCESSING

The general process for analyzing ethical issues within the design of machine learning- based systems has already been roughly outlined in [15]. Here the process is explained in detail and its exemplary application using a scenario of a food recommendation application is presented in Section V.

Fig. 1 shows the stages of data processing according to [17] as well as the aspects that are relevant for discovering potential ethical issues within this specific stage. Although, the *Apply* stage is not an integral part of the data processing pipeline, the effects of the application of the MLS on user behavior are important for its design and are thus included in this analysis. Furthermore, the ethical analysis differentiates between the MLS-user, i.e., a person using the MLS directly within a business process, and an MLS-*affected* user, i.e., a person or a stakeholder that is affected by the results of application of the MLS within a process.

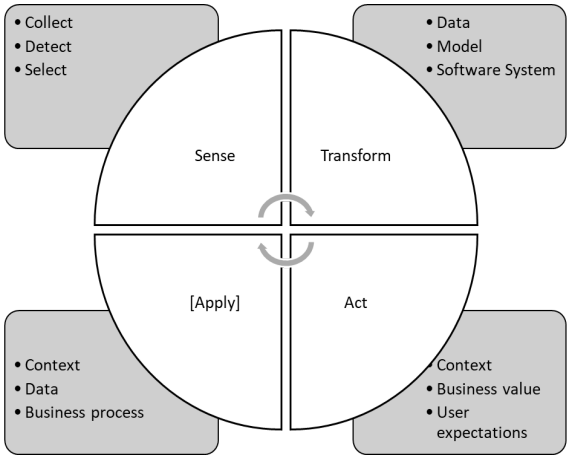


Figure 1. Data processing and relevant aspects for ethical design [18]

In the *Sense* phase the data needs to be collected, pre-processed and stored, i.e., detected for further application. The data features that are relevant to the business problem need to be selected. In the *Transform* phase data analysis methods, i.e., the mathematical model(s) used in the MLS, are in focus. Ethical issues can furthermore arise along the aspects of data manipulation, such as defining the test dataset and its features, as well as the entire software system in which the trained ML model is integrated and that has to be integrated into the business process to provide an added value. In the *Act* phase the integrated MLS in enacted within a business process to provide support for the selected tasks, i.e., to generate business value. To do so, the software system as a whole needs to adhere to the user's expectations towards usability, supported functions and expected output. The *Apply* phase is not included in the data processing pipeline. Nevertheless, to be able to analyze the potential effects of MLS applications, this phase needs to be included to reflect the view of the affected party, i.e., the external party that receives the results of the MLS application at the end of the (business) process.

Hence, this data-centered approach reflects different viewpoints on the data processing and use. While the first two phases, Sense and Transform, focus on the data and their sources, the last two phases are governed by the values of the (business) user and the affected user respectively. Thus, even when the input data for the MLS is machine-generated, the data processing phases require a socio-ethical approach to the requirements analysis and implementation.

In the following subsections the potential aspects that may arise in each phase are presented in more detail and structured along the three sub-categories: ethical aspects, technical aspects and existing methods of risk mitigation for raised technical or ethical questions (see Figures 2-5). While the ethical aspects address value-based issues within the data processing pipeline, technical issues address the technical means and tools that exist and can lead to the raise of ethical issues.

A. Sense Phase and potential ethical issues

In the sense phase of the data processing pipeline it needs to be assured that the data have been collected with the informed consent and voluntariness of the data subject. Hence, Fig. 2 shows the sub-division of the phase into the individual categories that can also be extended to accommodate further potential ethical issues.

The ethical value as defined by the European Commission in its ALTAI checklist [9] that is affected the most in the Sense phase, is the value of privacy and data governance. Being an issue that is subject to legislation and public debate, a research direction emerges in the philosophical community calling for empirical investigation of the effects of data collection under the term of ethics of influence [19]. It aims at further investigating of ethical questions in this data processing stage.

Issues associated with data collection have already been addressed in the legal form such as European GDPR legislation. Thus, legal compliance is part of the risk mitigation activities that can be taken by the enterprise applying the MLS. Risk mitigation activities may help to catch ethical issues that occur in the context of appropriation and necessity of data collection. They require, e.g., informed consent of the user to provide interaction or behavioral data. Informed consent also includes the statement of the purpose of data collection implying an opt-in function for data collection. What data is being collected is normally described in the *terms and conditions document of* the MLS. Nevertheless, data-based devices that can contain sensors and processing units might collect more data that the terms and conditions statement declare. These data can be considered a by-product of the service offered, but nevertheless, their collection and potential distribution need to be kept transparent for the future user and affected users.

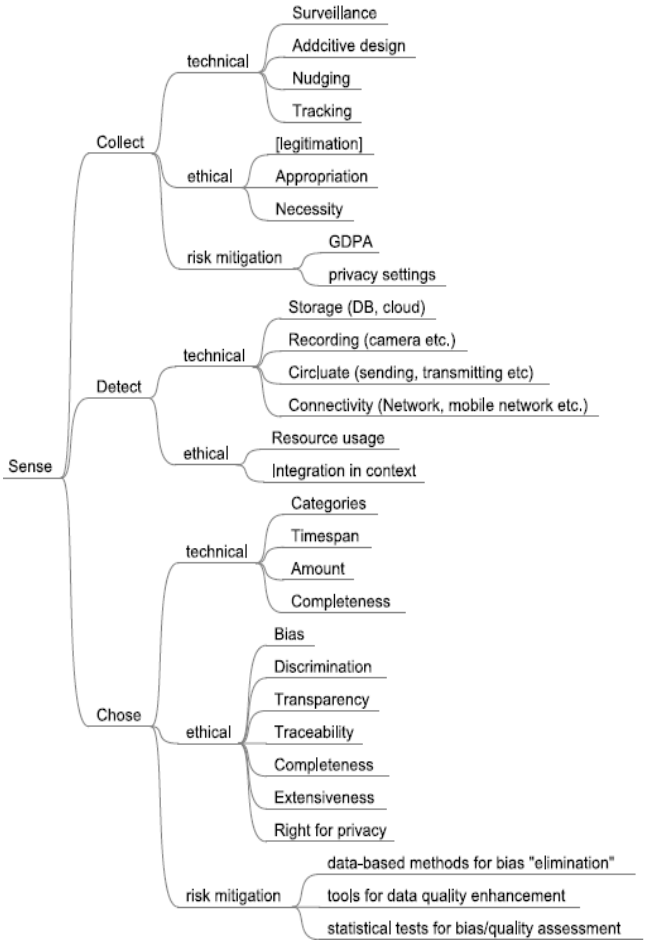


Figure 2. Potential ethical issues in the select stage

The technical realization of the data collection can thus be the origin of ethical issues. Such as the use of dark patterns in system design [20] is attempted at keeping the user engaged with the system are often aimed at collecting more data. The so called smart environments, such as the Internet of Things, are also potential sources for data collection with the focus on specific user behavior [21]. Tracking technologies, such as health tracker or internet cookies, are also technical mechanisms aimed at collecting behavioral data that might exceed the amount of the data required for the original purpose.

B. *Transform Phase and potential ethical issues*

In the Transform phase, technical aspects of model building and training are put into focus, while the ethical values that are most influenced in this phase are the values of transparency of the technical process as well as the societal and environmental well-being as described in the ALTAI checklist [9].

The model construction, i.e., the applied algorithms for pattern recognition, as well as the definition of the thresholds for the MLS results are decision points in the development process that carry potential for ethical issues. Depending on the choice of the algorithms, e.g., the energy efficiency of the performed computation is affected. Using pre-trained models to solve frequently occurring business problems, reduces the resources needed to train the model on the one hand, but on the other hand, this technique has also the potential to lead to homogenous and generalized results [22], i.e., potentially aggravating the ethical issues associated with the value of non-discrimination and fairness. The selection of the computational algorithm is also defined by the expected quality of the results [23]. Hence, the choice is partly made based on the expected quality metrics such as accuracy of the calculates prediction or recommendation by the MLS. Pursuing better accuracy can potentially mean choosing a more resource intensive mathematical model. Hence, this mathematical problem can directly relate to ethical issues.

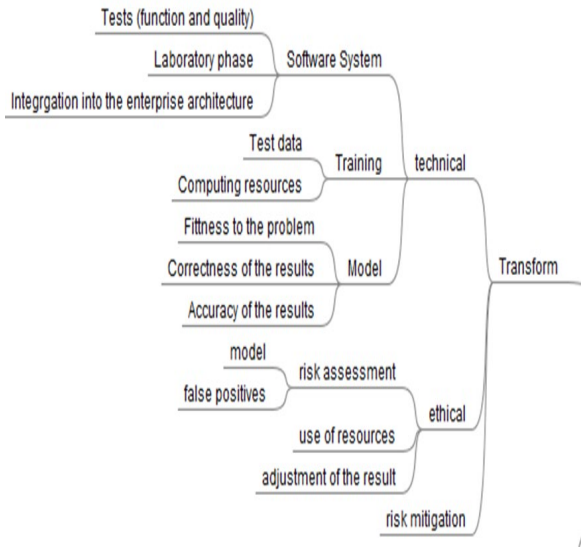


Figure 3. Potential ethical issues in the transform stage

Furthermore, the definition of the thresholds for accuracy or correctness require further ethical decisions [24], e.g., in favor of reduction of false negative or false positive results, depending on the problem at hand. To mitigate these issues, pre-trained models can be used that have already been applied on a similar problem, or industry standards can be addressed. Using industry standards bears nevertheless the negative potential, that the same thresholds would be applied in different use cases, leading to de facto standard values that might in the future lose their semantic correctness. Further potential risk mitigation measures might include meticulous description of the data set used to train the pre-trained model, description of model thresholds and parameters as well as stakeholders involved. Having this description may allow the software developers and model engineers to make an informed decision about the fitness of the model to the problem and data population at hand.

C. *Act Phase and potential ethical issues*

The Act phase focuses on the business process that is supported using the MLS in question. The results generated by the MLS and the usability of the MLS need thus to adhere to the expectations and requirements of its users. Hence, human-computer-interaction and usability aspects as well as

the control concepts such as human-in-the-loop are put in focus of the ethical analysis here. Human agency and oversight as well as communication are the values that are mostly affected in this phase. These values should guide the result integration from the social aspect as well as the technical system integration as the technical aspect of the MLS.

The value of human agency and oversight is also addressed by the handling and interpretation of the MLS results for the following process tasks. The information on the meaning of the calculated results in the context at hand as well as their interpretation is crucial for the meaningful application and generation of true added value of the system.

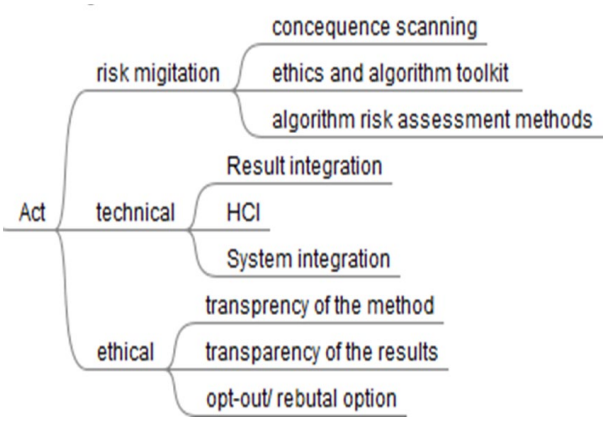


Figure 4. Potential ethical issues in the act stage

Failure to interpret the results might lead, e.g., to false decisions and thus negative consequences for the actors and stakeholders involved in the process. Also occurrence of false positive or negative results, their consequences for the actors involved as well as handling these errors should be included into the business process. Hence, a thorough user training for the process actors involved in the process using MLS is an essential tool for risk mitigation in the act as well as in the apply phase.

D. Apply Phase and potential ethical issues

While the apply phase does not belong to the technical data processing, it is involved in the ethical analysis of handling data in business context. The values accountability, communication and human agency and oversight as described in the ALTAI checklist [9] are mostly affected here, when the MLS application is realized.

The application of MLS in a business process requires good knowledge of the process and of the consequences that should be addressed, e.g., in user trainings. But it may also lead to the loss of previously present skills for the process actors such as moral [25] or decisional [7] de- skilling.

To mitigate these threads to human autonomy, control, expertise, and behavioral change [26] guidelines for MLS development can be applied as well as scheduled audits of the process and the effects on the process performance before and after MLS application can be helpful. Also, changes in the process environment should be monitored using, e.g., performance indicators from the domain of Green BPM [27].

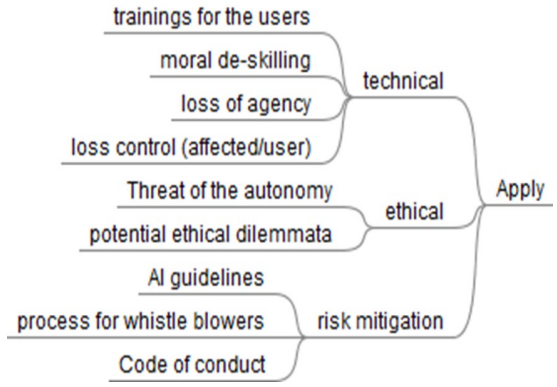


Figure 5. Potential ethical issues in the apply stage

V. AN EXAMPLE APPLICATION: FOODAPP- THE APPLICATION FOR MEAL DELIVERY

The FoodApp is a fictional application based on a three-sided digital platform that is implemented as a mobile app. It is a branch of a fictional large company Acima that offers on-demand individual transportation provided by freelancing drivers. To further explore the transportation market, Acima started FoodApp, a fast growing food delivery platform connecting the customer, restaurant owner and the delivery partner. It allows the customer to choose from a large database of participating restaurants and order a menu to be delivered to the customer's address via delivery partners. The eater can choose a specific delivery partner based on the ratings of the currently available partners. The payment process is integrated into the platform as is the real-time tracing of the order delivery.

The platform business goal is the "fast and easy food delivery whenever, wherever". To achieve this goal a MLS, a recommender system, is used to provide the best food suggestions for the user in accordance to the indicated preferences and the order history. The business performance indicators for the FoodApp include the return and re-order customer rates, as well as customer number growth rates. The implemented ML-model is thus optimized to drive user's re-ordering on the platform.

To use the FoodApp the customer downloads it on the mobile device granting permissions for it to access the location of the device. Further, a profile including information on delivery address, name, e-mail and phone number is required. Payment methods and login to the payment provider is further required. No manual modifications concerning the data collection by the app is possible. Then, the meal preferences such as preferred cuisine or menu item need to be indicated or a meal can be chosen from the provided suggestions. The first suggestions are based on the historical frequency of the orders made within the community in the area of eater's location. A rating

system for restaurant and delivery partner performance is implemented.

The platform gains revenues from the customer via convenience charge, fixed commissions and marketing feeds from the restaurants, while providing the assignments and the payment to the delivery partners, as well as the technical infrastructure for the platform participants. The application is a key driver of Acima's revenue and is a fast- growing meal delivery service with over 15 million users worldwide. Additionally, the platform includes an app for delivery partners that provides the possibility to accept or decline a specific delivery job, monitor the revenues, rate the restaurant's delivery process, as well as provide directions to the restaurant and to the eater.

VI. IDENTIFYING ETHICAL ISSUES USING THE ACENARIO-BASED APPROACH

To identify ethical issues in the MLS used by FoodApp, data process-oriented analysis [15] has been conducted. Since the core component of the FoodApp MLS converts (user) data into a food recommendation, ethical issues are explored using the data-centered ethical analysis approach described in Section IV. The ethical analysis looks at the data process within the system's design and identifies some of the relevant aspects, where ethical values are affected and ethical questions arise influencing the system design. To identify potential ethical issues, questions along the data-processing stages are asked, as suggested by [15].

First potential questions for the first stage, *sense*, are structured along the sub-stages: collect, detect and select. For the act of collection, some of the central questions are:

- Was the user aware of the mode or amount or content or context of data collection?
- Was the data collection conducted with sufficient legal compliance?
- Were any dark patterns [20] involved in the obtainment of the data?
- Are the data collected necessary for the MLS to function according to its purpose?
- Are there opt-in possibilities for different types of data collection?

FoodApp's business goal is to engage the user in the re-ordering of the food via FoodApps's digital platform. The user interacts with the app aiming for a comfortable provision of the favorite food in an efficient way. Therefore, as described in Section I, the user is inclined to give up some autonomy within this process. Nevertheless, in the digital realm the user is often not aware of what elements of his/her *autonomy* are jeopardized when the digital service, here food selection and ordering via a digital platform, is created [12][13]. E.g., in the FoodApp the location information of the device is transferred per default to the platform. Also the app has default access to the microphone and camera of the device. While the user can still change these settings, s/he is often unaware of the default access requirements of the app or does not know what access is needed for the app to function. Thus, the questions that arise in the collect sub-phase should address the actual data collection and their relation to the function of the service provided by the app. Also the questions on whether the data are stored permanently at the platform or have an expiration date are crucial in the detect sub-phase. In the select sub-phase, the questions about:

- data quality
- data sufficiency
- data sources
- representativeness for the solution of the given problem

will have to be addressed. Since data are the fundament for the further model building, their amount, quality, focus in relation to the problem solution (here: providing a food recommendation) as well as the rightfulness of its collection are essential for a mathematically good model design, representative training dataset as well as a socially-aware information system.

Additionally, the amount and sources of the collected data are mostly defined by the *business model*. As FoodApp would like their users to return to the app, it will need among other factors, very good recommendation results as well as a frictionless ordering process together with a reliable problem handling mechanisms to fulfill basic customer expectations [14]. The business model provides essential guidelines for the sense and transformation phases, including the type of information system that can be used to support the *business goals*. The first requirement, i.e., very good recommendation results in terms of user's preferences, can be realized using a recommendation algorithm based on the collected data from the user as well as from the users with similar preferences or history on the platform [14]. Since the user activity data might provide additional patterns for the recommendation, it also provides a potential reason to keep the user engaged on the app for the longest possible time, which might involve the use of dark patterns in the app design [20].

Beside from the user data, FoodApp database should include data on the restaurants available for ordering and delivery through the platform. Addressing the restaurants is part of the business model and might also be part of the business focus, as restaurants can be included on the platform according to specific *criteria*, e.g., reviews on other platforms, personal preferences, number of years in business, etc. leading to a potential pre-selection of available food choice on the platform. To avoid subjectivity in this dataset, a neutral source for the identification of the restaurants and confirmation of their availability should be considered. Additionally, the delivery network of partners that will pick up food at the restaurants and deliver it to the customer's door need to be established and equipped with the means to be contacted, payed and managed by the platform.

Hence, FoodApp needs to establish an *ecosystem*, similar to a classic supply chain, to be able to fulfill its business goal or even to be able to operate according to its business model. Building up such an ecosystem as well as the potential to manage the orders for delivery, provides Acima as a digital platform with a *specific power* over the delivery partners as well as the restaurants that can have extensional effects on the partners involved in the ecosystem as well as the bigger area of stakeholders. See [30] for the discussion of potential

ethical issues emerging from the digital platform as an ecosystem.

FoodApp's user profile provides the information that is, among others, needed for the algorithms in the MLS to derive food recommendations. The user does not have any information about the exact *purpose* of the provided datasets, the *data lifecycle*, nor about who has *access* to the (possibly) un-anonymized profile or historical data and about the *data state timeline*, i.e., when the data are transferred or deleted. These aspects can be categorized as "*transparency* issues", since the user does not have the information about FoodApp's processes s/he might need or would like to have.

The FoodApp is designed in a way that on the home screen the most frequent orders for eater's automatically identified location are presented. The user can filter the suggestions using the provided *filter categories*. These categories, defined by the MLS-engineers and designers, include cuisine and menu item names, as well as the ratings of the accordant restaurants. In future interactions with the FoodApp its home screen offers the meals and food items that are most frequently ordered by the eater or users that were identified to have a similar ordering behavior, thus nudging the eater to order the same or similar kind of food [31].

All these features and filtering categories were created as a part of the data *transform* phase, i.e., the model creation and training phase. The first and most significant question in the beginning of the transformation phase is:

- Is the use of machine learning techniques, especially the resource hungry ones such as the neural networks, essential for the solving of the business problem at hand?

The FoodApp has based its business model on the data-based provision of food recommendations and the forwarding of the recommendations to the restaurants and delivery partners. Thus, being data-based, these business questions would require the use of data analysis tools, although the added value of the neuronal networks for the recommendations depends on the quality of data and the accuracy thresholds defined by the product designers.

The *model quality* is in the center of the ethical inquiry in the transformation phase. The set thresholds define mathematical methods, e.g., neural networks vs., e.g., support vector machines, and thus the resources needed to train the model as well as to generate the recommendation. The transform phase does not only include the training and optimization of the models used for the recommendation, it also considers the inclusion of the ML-models into the information systems context.

While definition of food categories as well as the selection of the included cuisines and restaurants is part of the *sense* phase and especially the *select* sub-phase, questions in the transform phase focus on the mathematical transformation of these selected details. Inclusion of, e.g., nudging techniques is also part of the sense phase and the collect sub-phase, but it is strongly defined by the *business model*. Thus, for the transform phase ethical questions could be among others:

- What categories of the collected data are included in the statistical model?
- What is the category that the model is being optimized for?
- What are the quality criteria for the results derived by the MLS?
- What are the thresholds for the quality criteria?

In the *act phase* of data processing, the MLS is integrated into the business process such as its calculated results are being used to create business value and trigger the following business step. For Acima, the value is created when the food delivery order is completed in the FoodApp. Hence, the ordering process is organized in a way that no extended explications or additional information are given so that the user does not have to choose, decide or react during the interaction process. This design allows a *fast phase-out* between opening the FoodApp and ordering the food. This effect can be expected to contribute to user satisfaction and thus re-visiting the platform for the next order.

The process efficiency offered by the FoodApp is also built on the lack of decision possibilities and a limited items selection that is based on the historic and profile preferences for the user. Additionally, the gained comfort for the user in terms of food selection and delivery has implications on the *ecosystem* of the FoodApp. The restaurant partners will be faced with the increased amount of reviews from the delivery customers, potentially forcing them to concentrate on robust packaging to ensure the sound condition of the meal for delivery. More or more robust packaging means more damage to the *environment* but potentially better ratings from the FoodApp users [32].

Furthermore, the food recommendations based on historic and similar orders might lead to *homogenization* of the food offered and prepared in the participating restaurants, as menu items that are ordered less often might not be prepared by the restaurants anymore, potentially leading to the *decreasing of skills* of the cooking staff. The individual delivery of the food orders requires reliable and efficient delivery partners. Acima relies here on its network of drivers for personal transportation that are also incentivized to transport food orders via reward programs. This efficient and effortless process of ordering food for individual consumption can and does cause significant *environmental damage* in terms of air pollution through traffic and waste [32].

Further effects on the *social environment* can also occur. The eater rates the restaurant on the food quality and the delivery partner on the quality of the delivery. The rating is based on eaters' satisfaction with the end result, whereat the traffic situation and other external effects of the recommendation process are not considered. This relationship pattern causes societal effects that are visible in the traffic situation, environmental damages as well as reduction of labor costs and conditions [34][35]. Furthermore, usage of an MLS is probable to change user's behavior [35]. The questions that can be asked in this scenario to identify potential ethical issues are:

- Is it clear for the user that his/her choice of delivery partner would lead to potential loss of jobs for other delivery partners?
- Is it clear for the user what impact her or his order deliver has on traffic or environmental indicators?
- Is it clear for the user what consequences his or her ratings of the restaurant will have on the restaurant?
- Is it clear for the user what effects his or her order will have on his or her recommendations profile?
- Is it clear for the user what are the basics for the recommendations of food/ cuisine/ delivery partner or restaurant within the application?
- Is it possible for the user to change or manage the filtering categories in the app?
- Is it possible for the user to change the profile?

In the *apply* phase, the effects of the integrated MLS are in focus. Here the consequences of the provided food recommendations based on user's historic behavior could lead to *decisional de-skilling* [7] or in this case potentially *homogeneous* food preferences for the eater. Such an automated decision support can also potentially result in the *de-skilling of the evaluation* abilities [25] for the eater in the given context.

The *apply* phase demands for user training (see Fig. 5) or to a smaller degree an explanation of the mechanisms behind the results of the application. So that the model specifications as well as the usability and settings questions can be addressed. User training is here mostly out of scope, since it needs to be implemented as an inherent feature of the FoodApp and would affect the efficiency of the ordering process.

Rating of the delivery partners results in an increasing number of orders for high ranked drivers and in a reduction of delivery orders for the worse ranked drivers. Hence promoting the reviews into the main factor for job acquisition, and thus income, for the drivers. This type of job market is known as the *gig economy* [36]. It provides income potential for the workers while creating an interdependency between the platform customer and the gig worker. This relation seems to remain unclear for the platform customer and is often debated by the platform owners [18][19]. Consequently, the OECD stated in 2016 that digital platforms need social values to be reflected in the platform governance [39].

VII. INTEGRATING THE ETHICAL ISSUES INTO THE SYSTEM DESIGN

Based on the ethical analysis of the previous section, Table 1 provides a synthesis of the identified ethical issues and the hereby affected ethical values as defined by the ALTAI checklist. Also, an example how the identified issue can be integrated into the IT system is provided.

The recommendations are structured along the following levels: business level, User Interface (UI) and system level. While business level addresses the definition of the business model and business goals, the system level considers the systems design, including the design of the algorithms. The UI aspects can be used to balance the business goal, i.e., eater's re-ordering behaviour, and the eater's interaction expectations with the digital platform.

This paradigmatic nature allows an insight into the application of the ethical analysis during the MLS design. A more detailed analysis would be needed to provide specific insights on the algorithm level.

TABLE I. ETHICAL ISSUES OF THE FOODAPP AND SUGGESTIONS FOR THEIR IMPLEMENTATION

Ethical issues	Affected value(s)	Suggestion for implementation
No explanation on the data storage	Communication; Data governance	*System/UI*: Include clear and transparent information for the user about the data storage, in e.g., in individual contracts or in general terms and conditions. *System*: Develop a concept for deletion routines if the purpose of the data processing is no longer applicable. Accordant selection of the storage location.
No explanation on the purpose of data collection	Communication, Data governance	*UI*: Provide information, e.g., via mouse hover, about the purpose of the data collected in the field, as concrete as possible *System*: To collect data that are not essential for the provision of the service, provide opt-in options by asking the user directly, e.g., "Would you like to help us to improve our service by providing your automated location data?"
Lack of an opt-out for specific data type collection	Privacy, Accountability	*System/UI*: Privacy friendly default settings, e.g., opt-in function for every data item collected instead of the implementation of the "required" fields.
Lack of the possibility to manually adjust the collected data	Human agency and oversight	*System*: Possibility to add or correct data manually, e.g., to type the address for delivery. Establish a reporting system for customers if they wish to have data corrected.
The fact that stakeholders have access to the collected data by FoodApp	Privacy and Data govvernance, Communication	*Business*: No data exchange between other stakeholders without agreements; user can be asked if s/he wants specific data to be shared for a specific purpose with the specific partner (a reimbursement could be offered) Implement an accordant opt-in or rewarding mechanisms for the user in the settings.

Data life cycle is unclear for the user	*Communication, Accountability*	*System*: Describe the data life-cycle to the user, e.g., on the FAQ page. Integrate an automated deletion routines after the needed data are collected; inform the user about the routine in the FAQs, provide opt-ins for further data collection if needed; implement a reward system for additional data collection.
Data state: Is the data anonymized before the analysis?	Privacy, Accountability	*System*: Make clear in FAQ that data is processed anonymously. If this is fulfilled, the GDPR does not apply for the processing. Implement anonymization process via e.g., distributed data bases. If anonymization is not possible, secure data by pseudonymisation and encryption.
Lack of feedback from stakeholders.	Communication, Human Agency and Oversight	*System/business*: Provide a transparent feedback system from and to every actor in the ecosystem; provide an explanation of the ratings and their effects for the actors on the FAQ. Eliminate one-sided rating mechanisms.
Lack of tracing of (e.g.,societal) changes induced by the app.	Transparency, Accountability, Societal and Environmental well-being	Business: Schedule surveys regularly with eaters, restaurants and delivery partners to assess the changes induced in those ecosystems; perform simulations to define potential changes to the traffic in the delivery area; establish contact to the traffic agency; include actionable changes suggestions, e.g., provide contact to a sustainable packaging producer for the restaurant partners; make these actions transparent on the FAQ.
Optimizing the algorithm for user re-ordering	Privacy and Data Governance, Communication	System/Business: Include other stakeholders such as restaurants, delivery partners and the environmental effects with similar weights into the recommendation algorithm; evaluate the systems on a regular basis. UI: Provide different recommendations foci for the user, e.g., focus on preferences, focus on restaurant convenience, etc.
Lack of a test phase about the effects of app usage on the society	Accountability	*Business*: Include a laboratory phase, where the app is tested by the users and stakeholders with evaluation of the UI, UX, legal and ethical aspects plus relevant simulations on the ecosystem e.g., food and restaurant landscapes before release.
Definition of the parameters for food selection by the engineers	Accountability, Privacy and Data governance, Communicatio, Transparency	*Business/system*: Include a customer survey on which categories they would like to have; change categories or filters for sorting and extend these cate-gories regularly.
Live roll-out of changes to the MLS, i.e.,online experimentation	Accountability	*Business*: Perform changes roll-out during the laboratory phase and simu-lation; when approved, roll-out for the whole community.
Usage of power resources to train the (modifications to) ML-model recommendations are based on a selection of pre-set parameters	Societal and Environmental well-being	*Business*: Change and train the model as rarely as possible, e.g., once a year. *UI*: Provide information why the recommendation was generated and what impact the change of the parameters (e.g., delivery time) would have on the results; Provide possibilities to have parameters adjusted or included into the list.
Lack of understanding of the rating mechanism	Communication, Transparency, Human agency and oversight	*UI*: Provide an explanation of the rating mechanism containing a relative comparison to other ratings, as well as potential consequences (e.g., in a dialog: "Your rating will decrease the number of suggested orders to this delivery partner by 0.2% per cent").
Tendency of the user to accept the MLS suggestions	Communication, Accountability	*UI*: Include a "surprise me" function, where a product is suggested to the eater that does not adhere to his/her top preferences; add a reminder function: "you have already ordered this meal *n* times this month. Would you like to try Y (second choice) today instead?". System: Perform an assessment on how ML might impact user behaviour and present the results on the website.
Effects on the eco-system of the app are not clear for the user	Societal and Environmental well-being, Communication	*Business*: Make the results of the conducted surveys and traffic analysis accessible to the users on the website. *System*: Carry out an impact assessment on the rights of users and also on those of the stakeholders.
Individual food delivery	Human Agency and Oversight	*System*: Include environmental concerns into the algorithm evaluation; *Business*: Provide rewards for environmentally friendly behaviour of the partners (using e-vehicles, e.g., or using environmentally friendly packaging).

Recommendations presentation to optimize the business goal	*Accountability*, *Transparency*	*UI*: Change the UI to be more intuitive for the user with the goal of finding favourite food selection.

The suggestions provided in Table 1 are centred on mainly two aspects: providing information about every data element collected by the FoodApp, i.e., the facet of transparency, and establishing a reward system for the user in return to providing data to the company, i.e., a reward. The implementation of a reward system would implicitly make the data life-cycle more transparent for the user, as well as provide the user with more autonomy within the engagement with the service. It would also help the user to understand that the data is a resource that is traded and thus has a value.

Identified issues that go beyond the business processes might be subject to the interpretation of the regulation or the business ethics. Furthermore, due to the context of the example, some identified ethical aspects are due to the example being positioned in the platform economy and therefore are not specific for every MLS. Nevertheless, bigger negative effects such as the effects on the environment or the society are part of the social awareness and responsibility that are not (and maybe should not be) regulated, but can be supported by socially acceptable IT artefacts.

Therefore, the term of socially acceptable IT has been introduced in [1] to describe a system that considers and integrates ethical requirements into its design. The added effort but also the value of the implementation of the suggestions of the ethical considerations in Table 1 could lead to socially acceptable IT products and thus a realization of a socio-technical IT systems. To ensure the remaining and homogeneous quality adherence, inter-company assessment mechanisms, i.e., ethical quality audits, could be put in place.

VIII. CONCLUSION

Here a scenario of a fictional food ordering platform that uses a MLS for item recommendations was used to perform an ethical analysis of an MLS. This scenario was chosen as a realisation of a socio-technical system that incorporates system designers, users and stakeholders affected by the system design and process implementation.

The results showed that users of digital services need to be integrated into the design of a socio-technical system as they may have expectations and values that rely on the ethical awareness of the company and thus need to be implemented into the workflow. The examples of how to address these issues demonstrated that changes in the UI, system design but also in the business model can be realistically made to accommodate these challenges. Hence, designing socially acceptable socio-technical IT systems can be a chance to find a niche on the growing and competitive market of consumer-oriented digital services.

Although, the provided approach needs validation and verification in a rea-life environment, it can already be used by the designers and architects of information systems, business developers considering a data-based business model, as well as ISR scientists as it shows how ethical aspects can be incorporated into the context of IT design.

Therefore, future work will aim at establishing the criteria for the definition of the quality requirements for the social acceptable IT, evaluation of the suggested measures, as well as developing methods for the assessment of the effort of their implementation.

REFERENCES

[1] O. Levina, “Towards Implementation of Ethical Issues into the Recommender Systems Design,” in *ICCGI 2021, The Sixteenth International Multi-Conference on Computing in the Global Information Technology*, 2021, pp. 6–11, [Online]. Available: http://thinkmind.org/index.php?view=article&articleid=iccgi_2021_1_20_18002.

[2] L. Floridi and M. Taddeo, “What is data ethics?,” *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, vol. 374, no. 2083. The Royal Society, Dec. 28, 2016, doi: 10.1098/rsta.2016.0360.

[3] Comission on data ethics: "An assessment of the comission on the data ethics" (in German: Datenethikkommission, “Gutachten der Datenethikkommission,”), https://www.bmi.bund.de/SharedDocs/downloads/DE/publikationen/themen/it-digitalpolitik/gutachten-datenethikkommission.pdf?__blob=publicationFile&v=4 (Accessed, July 12 2021), 2018.

[4] G. Baxter and I. Sommerville, “Socio-technical systems: From design methods to systems engineering,” *Interact. Comput.*, vol. 23, no. 1, pp. 4–17, 2011, doi: 10.1016/j.intcom.2010.07.003.

[5] W. L. Robison, *Ethics Within Engineering: An Introduction*. Bloomsbury Academic, 2016.

[6] V. Dignum, “Responsible Artificial Intelligence: Ethical Thinking by and about AI,” 2019. Accessed: Oct. 07, 2019. [Online]. Available: https://icps.gwu.edu/sites/g/files/zaxdzs1736/f/downloads/Virginia%20Dignum_%20Responsible%20Artificial%20Intelligence%20(1).pdf.

[7] L. Floridi and M. Taddeo, “What is data ethics?,” *Philos. Trans. A. Math. Phys. Eng. Sci.*, vol. 374, no. 2083, 2016, doi: 10.1098/rsta.2016.0360.

[8] Doteveryone, “People, Power and Technology: The 2020 Digital Attitudes Report,” 2020. Accessed: May 13, 2020. [Online]. Available: https://www.doteveryone.org.uk/2020/05/people-power-and-technology-the-2020-digital-attitudes-report/.

[9] European Commission, “Assessment List for Trustworthy Artificial Intelligence (ALTAI) for self-assessment | Shaping Europe’s digital future,” 2020. https://ec.europa.eu/digital-single-market/en/news/assessment-list-trustworthy-artificial-intelligence-altai-self-assessment (accessed Feb. 07, 2021).

[10] D. Paraschakis, “Towards an ethical recommendation framework,” 2017, doi: 10.1109/RCIS.2017.7956539.

[11] D. Paraschakis, “Recommender Systems from an Industrial and Ethical Perspective,” in *10th ACM Conference on*

Recommender Systems - RecSys '16, 2016, pp. 463–466, doi: 10.1145/2959100.2959101.

[12] S. Milano, M. Taddeo, and L. Floridi, "Recommender Systems and their Ethical Challenges," *Minds Mach.*, vol. 2, pp. 187–191, 2019, Accessed: Aug. 29, 2019. [Online]. Available: https://philpapers.org/archive/MILRSA-3.pdf.

[13] A. Polonioli, "The ethics of scientific recommender systems," *Scientometrics*. Springer Science and Business Media B.V., pp. 1–8, Oct. 29, 2020, doi: 10.1007/s11192-020-03766-1.

[14] D. Karpati, A. Najjar, and D. Agustin Ambrossio, "Ethics of Food Recommender Applications," 2020, doi: 10.1145/3375627.3375874.

[15] O. Levina, "A Research Commentary- Integrating Ethical Issues into the Data Process," 2020, [Online]. Available: https://library.gito.de/open-access-pdf/Z3-EMoWI_2020_paper_3.pdf.

[16] A. Coravos, I. Chen, A. Gordhandas, and A. D. Stern, "We should treat algorithms like prescription drugs," *Quartz*, pp. 1–8, Apr. 2019, Accessed: Jan. 08, 2020. [Online]. Available: https://digital.hbs.edu/artificial-intelligence-machine-learning/we-should-treat-algorithms-like-prescription-drugs/.

[17] R. Schutt and C. O'Neil, *Doing Data Science- Straight Talk from the Frontline*. O'Reilly Media, 2013, p. 51.

[18] O. Levina, "AI and ethics in medical research. AI-based research in medicine and its ethical challenges," *gesundhyte.de*, pp. 8–11, 2020.

[19] D. Susser and V. Grimaldi, "Measuring Automated Influence: Between Empirical Evidence and Ethical Values," 2021.

[20] C. M. Gray, Y. Kou, B. Battles, J. Hoggatt, and A. L. Toombs, "The Dark (Patterns) Side of UX Design," *Proc. 2018 CHI Conf. Hum. Factors Comput. Syst.*, 2018, doi: 10.1145/3173574.

[21] G. Baldini, M. Botterman, R. Neisse, and M. Tallacchini, "Ethical Design in the Internet of Things," *Sci. Eng. Ethics*, vol. 24, no. 3, pp. 905–925, Jun. 2018, doi: 10.1007/s11948-016-9754-5.

[22] M. Sullivan, "Tech-industry AI is getting dangerously homogenized, say Stanford experts," *FastCompany*, 2021.

[23] Y. Wang, Y. Ning, I. Liu, and X. X. Zhang, "Food Discovery with Uber Eats: Recommending for the Marketplace | Uber Engineering Blog," *Uber Engineering*, 2018. https://eng.uber.com/uber-eats-recommending-marketplace/ (accessed Mar. 17, 2020).

[24] J. Stray, S. Adler, and D. Hadfield-Menell, "What are you optimizing for? Aligning Recommender Systems with Human Values," 2020, [Online]. Available: https://participatoryml.github.io/papers/2020/42.pdf.

[25] S. Vallor, "Moral Deskilling and Upskilling in a New Machine Age: Reflections on the Ambiguous Future of Character," *Philos. Technol.*, vol. 28, no. 1, pp. 107–124, 2015, doi: 10.1007/s13347-014-0156-9.

[26] X. Zhao, M. Maimaiti, M. Jia, Y. Ru, and S. Zhu, "How we eat determines what we become: opportunities and challenges brought by food delivery industry in a changing world in China," *Artic. Eur. J. Clin. Nutr.*, vol. 72, pp. 1282–1286, 2018, doi: 10.1038/s41430-018-0191-1.

[27] O. Levina and M. Behrend, "Assessing the Environmental Impact: A Case of Business Process Analysis in the Automotive Industry," 2016.

[28] A. Rao, F. Schaub, N. Sadeh, A. Acquisti, and R. Kang, *Expecting the Unexpected: Understanding Mismatched Privacy Expectations Online*. 2016.

[29] M. Hatamian, A. Kitkowska, J. Korunovska, and S. Kirrane, ""It's shocking!": Analysing the impact and reactions to the A3: Android apps behaviour analyser," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, Jul. 2018, vol. 10980 LNCS, pp. 198–215, doi: 10.1007/978-3-319-95729-6_13.

[30] O. Levina, "Digital Platforms and Digital Inequality-An Analysis from Information Ethics Perspective," 2019, Weizenbaum Conference, doi: https://doi.org/10.34669/wi.cp/2.4.

[31] R. Zhou, S. Khemmarat, and L. Gao, "The impact of YouTube recommendation system on video views," in *Proceedings of the ACM SIGCOMM Internet Measurement Conference, IMC*, 2010, pp. 404–410, doi: 10.1145/1879141.1879193.

[32] R. Zhong and C. Zhang, "Food Delivery apps are drowning china in plastic," *New York Times*, pp. 1–8, 2019, Accessed: Apr. 25, 2020. [Online]. Available: https://www.nytimes.com/2019/05/28/technology/china-food-delivery-trash.html.

[33] K. Griesbach, A. Reich, L. Elliott-Negri, and R. Milkman, "Algorithmic Control in Platform Food Delivery Work," *Socius Sociol. Res. a Dyn. World*, vol. 5, p. 237802311987004, Jan. 2019, doi: 10.1177/2378023119870041.

[34] V. Dorner, O. Ivanova, and M. Scholz, "Think twice before you buy! how recommendations affect three-stage purchase decision processes," in *International Conference on Information Systems (ICIS 2013): Reshaping Society Through Information Systems Design*, 2013, vol. 5, pp. 4278–4297.

[35] M. De-Arteaga, R. Fogliato, and A. Chouldechova, "A Case for Humans-in-the-Loop: Decisions in the Presence of Erroneous Algorithmic Scores," *Conf. Hum. Factors Comput. Syst. - Proc.*, Apr. 2020, doi: 10.1145/3313831.3376638.

[36] G. Friedman, "Workers without employers: Shadow corporations and the rise of the gig economy," *Rev. Keynes. Econ.*, vol. 2, no. 2, pp. 171–188, Apr. 2014, doi: 10.4337/roke.2014.02.03.

[37] J. Staufenberg, "Deliveroo courier strike: Employers cannot 'simply opt out of the National Living Wage', says Government," *The Independent*, 2016.

[38] M. Dewhurst, "Deliveroo couriers are right to strike: the company's claims of freedom are a sham," *The Guardian*, 2016. https://www.theguardian.com/commentisfree/2016/aug/16/deliveroo-couriers-strike-freedom-william-shu (accessed May 11, 2020).

[39] OECD, "Protecting Consumers In Peer Platform Markets: Exploring The Issues Background report for Ministerial Panel 3.1," Nov. 2016.

Deep Reinforcement Learning for Spatial Motion Planning in 3D Urban Environments

Oren Gal and Yerach Doytsher
Mapping and Geo-information Engineering
Technion - Israel Institute of Technology
Haifa, Israel
e-mails: {orengal@alumni.technion.ac.il, doytsher@technion.ac.il}

Abstract—**In this paper, we present spatial motion planner in 3D environments based on Deep Reinforcement Learning (DRL) algorithms. We tackle 3D motion planning problem by using Deep Reinforcement Learning (DRL) approach, which learns agent's and environment constraints. Spatial analysis focuses on visibility analysis in 3D setting an optimal motion primitive considering agent's dynamic model based on fast and exact visibility analysis for each motion primitives. Based on optimized reward function, which consist of generated 3D visibility analysis and obstacle avoidance trajectories, we introduce DRL formulation, which learns the value function of the planner and generates an optimal spatial visibility trajectory. We demonstrate our planner in simulations for Unmanned Aerial Vehicles (UAV) in 3D urban environments. Our spatial analysis is based on a fast and exact spatial visibility analysis of the 3D visibility problem from a viewpoint in 3D urban environments. We present DRL architecture generating the most visible trajectory in a known 3D urban environment model, as time-optimal one with obstacle avoidance capability.**

Keywords - Deep Reinforcement Learning; Visibility; 3D; Spatial analysis; Motion Planning.

I. INTRODUCTION AND RELATED WORK

Spatial clustering in urban environments is a new spatial field from trajectory planning aspects [1]. The motion and trajectory planning fields have been extensively studied over the last two decades [2][4][6]. The main effort has focused on finding a collision-free path in static or dynamic environments, i.e., in moving or static obstacles, using roadmap, cell decomposition, and potential field methods [11].

The path-planning problem becomes an NP-hard one, even for simple cases such as time-optimal trajectories for a system with point-mass dynamics and bounded velocity and acceleration with polyhedral obstacles [7].

Path planning algorithms can be distinguished as local and global planners. The local planner generates one, or a few, steps at every time step, whereas the global planner uses a global search to the goal over a time-spanned tree. Examples of local (reactive) planners are [9][14]. These planners are too slow, do not guarantee safety and neglect spatial aspects.

Efficient solutions for an approximated problem were investigated by LaValle and Kuffner, addressing non-holonomic constraints by using the Rapidly Random Trees (RRT) method [15][16]. Over the years, many other semi-randomized methods were proposed, using evolutionary programming [5][18].

The randomized sampling algorithms planner, such as RRT, explores the action space stochastically. The RRT algorithm is probabilistically complete, but not asymptotically optimal [13]. The RRT* planner challenges optimality by a rewiring process each time a node is added to the tree. However, in cluttered environments, RRT* may behave poorly since it spends too much time deciding whether to rewire or not.

Overall, only a few works have focused on spatial analysis characters integrated into trajectory planning methods such as visibility analysis or spatial clustering methods [11].

Analyzing pedestrian's mobility from a spatial point of view mainly focused on route choice [3], simulation model [19] and agent-based modeling [12].

The efficient computation of visible surfaces and volumes in 3D environments is not a trivial task. The visibility problem has been extensively studied over the last twenty years, due to the importance of visibility in GIS and Geomatics, computer graphics and computer vision, and robotics. Accurate visibility computation in 3D environments is a very complicated task demanding a high computational effort, which could hardly have been done in a very short time using traditional well-known visibility methods.

The exact visibility methods are highly complex, and cannot be used for fast applications due to their long computation time. Previous research in visibility computation has been devoted to open environments using Digital Elevation Model (DEM), representing raster data in 2.5D (Polyhedral model), and do not address, or suggest solutions for dense built-up areas.

Most of these works have focused on approximate visibility computation, enabling fast results using interpolations of visibility values between points, calculating

point visibility with the Line of Sight (LOS) method [7]. Lately, fast and accurate visibility analysis computation in 3D environments has been presented [10].

In this paper, we present unique spatial trajectory planning method based on DRL algorithm based on exact visibility analysis in urban environment. The generated trajectories are based on visibility motion primitives as part of the planned trajectory, which takes into account exact 3D visible volumes analysis clustering in urban environments.

The proposed planner includes obstacle avoidance capabilities, satisfying dynamics' and kinematics' agent model constraints in 3D environments, using Velocity Obstacles (VO) in 3D for Unmanned Aerial Vehicle (UAV) model.

In the following sections, we first introduce the DRL algorithm and method and our extension for a spatial analysis case, such as 3D visibility. Later on, we present the our planner, using VO method and planner model. In the last part of the paper, with planner simulation using DRL method.

II. PROBLEM STATEMENT

We consider the basic visibility problem in a 3D urban environment, consisting of 3D buildings modeled as 3D cubic parameterization $\sum_{i=1}^{N} C_i(x, y, z = {}^{h_{\max}}_{h_{\min}})$, and viewpoint $V(x_0, y_0, z_0)$.

Given:

- Parameterizations of N objects $\sum_{i=1}^{N} C_i(x, y, z = {}^{h_{\max}}_{h_{\min}})$ describing a 3D urban environment model

Compute:

- *Trajectory*, which consist of optimal set of all visible points, i.e., most visible points of $\sum_{i=1}^{N} C_i(x, y, z = {}^{h_{\max}}_{h_{\min}})$, from starting point ,$q_s$, to the goal, q_g, without collision.

This problem seems to be solved by conventional geometric methods, but as mentioned before, it demands a long computation time. We introduce a fast and efficient computation solution for a schematic structure of an urban environment that demonstrates our method based on Deep Reinforcement Learning method (DRL).

On the first part, we present the DRL algorithm, formulated to our planning problem, and the visibility analysis along with obstacles avoidance planner.

III. DEEP REINFORCEMENT LEARNING (DRL) ALGORITHM

In most Deep Reinforcement Learning (DRL) systems, the state is basically agent's observation of the environment. At any given state the agent chooses its action according to a policy. Hence, a policy is a road map for the agent, which determines the action to take at each state. Once the agent takes an action, the environment returns the new state and the immediate reward. Then, the agent uses this information, together with the discount factor to update its internal understanding of the environment, which, in our case, is accomplished by updating a value function. Most methods are using the use well-known simple and efficient greedy exploration method maximizing Q-value.

In case of velocity planning space as part of spatial analysis planning, each possible action is a possible velocity in the next time step, which also represent a viewpoint. The Q-value function is based on greedy search velocity, with greedy local search method. Based on that, TD and SARSA methods for DRL can be used, generating visible trajectory in 3D urban environment.

A. Markov Decision Processes (MDP)

The standard Reinforcement Learning set-up can be described as a MDP as can be seen in Figure 1, consisting of:

- **A finite set of states** *S*, comprising all possible representations of the environment.
- **A finite set of actions** *A*, containing all possible actions available to the agent at any given time.
- **A reward function** $R = \psi(s_t, a_t, s_{t+1})$, determining the immediate reward of performing an action at from a state s_t, resulting in s_{t+1}.
- **A transition model** $T(s_t, a_t, s_{t+1}) = p(s_{t+1} | s_t, a_t)$, describing the probability of transition between states s_t and $s_{t+1}$when performing an action a_t.

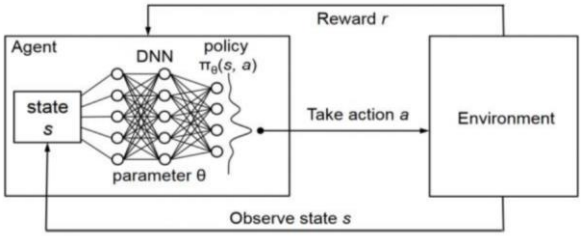


Figure 1. Standard Reinforcement Learning Methology

B. Temporal Difference Learning

Temporal-difference learning (or TD) interpolates ideas from Dynamic Programming (DP) and Monte Carlo methods. TD algorithms can learn directly from raw experiences without any model of the environment.

Whether in Monte Carlo methods, an episode needs to reach completion to update a value function, Temporal-difference learning can learn (update) the value function within each experience (or step). The price paid for being

able to regularly change the value function is the need to update estimations based on other learned estimations (recalling DP ideas). Whereas in DP a model of the environment's dynamic is needed, both Monte Carlo and TD approaches are more suitable for uncertain and unpredictable tasks.

Since TD learns from every transition (state, reward, action, next state, next reward) there is no need to ignore/discount some episodes as in Monte Carlo algorithms.

C. Spatial Planning Using DRL

In this section, we present DRL approach based on the proposed spatial planning method. It considers that the value function *f* related to each point *x*. The spatial planner seeks to obtain the trajectory *T** that based on visibility motion primitives set as part of the planned trajectory, which takes into account exact 3D visible volumes analysis clustering in urban environments, based on optimizing value function *f* along *T*.

The generated trajectories are then represented by a set of discrete configuration points:

$$T = \{x_1, x_2, \cdots, x_N\} \quad (1)$$

Without loss of generality, we can assume that the value function for each point can be expressed as a linear combination of a set of sub-value functions, that will be called features $c(x) = \sum c_j f_j(x)$. The cost of path *T* is then the sum of the cost for all points in the path. Particularly, in the Velocity Obstacles as will be presented later on, the value is the sum of the sub-values of moving between pairs of states in the path:

$$c(\zeta) = \sum_{i=1}^{N-1} c(x_i, x_{i+1}) = \sum_{i=1}^{N-1} \frac{c(x_i) + c(x_{i+1})}{2} \|x_{i+1} - x_i\| = \omega^T \sum_{i=1}^{N-1} \frac{f(x_i) + f(x_{i+1})}{2} \|x_{i+1} - x_i\| = \omega^T f(\zeta) \quad (2)$$

Based on number of demonstration trajectories D, $D = \{\zeta_1, \zeta_2, \cdots, \zeta_D\}$, by using DRL, weights ω can be set for learning from demonstrations and setting similar planning behavior. As was shown by [23,24], this similarity is achieved when the expected value of the features for the trajectories generated by the planner is the same as the expected value of the features for the given demonstrated trajectories:

$$\mathbb{E}(f(\zeta)) = \frac{1}{D} \sum_{i=1}^{D} f(\zeta_i) \quad (3)$$

Applying the Maximum Entropy Principle [25] to the DRL problem leads to the following form for the probability density for the trajectories returned by the demonstrator:

$$p(\zeta|\omega) = \frac{1}{Z(\omega)} e^{-\omega^T f(\zeta)} \quad (4)$$

$Z(\omega)$ is a normalization function that does not depend on ζ. One way to determine ω is maximizing the log-likelihood of the demonstrated trajectories under the previous model:

$$L(D|\omega) = -D\log(Z(\omega)) \quad +\sum_{i=1}^{D} (-w^T f(\zeta_i)) \quad (5)$$

The gradient of the previous log-likelihood with respect to ω is given by:

$$\nabla \mathcal{L} = \frac{\partial \mathcal{L}(\mathcal{D}|\omega)}{\partial \omega} = \mathbb{E}(f(\zeta)) - \frac{1}{D} \sum_{i=1}^{D} f(\zeta_i) \quad (6)$$

As mentioned in [23], this gradient can be intuitively explained. If the value of one of the features for the trajectories returned by the planner are higher from the value in the demonstrated trajectories, the corresponding weight should be increased to increase the value of those trajectories.

The main problem with the computation of the previous gradient is that it requires to compute the expected value of the features $E(f(\zeta))$ for the generative distribution (4).

We suggest setting large amount of D cased, setting the relative *w* values for our planner characters.

TABLE I. DRL PLANNER PSEUDO CODE

DRL Planner
Setting Trajectory S Examples D, D= T.init* (x_{init});
Calculate function features Weight, w
$f_D \leftarrow$ *AverageFeatureCount(D)*;
$w \leftarrow$ *random_init()*;
Repeat
  for each T do*
  for VelocityObstacles_repetitions do
    $\zeta_i \leftarrow$ *getVOstarPath(T*,ω)*
    $f(\zeta_i) \leftarrow$ *calculeFeatureCounts*(ζ_i)
  end for
  f_{VO} *(T*)*$\leftarrow \sum_{i=1}^{VO\_repetitions} f(\zeta_i))/VO\_repetitions$
  end for
  $f_{VO} \leftarrow (\sum_{i=1}^{S} f_{VO})/s$
  $\nabla L \leftarrow f_{VO} - f_D$
  $w \leftarrow$ *UpdatedWeigths* (∇L)
Until convergence
Return w

IV. UAV MODEL

We introduce an Unmanned Aerial Vehicle (UAV) model, based on the well-known simple car and Dubins airplane [26]. Dubins airplane [27] model extends Dubins

car model with continuous change of altitude without reverse gear, avoiding sudden altitude speed rate variation. Our UAV model includes kinematic and dynamic constraints which ignore pitch and roll rotation or winds disturbances.

A. Kinematic Constraints

We use a simple UAV model with four dimensions, each configuration is $q=(x,y,z,\theta)$, when x,y,z are the coordinates of the origin, and θ is the orientation, in x-y plane relative to x-axis, as can be seen in Figure 2 for a simple car-like model.

The steering angle is denoted as ϕ . The distance between front and rear axles is equal to *L*. The kinematic equations of a simple UAV model can be written as:

$$\begin{aligned} \dot{x} &= u_s \cos\theta, \\ \dot{y} &= u_s \sin\theta, \\ \dot{z} &= u_{z,} \\ \dot{\theta} &= u_s \tan u_\phi \end{aligned} \tag{7}$$

Where u_s is the speed parallel to x-y plane, climb rate (speed parallel to z-axis) is u_z and the control on steering angle u_ϕ. We denote the control vector as $u=(u_s,u_z,u_\phi)$. Each of the controllers is bounded, $u_\phi \in [-\phi^{\max},\phi^{\max}]$ where $\phi^{\max}<\pi/2$, the speed $u_s \in [u_s^{\min},u_s^{\max}]$ and climb rate $u_z \in [-u_z^{\max},u_z^{\max}]$. $u_s^{\min}>0$, so UAV cannot stop.

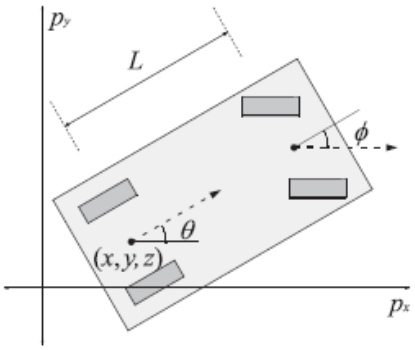


Figure 2. The Simple Car Model. The z-axis can be changed for a Simple -Airplane (Source [26])

B. Dynamic Constraints

The UAV model has to take into account the dynamic constraints, preventing instantaneous changes (increase or decrease) of the control vector $u=(u_s,u_z,u_\phi)$.

UAV model also includes dynamic constraints, $\dot{u}_s \in [-a_s,a_s], \dot{u}_z \in [-a_z,a_z]$ and $\dot{u}_\phi \in [-a_\phi,a_\phi]$.

V. Analytic Visibility Computation

A. Analytic Solution for a Single Object

In this section, we first introduce the visibility solution from a single point to a single 3D object. This solution is based on an analytic expression, which significantly improves time computation by generating the visibility boundary of the object without the need to scan the entire object's points.

Our analytic solution for a 3D building model is an extension of the visibility chart in 2D introduced by Elber et al. [26] for continuous curves. For such a curve, the silhouette points, i.e., the visibility boundary of the object, can be seen in Figure 3:

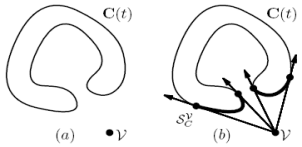


Figure 3. Visible Silhouette Points $S_C{}^V$ from viewpoint *V* to curve *C(t)* (source: [26]).

The visibility chart solution was originally developed for dealing with the Art Gallery Problem for infinite viewpoint; it is limited to 2D continuous curves using multivariate solver [26], and cannot be used for on-line application in a 3D environment.

Based on this concept, we define the visibility problem in a 3D environment for more complex objects as:

$$C'(x,y)_{z_{const}} \times (C(x,y)_{z_{const}} - V(x_0,y_0,z_0)) = 0 \tag{8}$$

3D model parameterization is $C(x,y)_{z_{const}}$, and the viewpoint is given as $V(x_0,y_0,z_0)$. Solutions to equation (8) generate a visibility boundary from the viewpoint to an object, based on basic relations between viewing directions from *V* to $C(x,y)_{z_{const}}$ using cross-product characters.

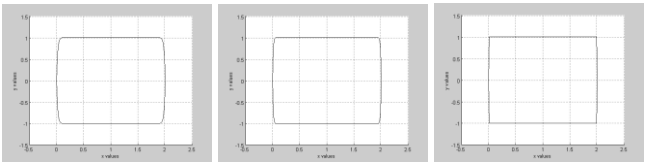
A three-dimension urban environment consists mainly of rectangular buildings, which can hardly be modeled as continuous curves. Moreover, an analytic solution for a single 3D model becomes more complicated due to the higher dimension of the problem and is not always possible. Object parameterization is therefore a critical issue, allowing us to find an analytic solution and, using that, to generate the visibility boundary very fast.

1) 3D Building Model: Most of the common 3D City Models are based on object-oriented topologies, such as 3D Formal Data Structure (3D FDS), Simplified Spatial Model (SSS) and Urban Data Model (UDM) [26]. These models are very efficient for web-oriented applications. However, the fact that a building consists of several different basic

features makes it almost impossible to generate analytic representation. A three-dimension building model should be, on the one hand, simple enabling analytic solution, and on the other hand, as accurate as possible. We examined several building object parameterizations, and the preferred candidate was an extended n order sphere coordinates parameterization, even though such a model is a very complex, and will necessitate a special analytic solution. We introduce a model that can be used for analytic solution of the current problem. The basic building model can be described as:

$$x = t, y = \begin{pmatrix} x^n - 1 \\ 1 - x^n \end{pmatrix}, z = c \qquad (9)$$
$$-1 \le t \le 1, n = 350, c = c + 1$$

This mathematical model approximates building corners, not as singular points, but as continuous curves. This building model is described by equation (9), with the lower order badly approximating the building corners, as depicted in Figure 4. Corner approximation becomes more accurate using *n=350* or higher. This approximation enables us to define an analytic solution to the problem.



(a) (b) (c)

Figure 4. Topside view of the building model using equation (2) - (a) n=50; (b) *n=200*; (c) *n=350*.

We introduce the basic building structure that can be rotated and extracted using simple matrix operators (Figure 4). Using a rotation matrix does not affect our visibility algorithm, and for simple demonstration of our method we present samples of parallel buildings.

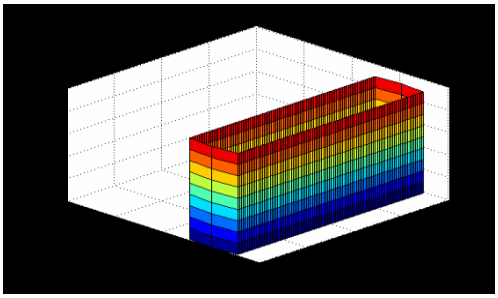


Figure 5. A Three-dimension Analytic Building Model with Equation (8), where $z_{h_{\min}=0}^{h_{\max}=9}$

2) Analytic Solution for a Single Building: In this part we demonstrate the analytic solution for a single 3D building model. As mentioned above, we should integrate building model parameterization to the visibility statement. After integrating eqs. (8) and (9):

$$\begin{aligned} & C'(x,y)_{z_{const}} \times (C(x,y)_{z_{const}} - V(x_0, y_0, z_0)) = 0 \rightarrow \\ & x^n - V_{y_0} - n \cdot x^{n-1}(x - V_{x_0}) - 1 = 0 \\ & x^n + V_{y_0} - n \cdot x^{n-1}(x - V_{x_0}) - 1 = 0 \\ & n = 350, -1 \le x \le 1 \end{aligned} \qquad (10)$$

where the visibility boundary is the solution for these coupled equations. As can be noticed, these equations are not related to Z axis, and the visibility boundary points are the same ones for each x-y surface due to the model's characteristics. Later on, we treat the relations between a building's roof and visibility height in our visibility algorithm, as part of the visibility computation.

The visibility statement leads to two polynomial *N* order equations, which appear to be a complex computational task. The real roots of these polynomial equations are the solution to the visibility boundary. These equations can be solved efficiently by finding where the polynomial equation changes its sign and cross zero value; generating the real roots in a very short time computation (these functions are available in Matlab, Maple and other mathematical programs languages). Based on the polynomial cross zero solution, we can compute a fast and exact analytic solution for the visibility problem from a viewpoint to a 3D building model. This solution allows us to easily define the Visible Boundary Points.

Visible Boundary Points (VBP) - we define VBP of the object *i* as a set of boundary points $j=1..N_{bound}$ of the visible surfaces of the object, from viewpoint $V(x_0, y_0, z_0)$.

$$VBP_{i=1}^{j=1..N_{bound}}(x_0, y_0, z_0) = \begin{bmatrix} x_1, y_1, z_1 \\ x_2, y_2, z_2 \\ .. \\ x_{N_{bound}}, y_{N_{bound}}, z_{N_{bound}} \end{bmatrix} \qquad (11)$$

Roof Visibility – The analytic solution in equation (10) does not treat the roof visibility of a building. We simply check if viewpoint height $V(z_0)$ is lower or higher than the building height $h_{\max_{C_i}}$ and use this to decide if the roof is visible or not:

$$V_{z_0} \ge Z = h_{\max_{C_i}} \qquad (12)$$

If the roof is visible, roof surface boundary points are added to VBP. Roof visibility is an integral part of VBP computation for each building.

Two simple cases using the analytic solution from a visibility point to a building can be seen in Figure 6. The

visibility point is marked in black, the visible parts colored in red, and the invisible parts colored in blue. The visible volumes are computed immediately with very low computation effort, without scanning all the model's points, as is necessary in LOS-based methods for such a case.

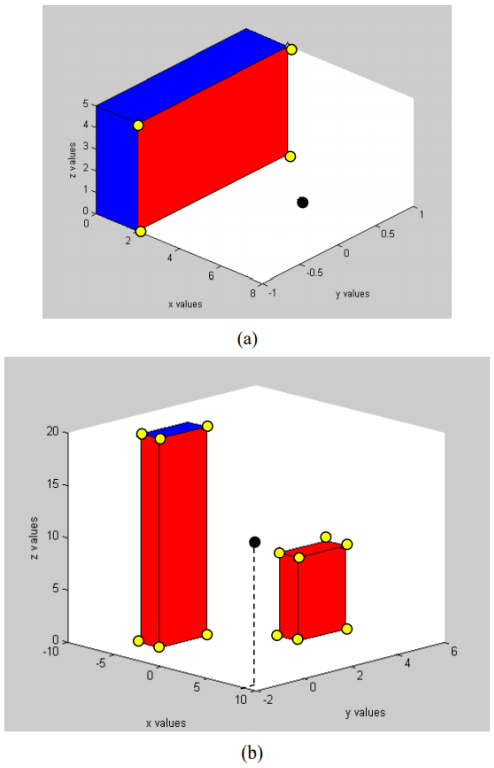


Figure 6. Visibility Volume computed with the Analytic Solution. Viewpoint is marked in black, visible parts colored in red, and invisible parts colored in blue. VBP marked with yellow circles - (a) single building; (b) two non-overlapping buildings.

B. Visibility Computation in Urban Environments

In the previous sections, we treated a single building case, without considering hidden surfaces between buildings, i.e., building surface occluded by other buildings, which directly affect the visibility volumes solution. In this section, we introduce our concept for dealing with these spatial relations between buildings, based on our ability to rapidly compute visibility volume for a single building generating VBP set.

Hidden surfaces between buildings are simply computed based on intersections of the visible volumes for each object. The visible volumes are defined easily using VBP, and are defined, in our case, as Visible Pyramids. The invisible components of the far building are computed by intersecting the projection of the closer buildings' VP base to the far building's VP base.

1) The Visible Pyramid (VP): we define $VP_i^{j=1..Nsurf}(x_0, y_0, z_0)$ of the object i as a 3D pyramid generated by connecting VBP of specific surface j to a viewpoint $V(x_0, y_0, z_0)$. Maximum number of N_{surf} for a single object is three. VP boundary, colored with green arrows, can be seen in Figure 6. The intersection of VPs allows us to efficiently compute the hidden surfaces in urban environments, as can be seen in the next sub-section.

2) Hidden Surfaces between Buildings: As we mentioned earlier, invisible parts of the far buildings are computed by intersecting the projection of the closer buildings' VP to the far buildings' VP base.

For simplicity, we demonstrate the method with two buildings from a viewpoint $V(x_0, y_0, z_0)$ one (denoted as the first one) of which hides, fully or partially, the other (the second one).

As can be seen in Figure 7, in this case, we first compute VBP for each building separately, $VBP_1^{1..4}$, $VBP_2^{1..4}$, based on these VBPs, we generate VPs for each building, VP_1^1, VP_2^1. After that, we project VP_1^1 base to VP_2^1 base plane, as seen in Figure 8, if existing. At this point, we intersect the projected surface in VP_2^1 base plane and update $VBP_2^{1..4}$ and VP_2^1 (decreasing the intersected part).

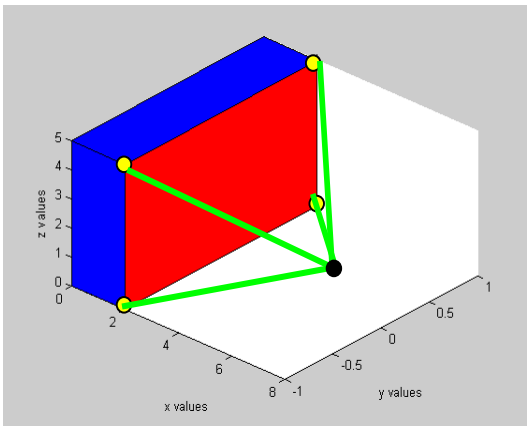


Figure 7. A Visible Pyramid from a viewpoint (marked as a black point) to VBP of a specific surface

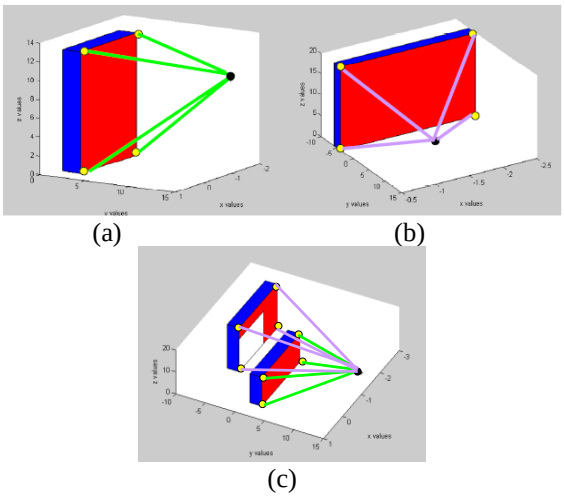


Figure 8. Generating VP - (a) VP_1^1 boundary colored in green arrows; (b) VP_2^1 boundary colored in purple lines; (c) the two buildings - VP_1^1in green and VP_2^1 in purple, from the viewpoint.

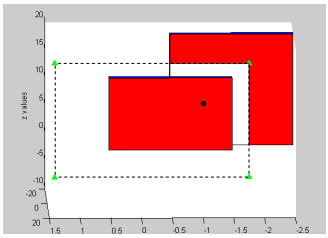


Figure 9. Projection of VP_1^1 to VP_2^1 base plane marked with dotted lines.

The intersected part is the invisible part of the second building from viewpoint $V(x_0, y_0, z_0)$ hidden by the first building, which is marked in white in Figure 9.

In the case of a third building, in addition to the buildings introduced in Figure 9, the projected VP will only be the visible ones, and the VBP and VP of the second building will be updated accordingly.

We demonstrated a simple case of an occluded building. A general algorithm for more a complex scenario, which contains the same actions between all the combinations of VP between the objects, is detailed in the next sub-section. Projection and intersection of 3D pyramids can be done with simple computational geometry elements, which demand a very low computation effort.

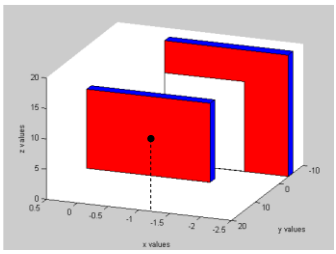


Figure 10. Computing Hidden Surfaces between Buildings by using the Visible Pyramid Colored in White on VP_2^1 Base Plane.

C. Viewpoint Invisibility Value

Planning UAVs visible trajectory is based on the ability to accumulate the visibility value of each viewpoint explored as part of the planner algorithm. We calculate the exact invisible value of a specific viewpoint, i.e., the total sum of the invisible surfaces and roofs from viewpoint.

We divide point invisibility value into Invisible Surfaces Value (ISV) and Invisible Roofs Value (IRV). This classification allows us to plan delicate and accurate trajectory upon demand. We define ISV and IRS as the total sum of the invisible roofs and surfaces (respectively).

Invisible Surfaces Value (ISV) of a viewpoint is defined as the total sum of the invisible surfaces of all the objects in a 3D environment, as described in equation (13):

$$ISV(x_0, y_0, z_0) = \sum_{i=1}^{N_{obj}} IS_{VP_i^{j=1..N_{bound}-1}}^{VP_i^{j=1..N_{bound}-1}} \quad (13)$$

In the same way, we define Invisible Roofs Value (IRV) value as the total sum of all the invisible roofs surfaces:

$$IRV(x_0, y_0, z_0) = \sum_{i=1}^{N_{obj}} IS_{VP_i^{j=N_{bound}}}^{VP_i^{j=N_{bound}}} \quad (14)$$

VI. Deep Reinforcement Learning (DRL) Planner

Our planner, as described in Table 1, based on DRL method, generate visible sequence of optimal-visible waypoints as a candidate trajectory. We extend previous planners which take into account kinematic and dynamic constraints [26][27] and present a local planner for UAV with these constraints, which for the first time generates fast and exact visible trajectories based on analytic solution. The fast and efficient visibility analysis of our method presented above, allows us to generate the most visible trajectory from a start state to the goal state in 3D urban environments, and demonstrates our capability, which can be extended to real performances in the future. We assume knowledge of the 3D urban environment model and use the well-known Velocity Obstacles (VO) method to avoid collision with buildings presented as static obstacles.

For obstacle avoidance capability, at each time step, the planner computes the next eighth Attainable Velocities (AV). The safe nodes not colliding with buildings, i.e., nodes outside Velocity Obstacles [25], are explored. The planner computes the cost for these safe nodes and chooses the node with the lowest cost. Trajectory can be characterized by the most visible roofs only, surfaces only, or another combination of these kinds of visibility types. We repeat this procedure while generating the most visible trajectory.

A. Velocity Obstacles

The Velocity Obstacles (VO) [25] is a well-known method for obstacle avoidance in static and dynamic environments, used in our planner to prevent collision between UAV and the buildings (as static obstacles), as part of the trajectory planning method.

The VO represents the set of all colliding velocities of the UAV with each of the neighboring obstacles, in our case static obstacles - buildings. Each building is bounded by cylinder instead of circle in 2D case [25] and mapped as static obstacle into the UAV's velocity space.

We introduce the velocity obstacles of a planar circular obstacle, B, which is moving at a constant velocity v_b, as a cone in the velocity space of UAV ,*A*, reduced to a point by correspondingly enlarging obstacle B.

Each point in VO represents a velocity vector that originates at *A*. Any velocity of *A* that penetrates VO is a colliding velocity that would result in a collision between *A* and *B* at some future time. Figure 10 shows two velocities of *A*: one that penetrates VO, v_{a1}, and is hence a colliding velocity, and one that does not, v_{a2}.

All velocities of A that are outside of VO are safe as long as B stays on its current course or in our case a static

one. The velocity obstacles thus allows us to determine if a given UAV velocity will cause a collision.

B. Attainable Velocities

Based on the dynamic and kinematic constraints, UAVs velocities at the next time step are limited. At each time step during the trajectory planning, we map the Attainable Velocities (AV), the velocities set at the next time step $t+\tau$, which generate the optimal trajectory, as is well-known from Dubins theory [27].

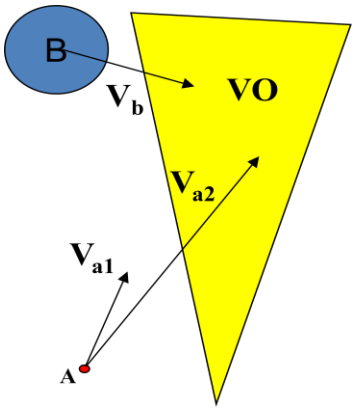


Figure 11. Linear Velocity Obstacles

We denote the allowable controls as $u=(u_s,u_z,u_\phi)$ as U, where $V \in U$.

We denote the set of dynamic constraints bounding control's rate of change as $\dot{u}=(\dot{u}_s,\dot{u}_z,\dot{u}_\phi) \in U'$.

Considering the extremals controllers as part of the motion primitives of the trajectory cannot ensure time-optimal trajectory for Dubin's airplane model [27], but is still a suitable heuristic based on time-optimal trajectories of Dubin - car and point mass models.

We calculate the next time step's feasible velocities $\tilde{U}(t+\tau)$, between $(t,t+\tau)$:

$$\tilde{U}(t+\tau)=U\cap\{u\,|\,u=u(t)\oplus\tau\cdot U'\} \quad (15)$$

Integrating $\tilde{U}(t+\tau)$ with UAV model yields the next eight possible nodes for the following combinations:

$$\tilde{U}(t+\tau)=\begin{pmatrix}\tilde{U}_s(t+\tau)\\ \tilde{U}_z(t+\tau)\\ \tilde{U}_\phi(t+\tau)\end{pmatrix}=\begin{pmatrix}u_s^{\min},u_s(t)+a_s\tau\\ -u_s^{\max}\tan\phi^{\max},u_s(t)\tan u_\phi(t)+u_s^{\max}\tan a_\phi\\ u_z^{\max}u_z(t)-a_z\tau\end{pmatrix} \quad (16)$$

At each time step, we explore the next eight AV at the next time step as part of our tree search. Each node $(q,\dot{q})$,where $q=(x,y,z,\theta)$, consist of the current UAVs position and velocity at the current time step. At each state, the planner computes the set of Attainable Velocities (AV), $\tilde{U}(t+\tau)$, from the current UAV velocity, $U(t)$, as shown in Figure 12. We ensure the safety of nodes by computing a set of Velocity Obstacles (VO).

In Figure 12, nodes inside VO, marked in red, are inadmissible. Nodes out of VO are further evaluated; safe nodes are colored in blue. The safe node with the lowest cost, which is the next most visible node, is explored in the next time step. This is repeated while generating the most visible trajectory.

Attainable velocities profile is similar to a trunked cake slice, as seen in Figure 12, due to the Dubins airplane model with one time step integration ahead. Simple models attainable velocities, such as point mass, create rectangular profile [25].

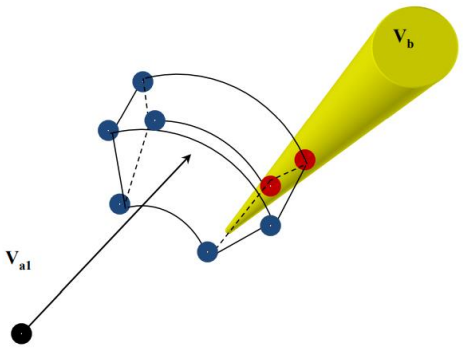


Figure 12. Tree Search Method. Attainable Velocities marked in Blue and Red Circles; Nodes inside VO (marked Red) are Inattainable; Nodes outside VO, Colored in Blue with Lowest Cost, are Explored

C. Cost Function

Our search is guided by minimum invisible parts from viewpoint V to the 3D urban environment model. The cost function for each node is a combination of IRV and ISV, with different weights as functions of the required task.

The cost function is computed for each safe node $(q,\dot{q})\notin VO$, i.e., node outside *VO*, considering UAV position at the next time step $(x(t+\tau),y(t+\tau),z(t+\tau))$ as viewpoint:

$$w(q(t+\tau))=\alpha\cdot ISV(q(t+\tau))+\beta\cdot IRV(q(t+\tau)) \quad (17)$$

Where α,β are coefficients, effecting the trajectory character. The cost function $w(q(t+\tau))$ produces the total sum of invisible parts from the viewpoint to the 3D urban environment, meaning that the velocity at the next time step

with the minimum cost function value is the most visible node in our local search.

D. Planner Neural Network

In our DRL model, we are using fully-connected layers, consisting of:

- the state space of 37 dimensions
- Two hidden layers (64 nodes each)
- An output of four actions

Our network structure can be seen in Figure 13.

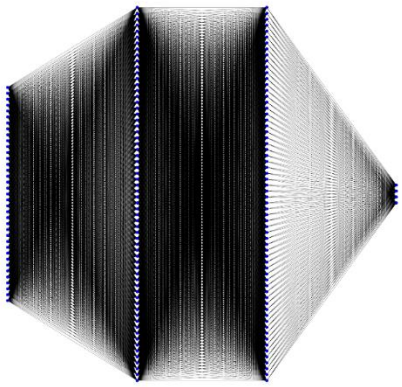


Figure 13. DRL planner network model based on fully-connected layers

E. Simulation Results

We have implemented the presented algorithm and tested some urban environments. We computed the visible trajectories using our DRL planner, as described above. We used the proposed UAV model with several types of trajectories consisting of roof and surfaces visibility, based on the introduced visibility computation method. Obstacle avoidance capability tested by VO method.

The initial parameters values are: $u_s(t=0)=10$ [m/s], u_z $\theta(t=0)=5[\text{deg}]$. UAV dynamic and kinematic constraints are $\phi^{\max}=\pi/4$, $u_z^{\max}=0.3[m/s]$. $u_s^{\min}=1$ [m/s], $u_s^{\max}=15$ [m/s].

In the following figures the start and goal points are marked, in number of scenarios with various start's and goal's points location.

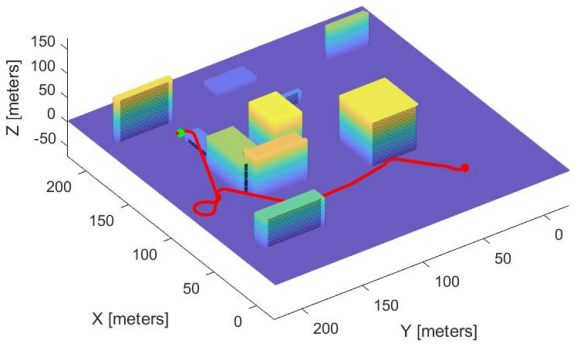


Figure 14. Trajectory Planning in Urban Environment Using DRL. Start and Goal Points with Scenario Demonstration.

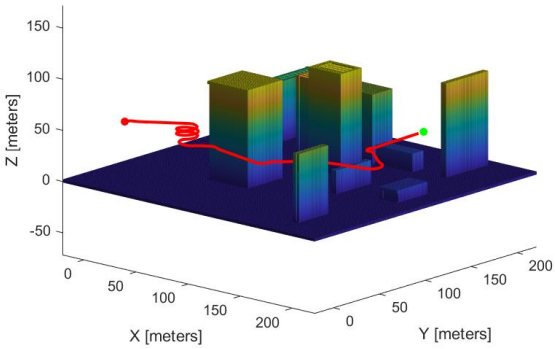


Figure 15. Trajectory Planning in Urban Environment Using DRL. Setting other Start and Goal Points with Scenario Demonstration.

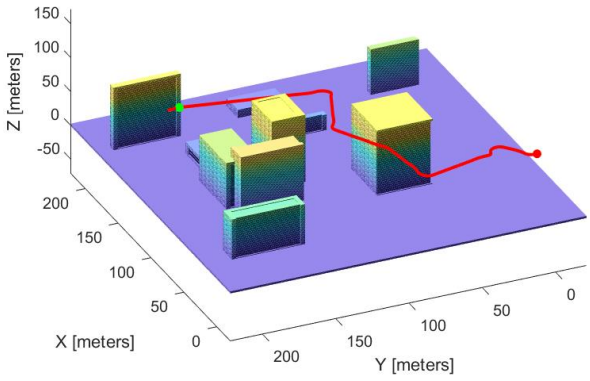


Figure 16. Trajectory Planning in Urban Environment Using DRL. Setting other Start and Goal Points with Scenario Demonstration.

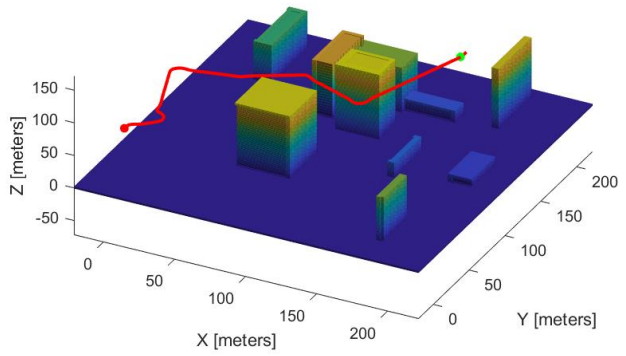


Figure 17. Trajectory Planning in Urban Environment Using DRL. Setting other Start and Goal Points with Scenario Demonstration.

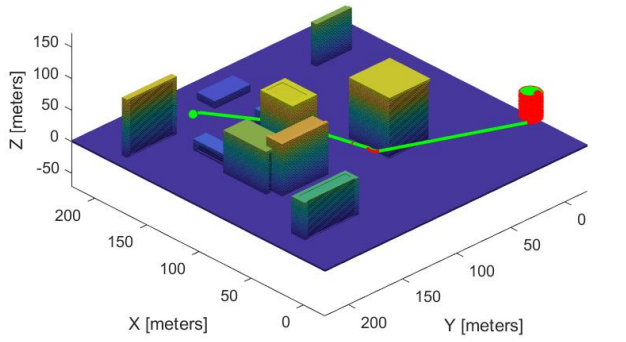


Figure 18. Trajectory Planning in Urban Environment Using DRL. Setting other Start and Goal Points with Scenario Demonstration.

VII. CONCLUSIONS

In this paper, we present spatial motion planner in 3D environments based on Deep Reinforcement Learning (DRL) algorithms. We tackled 3D motion planning problem by using Deep Reinforcement Learning (DRL) approach, which learns agent's and environment constraints.

Spatial analysis focuses on visibility analysis in 3D setting an optimal motion primitive considering agent's dynamic model based on fast and exact visibility analysis for each motion primitives. Based on optimized reward function, which consist of generated 3D visibility analysis and obstacle avoidance trajectories, we introduced DRL formulation, which learns the value function of the planner and generates an optimal spatial visibility trajectory.

We demonstrated our planner in simulations for Unmanned Aerial Vehicles (UAV) in 3D urban environments.

Our spatial analysis is based on a fast and exact spatial visibility analysis of the 3D visibility problem from a viewpoint in 3D urban environments.

We presented DRL architecture generating the most visible trajectory in a known 3D urban environment model, as time-optimal one with obstacle avoidance capability.

VIII. REFERENCES

[1] O. Gal and Y. Doytsher, "Spatial Visibility Clustering Analysis In Urban Environments Based on Pedestrians' Mobility Datasets," The Sixth International Conference on Advanced Geographic Information Systems, Applications, and Services, pp. 38-44, 2014.

[2] J. Bellingham, A. Richards, and J. How, "Receding Horizon Control of Autonomous Aerial Vehicles," in Proceedings ofthe IEEE American Control Conference, Anchorage, AK, pp. 3741–3746, 2002.

[3] A. Borgers and H. Timmermans, "A model of pedestrian route choice and demand for retail facilities within inner-city shopping areas," Geographical Analysis, vol. 18, No. 2, pp. 115-128, 1996.

[4] S. A. Bortoff, "Path planning for UAVs," In Proc. of the American Control Conference, Chicago, IL, pp. 364–368, 2000.

[5] B. J. Capozzi and J. Vagners, "Navigating Annoying Environments Through Evolution," Proceedings of the 40th IEEE Conference on Decision and Control, University of Washington, Orlando, FL, 2001.

[6] H. Chitsaz and S. M. LaValle, "Time-optimal paths for a Dubins airplane," in Proc. IEEE Conf. Decision. and Control., USA, pp. 2379–2384, 2007.

[7] B. Donald, P. Xavier, J. Canny, and J. Reif, "Kinodynamic Motion Planning," Journal of the Association for Computing Machinery, pp. 1048–1066, 1993.

[8] Y. Doytsher and B. Shmutter, "Digital Elevation Model of Dead Ground," Symposium on Mapping and Geographic Information Systems (Commission IV of the International Society for Photogrammetry and Remote Sensing), Athens, Georgia, USA, 1994.

[9] W. Fox, D. Burgard, and S. Thrun, "The dynamic window approach to collision avoidance," IEEE Robotics and Automation Magazine, vol. 4, pp. 23–33, 1997.

[10] O. Gal and Y. Doytsher, "Fast and Accurate Visibility Computation in a 3D Urban Environment," in Proc. of the Fourth International Conference on Advanced Geographic Information Systems, Applications, and Services, Valencia, Spain, pp. 105-110, 2012 [accessed February 2014].

[11] O. Gal and Y. Doytsher, "Fast and Efficient Visible Trajectories Planning for Dubins UAV model in 3D Built-up Environments," Robotica, FirstView, Article pp. 1-21 Cambridge University Press 2013 DOI: http://dx.doi.org/10.1017/S0263574713000787, [accessed February 2014].

[12] M. Haklay, D. O'Sullivan, and M.T. Goodwin, "So go down town: simulating pedestrian movement in town centres," Environment and Planning B: Planning & Design, vol. 28, no. 3, pp. 343-359, 2001.

[13] S. Karaman and E. Frazzoli, "Sampling-based algorithms for optimal motion planning," Int. J. Robot. Res., vol. 30, no. 7, pp. 846–894, 2011.

[14] N.Y. Ko and R. Simmons, "The lane-curvature method for local obstacle avoidance," In International Conference on Intelligence Robots and Systems, 1998.

[15] S. M. LaValle, "Rapidly-exploring random trees: A new tool for path planning," TR 98-11, Computer Science Dept., Iowa State University, 1998.

[16] S. M. LaValle and J. Kuffner. "Randomized kinodynamic planning," In Proc. IEEE Int. Conf. on Robotics and Automation, Detroit, MI, pp. 473–479, 1999.

[17] L.R. Lewis, "Rapid Motion Planning and Autonomous Obstacle Avoidance for Unmanned Vehicles," Master's Thesis, Naval Postgraduate School, Monterey, CA, December 2006.

[18] C. W. Lum, R. T. Rysdyk, and A. Pongpunwattana, "Occupancy Based Map Searching Using Heterogeneous Teams of Autonomous Vehicles," Proceedings of the 2006 Guidance, Navigation, and Control Conference, Autonomous Flight Systems Laboratory, Keystone, CO, August 2006.

[19] S. Okazaki and S. Matsushita, "A study of simulation model for pedestrian movement with evacuation and queuing," Proceedings of the International Conference on Engineering for Crowd Safety, London, UK, pp. 17-18, March 1993.

[20] P. Abbeel and P. Ng, "Apprenticeship learning via inverse reinforcement learning," in Proceedings of the twenty-first international conference on Machine learning, ICML '04, ACM, New York, NY, USA, http://doi.acm.org/10.1145/1015330.1015430, 2004.

[21] M. Kuderer, S. Gulati, and W. Burgard, "Learning driving styles for autonomous vehicles from demonstration," in Proceedings of the IEEE International Conference on Robotics & Automation (ICRA), Seattle, USA. vol. 134, 2015

[22] B. Ziebart, A. Maas, J. Bagnell, and A. Dey, "Maximum entropy inverse reinforcement learning," in Proc. of the National Conference on Artificial Intelligence (AAAI), 2008.

[23] S. M. LaValle, "Planning Algorithms," Cambridge,U.K.:Cambridge Univ. Pr., 2006.

[24] H. Chitsaz and S. M. LaValle, Time-optimal paths for a Dubins airplane, in Proc. IEEE Conf. Decision. and Control., USA, pp. 2379–2384, 2007.

[25] P. Fiorini and Z. Shiller, "Motion planning in dynamic environments using velocity obstacles," Int. J. Robot. Res.17, pp. 760–772, 1998.

[26] G. Elber, R. Sayegh, G. Barequet and R. Martin. "Two-Dimensional Visibility Charts for Continuous Curves," in Proc. Shape Modeling, MIT, Boston, USA, pp. 206-215, 2005.

[27] S. A. Bortoff, "Path planning for UAVs," In Proc. of the American Control Conference, Chicago, IL, pp. 364–368 ,2000.

Collective Interpretation Controlled by Simplified Selective Information-Driven Learning for Interpreting Multi-Layered Neural Networks

Ryotaro Kamimura
Kumamoto Drone Technology and Development Foundation
Techno Research Park, Techno Lab 203
1155-12 Tabaru Shimomashiki-Gun Kumamoto 861-2202
and IT Education Center, Tokai University
4-1-1 Kitakaname, Hiratsuka, Kanagawa 259-1292, Japan
Email: ryotarokami@gmail.com

***Abstract*—The present paper aims to interpret multi-layered neural networks by considering as many possible internal representations as possible, which is called "collective interpretation." The interpretation is performed in a syntagmatic and paradigmatic way. In the syntagmatic processing, all representations created in each step of the learning processes from the beginning to the final stage are considered. Then, in the paradigmatic approach, we try to deal with all possible representations by the syntagmatic processing. In addition, to make this collective interpretation easier, we control collective interpretation by the selective information, which is simplified to control the cost in terms of the strength of connection weights. The collective interpretation with the simplified selective information augmentation by the cost control was applied to three actual data sets: the traffic, facility for the elderly, and wine data sets. With the first two data sets, we could observe that the networks tried to extract simple and clear relations between inputs and outputs. For the wine data set, because the simple cost reduction could not be effective, the cost was first augmented to reduce the selective information, and then it was increased. The final compressed weights were also simplified for clearer interpretation. The results showed that the collective interpretation with the simple selective information control by the cost control could flexibly deal with input and output information for producing simple and interpretable representations.**

Keywords-collective interpretation; selective information; cost; partial compression; generalization

I. Introduction

The present paper aims to propose a new interpretation method composed of collective interpretation and selective information control [1], [2]. We discuss here several problems related to the conventional interpretation methods and then introduce a concept of collective interpretation. This interpretation tries to take into account as many internal representations as possible with a method of selective information to make the collective interpretation clearer and easier.

A. Interpretation Problem

As has been well known, neural networks have been notorious as one of the typical black-box models in machine learning, though there have been many attempts to interpret their internal representations from the beginning of the research [3]–[7]. Even if neural networks can show good performance in generalization, they have not been accepted as reliable models, because there have been serious risks we must face in unexpected ways. In addition, neural networks have been used to explain and understand human cognitive processes, as was done in the name of connectionism [3], [8], [9]. In this approach, the interpretation of internal representations obtained by neural networks is the objective of the research, and generalization performance, which is nowadays one of the main objectives in neural networks, is only one aspect among many to be explained.

Meanwhile, the massive invasion of neural networks as well as other machine learning techniques into our daily life has caused some concern about their use for our critical decision making. Then, due to the urgent need to respond to the right to explanation [10], there have been many different types of interpretation, in the field of convolutional neural networks (CNN) in particular. Those conventional interpretation methods can be classified into three types: conditional, individual, and intuitive.

First, the interpretation has been based only on a specific condition. Usually, we have tried to interpret an instance of network behavior only when an initial condition is applied. Actually, with a specific initial condition, for example, with a specific set of initial weights, learning is performed, followed by the interpretation of obtained representations. However, the final internal representations are greatly variable, depending on different initial conditions; it is almost impossible to give fixed and stable meanings to those different representations, and furthermore, some contradictory interpretations can be obtained. In particular, when we have tried to apply logical and linguistic rules to the interpretation [11]–[14], we have faced much difficulty in interpreting the different rules. With those formal methods for interpretation, we can produce a number of different formal and logical rules for interpretation. Certainly, we can determine a specific representation for interpretation. For example, we should interpret a representation related to the best generalization. Generalization is an important property to be pursued, but we need to consider many other factors for neural networks when we try to make them as close as possible to human intelligence. Those types of interpretation can be valid only under some specific conditions, such as specific initial conditions, best generalization, and so on. In the present paper, it is supposed that the interpretation should be

as independent as possible of any specific conditions. Second, the conventional methods tend to interpret network behaviors individually, which is closely related to the conditional interpretation. This means that the interpretation has been restricted to an interpretation responding to a specific input or a specific output. In particular, in the convolutional neural network (CNN), many individual interpretations or visualization methods have been developed with much success, for example, the activation maximization [15]–[20], the sensitivity detection [21]–[25], the layer-wise relevance propagation (LRP) [26]–[31], and so on. This is because the intuitive interpretation of image data sets to be discussed immediately below, dealt with by the CNN, has made it possible to understand an instance of network behavior seemingly where the interpretation has been replaced by the intuitive one for a specific image data set. This individual interpretation seems to be successfully applied to many data sets. However, one of the main problems is that the individual interpretation can produce a number of different types of interpretation on one data set, which can be contradictory from each other in some cases. We can say that the sum of individual interpretations cannot necessarily lead us to the full understanding of data sets, because there should be a number of cases in which some interpretations are contradictory to others [32].

The third one is also close to the first and second one, where the interpretation tends to be heavily dependent on our intuitive knowledge of data sets, in particular, when the method is applied to image data sets, as discussed above. Intuition is naturally one of the most important techniques in the explanation, because it is easy to persuade people how neural networks can understand inputs and produce outputs. However, this intuition has prevented us from understanding the true inference mechanism of neural networks. The inference mechanism of neural networks should be different from that of human beings, because the inference mechanism of human beings should be severely constrained culturally and physically so as to maintain their stability and existence [33]. Neural networks have been well known to produce unexpected final outputs, which have been called "adversarial examples" [34], [35]. The adversarial examples can be explained when we can interpret the inference mechanism without human intuition or human cultural bias toward or against the data sets. The neural network can deal with the data sets from a viewpoint that is different from that of human beings. More strongly, the viewpoint cannot be accepted by human beings due to the cultural and physical constraints on their inference mechanism [33], [36]. It should be repeated that human beings are strictly restricted by their physical or cultural conditions that might threaten their existence. Their inference mechanism has been acquired in those severe conditions, which naturally provides strong bias for the interpretation. Thus, the human inference can be only one of many different ways to deal with given inputs appropriately. In short, the adversarial attacks may show one truth about human intuition that the data sets cannot be necessarily well suited to interpret by the inference mechanism of neural networks.

Those limitations and conditions of interpretation seem to be related to the severe shortcomings of neural networks. However, when different types of explanations can be unified, neural networks can be ironically well suited for dealing with unstable and multiple interpretations, compared with the conventional statistical methods. As mentioned above, we have a problem of conditional interpretation, where one of the main problems of neural networks is that they are seriously dependent on initial conditions and where different initial conditions can produce completely different internal representations. Though this phenomenon seems to be one of the main drawbacks of neural networks, it can also be one of the merits of neural networks. This is because they can explain many different aspects of given tasks and data sets just by using different initial conditions. Conventional statistical methods have tried to obtain a representation fixed by a corresponding idealized model, while neural networks try to produce as many different types of representations as possible by using different initial conditions. At this point, all we have to do is to propose a method to unify those different types of representations created by neural networks.

B. Collective Interpretation

In this context, we present here a new type of interpretation method called "collective interpretation," aiming to consider all possible internal representations generated by neural networks. On the contrary, the conventional interpretation method in neural networks is an attempt to interpret only one internal representation. First, as mentioned above, we suppose that different results by different initial conditions should be considered one of the main merits of neural networks. The different results can be produced by an effort to see a given task or data set from a number of different viewpoints and in a number of different ways. All results by different initial conditions should have some meaning to explain the task. Our hypothesis is that all representations by different initial conditions should be taken into account to reach the full understanding of the inference mechanism of neural networks.

In collective interpretation, there are two components: network compression and selective information control. First, we introduce the network compression to simplify multi-layered neural networks. Our method of compression lies in compressing as many representations as possible into the simplest form for explaining the core knowledge obtained by neural networks. Model compression has received due attention recently to simplify multi-layered neural networks [37]–[44]. Those conventional methods have aimed to replace complicated multi-layered networks with simpler ones, keeping the same generalization performance as much as possible. Thus, the internal representations obtained by those methods cannot inherit the original representations of complex multi-layered neural networks. On the contrary, we have proposed a method to compress multi-layered neural networks [45], keeping information stored in weights in multi-layered neural networks as much as possible.

In addition to different types of internal representations, we consider connection weights in syntagmatic and paradigmatic ways. First, by restricting learning individually and conditionally, we train neural networks with a specific set of initial conditions and input patterns, and all representations in the course of learning are collected. This process is called "syntagmatic processing," where all representations, obtained in each learning step, are taken into account to produce collected representations. This syntagmatic processing should be performed for different initial conditions and input patterns, producing a number of different types of representations. Then, we should collect all those different compressed representations, which is called "paradigmatic processing." Collective interpretation is composed of the network compression where syntagmatic processing is first applied, followed by paradigmatic processing to deal with all possible representations.

An ideal collective interpretation should consider all instances obtained in neural learning, and it should extract some core structure by which all instances can be generated. The present paper uses a kind of partial conditional collective interpretation, where one condition is assumed for the collective interpretation. The condition is that information per cost should be maximized for simplification. Information-theoretic methods have been introduced from the beginning of research into neural networks, producing many principles affecting studies on neural information processing. For example, Linsker's maximum information preservation principle has had much influence on neural computing [46]–[49], in which some visual processing can be explained by the maximum information principle. Intuitively, we humans try to collect surrounding information to secure our existence and to keep it as secure as possible. Thus, though the principle should play more important roles in extracting some principles in neural computing, there have been few attempts made to use information-theoretic principles following important past studies [50]–[55]. In this context, we try to show how neural networks are transformed under the condition that information per cost is maximized. Then, we try to show that we can disentangle complicated representations into the simplest ones when the information per cost is maximized. For this, we introduce a method to increase the selective information for connection weights, expecting that those weights will be selected to be disentangled from each other.

However, when the information is formulated in the classical form of information measures such as entropy and mutual information, it is not so easy to understand how those measures are concretely related to the disentanglement of representation. This is because the abstract and ambiguous property of information, accompanied by the need for much computational resources, has prevented us from using them appropriately for the actual formulation. The present paper proposes a more simplified method to compute the selective information, which is not the abstract measure of information but which has the actual meaning of the number of important connection weights. Thus, the selective information can be applied to neural networks and to understanding how information can be stored in terms of the number of connection weights.

The selectivity has played important roles in neural networks, in particular, in generalization [56]–[61]. We should choose a small number of important connection weights, based on some criteria on the importance. However, it is impossible to know the importance of connection weights, and it has been stressed that the selectivity is of no use in generalization [56], [59], [61]. For coping with this problem of selectivity, we use the passive method to extract important ones. We use the concept of cost [62] in terms of strength of connection weights. We consider a connection weight important only when this weight remains strong by introducing the cost reduction method. This method is closely related to the conventional weights decay, but the fundamental difference is that the cost reduction is performed independently of error minimization. This simple and independent cost reduction method can eventually produce a small number of important weights.

Selective information can be maximized to simplify neural networks, but this simplification is not necessarily successful. The information on the given data set should be naturally obtained through inputs. However, those inputs are artificially prepared by our knowledge on the data set. When these inputs cannot be used to transmit information on inputs, simplified networks cannot necessarily represent information on relations between inputs and outputs. In this case, we must decrease information on inputs as much as possible. Thus, we first try to increase the selective information to simplify networks. Then, if it is impossible to simplify them, we try to decrease selective information in the first place and then increase selective information.

C. The Purpose of the Present Study

Considering the above problems, the present paper aims to propose a new interpretation method with three properties. First, connection weights produced by neural networks are exhaustively considered in syntagmatic and pragmatic ways. This tries to take into account all possible representations by the neural network. Second, the network simplification is performed not by the selective information maximization directly but by the corresponding cost minimization. Thus, the learning procedures are greatly simplified. Third, when the selective information maximization cannot give acceptable results, we first minimize the selective information by increasing the cost. Then, the ordinary selective information minimization is applied. This can be used to eliminate harmful information obtained through inputs.

D. Paper Organization

The paper has been organized as follows. In Section 2, after briefly explaining the concept of collective interpretation, we try to explain full compression with syntagmatic and paradigmatic compression. In addition, to see the intermediate states of learning, we introduce partial compression and how to partially compress intermediate layers. Then, we introduce the potentiality and corresponding selective information, followed

by the computational methods of cost reduction and augmentation. We applied the method to three data sets, namely, the traffic, facility for the elderly, and wine data sets. In all cases, we first tried to show that syntagmatic and paradigmatic compression could produce compressed weights close to the correlation coefficients between inputs and targets. In the first two data sets, by the cost reduction, we could increase the ratio of selective information to its cost. By the partial compression, we could see that the present method tried to deal with inputs from the lower hidden layers, while the other conventional methods could not consider inputs well. Because the wine data set could not produce reasonably good interpretation results, we first decreased the selective information by increasing the corresponding cost, and then, the selective information was increased by decreasing the cost. In all experimental results, the final collective interpretation showed that the main characteristics were based on the correlation coefficients between inputs and targets. The differences between compressed weights and correlation could be used to detect the effects of non-linear relations between inputs and outputs.

II. Theory and Computational Methods

We explain here the concept of collective interpretation, taking into account all internal representations by the syntagmatic and paradigmatic compression. In addition to the full network compression, we introduce the partial compression to see the states of intermediate layers. Then, we introduce the simplified information-theoretic method by the selective information, representing how many connection weights are combined with neurons. After formulating the potentiality and selective information, we introduce a practical method to control selective information, where instead of direct control of selective information, we try to control the cost in terms of strength of weights. Finally, we present a two-step learning method in which the cost is first increased, and then decreased when the simple cost reduction could not be applied.

A. Compression

1) Collective Interpretation: We introduce here a concept of collective interpretation in which we try to take into account all possible internal representations, assumed to have equal importance, created by the neural network. One of the main shortcomings of neural networks is that their learning behaviors are sometimes completely different when different initial conditions and different subsets of a data set are given, as shown on the left-hand side of Figure 1. However, we suppose here that this shortcoming of different learning behaviors should be one of the most important merits of neural networks. This means that a neural network tries to see a target object from many different points of view, corresponding to different initial conditions and different subsets of a data set. Then, we suppose that all representations created by the neural network should have some meaning related to the properties of the target objects. We more strongly assume that all possible representations should have the same importance, at least, in terms of interpretation, dealt with in this paper. As shown in Figure 1, a neural network can produce many different final representations by different initial conditions and different subsets of a data set. Then, we should interpret how a neural network tries to produce outputs, based on the corresponding inputs, considering all possible internal representations they create. The interpretation, taking into account all possible internal representations, can be called "collective interpretation" in this paper.

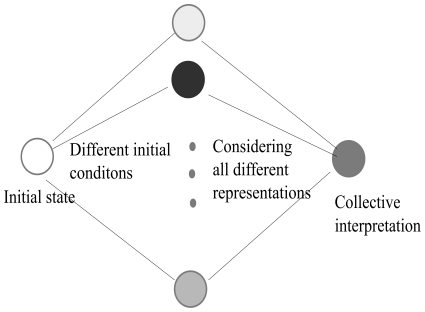


Fig. 1. Collective interpretation aiming to consider all possible internal representations created by a neural network.

2) Full Compression: For interpreting multi-layered neural networks, we first compress them into the simplest ones, as shown in Figure 2. We try here to trace all routes from inputs to the corresponding outputs by multiplying and summing all corresponding connection weights.

First, we compress connection weights from the first to the second layer, denoted by (1,2), and from the second to the third layer (2,3) for an initial condition and a subset of a data set. Then, we have the compressed weights between the first and the third layer, denoted by (1,3).

$$w_{ik}^{(1,3)} = \sum_{j=1}^{n_2} w_{ij}^{(1,2)} w_{jk}^{(2,3)} \tag{1}$$

Those compressed weights are further combined with weights from the third to the fourth layer (3,4), and we have the compressed weights between the first and the fourth layer (1,4).

$$w_{ik}^{(1,4)} = \sum_{k=1}^{n_3} w_{ik}^{(1,3)} w_{kl}^{(3,4)} \tag{2}$$

By repeating these processes, we have the compressed weights between the first and sixth layer, denoted by $w_{iq}^{(1,6)}$. Using those connection weights, we have the final and fully compressed weights (1,7).

$$w_{ir}^{(1,7)} = \sum_{q=1}^{n_6} w_{iq}^{(1,6)} w_{qr}^{(6,7)} \tag{3}$$

Because we consider all routes from the inputs to the outputs, the final connection weights should represent the overall characteristics of connection weights of the original multi-layered neural networks.

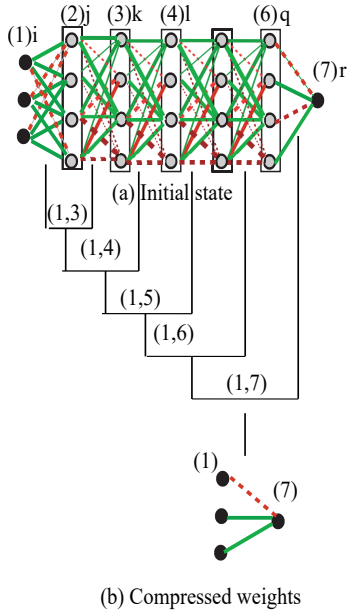


Fig. 2. Full compression for an initial condition and a subset of a data set from a seven-layered to a two-layered network without hidden layers.

3) Syntagmatic and Paradigmatic Compression: The full compression actually is composed of syntagmatic and paradigmatic compression in Figure 3. With an initial condition and a set of input patterns, we train a neural network, taking into account all internal representations by all possible conditions and subsets of a data set. For simplicity's sake, we suppose that only initial conditions are changed, but actually, the subset of the data set can be changed. Then, we average obtained connection weights over all weights obtained in a process of learning for the initial condition. Let us take an example of connection weights from the sixth to the seventh layer only and the maximum number of training steps for the sth initial condition in Figure 3(a4). Then, we can average all possible weights for all training epochs. For the weights from the sixth to the seventh weights $(6, 7; t)$ for the tth learning epoch, we can average all possible weights

$$\bar{w}_{qr}^{(6,7)} = \frac{1}{t_s} \sum_{t=1}^{t_s} w_{ir}^{(6,7;t)} \tag{4}$$

where t_s denotes the maximum number of learning steps for the sth initial condition. All other connection weights are averaged in the same way. Then, we compress those average weights in full compression.

$$\bar{w}_{ir}^{(1,7)} = \sum_{q=1}^{n_6} \bar{w}_{iq}^{(1,6)} \bar{w}_{qr}^{(6,7)} \tag{5}$$

where $\bar{w}_{iq}^{(1,6)}$ denote the compressed averaged weights up to the sixth layer. This compression can be called "syntagmatic compression" in Figure 3, because it tries to compress all connection weights obtained for all learning steps.

Finally, the syntagmatically compressed weights are averaged over all initial conditions and subsets of the data sets. For simplicity's sake, we restrict the compression for an initial condition, and we have the paradigmatic compression in Figure 3(b).

$$\bar{w}_{ir} = \frac{1}{s_m} \sum_{s=1}^{s_m} \bar{w}_{ir}^{(1,7)} \tag{6}$$

where s_m denotes the maximum number of initial conditions. We should repeat that we try to consider all possible representations created by neural networks. Thus, we can deal with all connection weights for all learning steps and by all different initial conditions and input patterns.

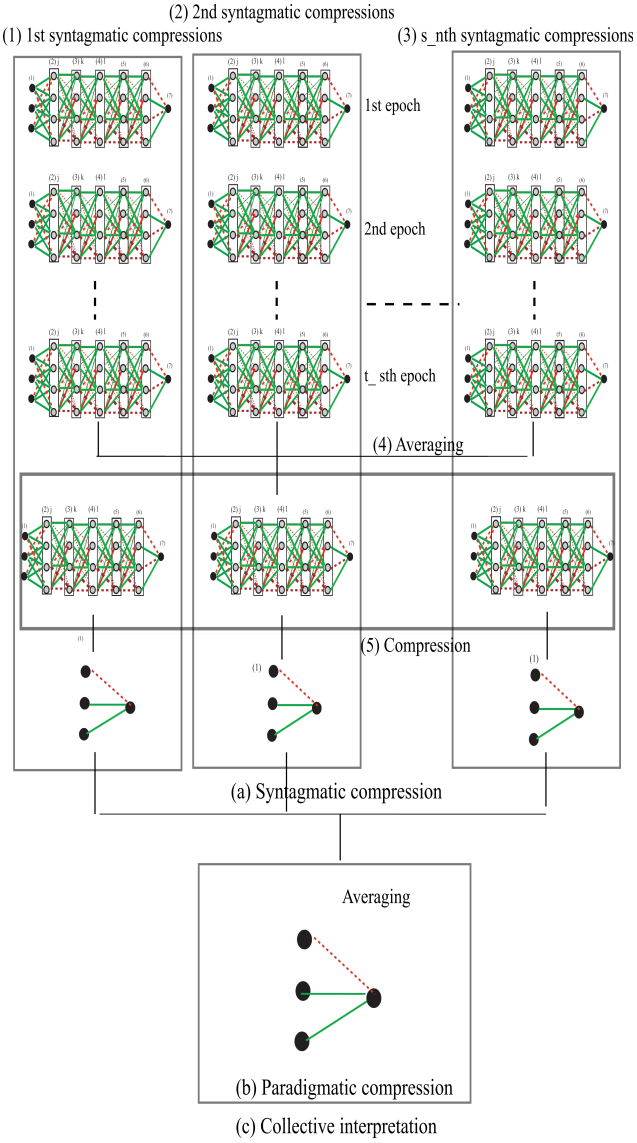


Fig. 3. Collective compression composed of syntagmatic (a) and paradigmatic (b) compression for collective interpretation (c).

4) Partial Compression : In addition to the full compression, we need to examine the outputs from the intermediate layers. For this purpose, we introduce the partial compression, in which compression is applied up to a specific layer. As shown in Figure 4, we illustrate the partial compression up

to the fourth layer. Now, let us assume that we have already compressed weights up to the fourth layer, denoted by $w_{il}^{(1,4)}$. In addition, the number of neurons in all hidden layers is supposed to be the same. The partially compressed weights up to the fourth layer can be computed by

$$w_{ir}^{(1,4,7)} = \sum_{q=1}^{n_6} w_{iq}^{(1,4)} w_{qr}^{(6,7)} \tag{7}$$

where $w_{iq}^{(1,4)}$ denote connection weights, compressed up to the fourth layer. For the other intermediate layers, we can compute the same partially compressed weights. The partial compression aims to examine to what degree the intermediate layers contain information on inputs as well as outputs.

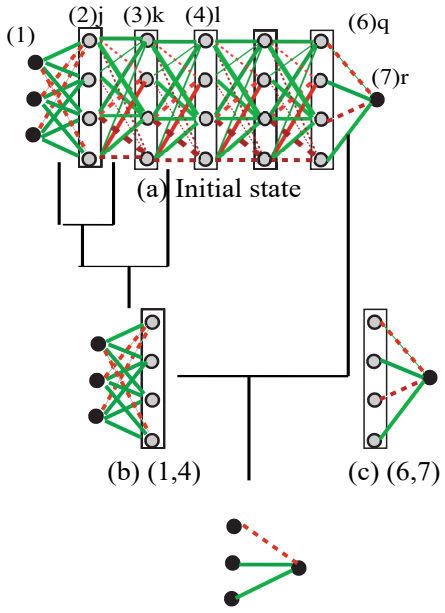


Fig. 4. An example of partial compression where only weights up to the fourth layer are compressed.

B. Reduction and Augmentation of Selective Information

1) Selective Information and Its Cost: The selective information can be defined by using the selective potentiality of connection weights. When the selective information increases, a small number of connection weights tends to be connected with some specific neurons. The individual potentiality of connection weights can be defined by the absolute values of weights, for example, from the second to the third layer, represented by (2,3), which is computed by

$$u_{jk}^{(2,3)} = \mid w_{jk}^{(2,3)} \mid \tag{8}$$

Then, we normalize these values by their maximum ones.

$$h_{jk}^{(2,3)} = \frac{u_{jk}^{(2,3)}}{\max_{j'k'} u_{j'k'}^{(2,3)}} \tag{9}$$

where the maximum operation is over all connection weights between two layers. Then, summing all these normalized values, the selective potentiality can be defined by

$$H^{(2,3)} = \beta_1 \sum_{j=1}^{n_2} \sum_{k=1}^{n_3} \left[\frac{u_{jk}^{(2,3)}}{\max_{j'k'} u_{j'k'}^{(2,3)}} \right] \tag{10}$$

where n_2 and n_3 denote the number of neurons in the second and the third layer, and β_1 is a parameter to control the strength. It should be larger than zero. Then, the complementary potentiality is defined by

$$g_{jk}^{(2,3)} = 1 - \frac{u_{jk}^{(2,3)}}{\max_{j'k'} u_{j'k'}^{(2,3)}} \tag{11}$$

Summing all these normalized values, the selective information can be defined by

$$G^{(2,3)} = \beta_2 \sum_{j=1}^{n_2} \sum_{k=1}^{n_3} \left[1 - \frac{u_{jk}^{(2,3)}}{\max_{j'k'} u_{j'k'}^{(2,3)}} \right] \tag{12}$$

In addition, we need to define the corresponding cost to represent the potentiality and information. In this paper, the cost is simply the sum of all the absolute weights.

$$C^{(2,3)} = \sum_{j=1}^{n_2} \sum_{k=1}^{n_3} u_{jk}^{(2,3)} \tag{13}$$

We suppose that the cost representing the information should be as small as possible, and then the final function to be controlled for the selective potentiality is

$$R^{(2,3)} = \frac{H^{(2,3)}}{C^{(2,3)}} \tag{14}$$

Then, for the selective information, the function to be controlled is

$$R^{(2,3)} = \frac{G^{(2,3)}}{C^{(2,3)}} \tag{15}$$

2) Cost Control for Sensitive Selective Information: The selective information should be augmented and the corresponding cost should be reduced in the majority of data sets. When we try to control the ratio of selective information to its cost, we have two possible ways to do so: selective information control or cost control. Because it is sometimes difficult to directly control the selective information, we focus on the cost and try to control it. In addition, when we try to increase the selective information, one of the major problems is that we cannot identify important connection weights or see whether a weight plays a major role in interpretation or generalization. Thus, we pay attention to the corresponding cost, and we try to reduce the cost as much as possible, which is expected to increase the selective information eventually. For this case, the connection weights at the $t+1$th learning step are simply computed by

$$w_{jk}^{(2,3)}(t+1) = \beta_1 \, w_{jk}^{(2,3)}(t) \tag{16}$$

where β_1 should range between zero and one, because we try to reduce the cost or the strength of connection weights.

However, for some data sets, we have found that the selective information augmentation and the corresponding cost reduction cannot be accompanied by disentangling connection weights into simplified ones for better interpretation. In those cases, we first reduce the selectivity at the expense of higher cost. Then, we try to increase the selective information and to decrease the corresponding cost. Figure 5 shows the process of a two-step method of selective information reduction and augmentation. In the initial state in Figure 5(a), connection weights are randomly initialized with the intermediate selectivity. Then, we try to decrease the selectivity at the expense of larger connection weights or higher cost in Figure 5(b). Finally, we try to decrease the cost and at the same time increase the selective information in Figure 5(c). In this case, for the initial learning steps, we have

$$w_{jk}^{(2,3)}(t+1) = \beta_2 \, w_{jk}^{(2,3)}(t) \tag{17}$$

The parameter β_2 should be larger than one. We try to increase the strength of connection weights. This leads us to the augmentation of selective potentiality at the expense of cost. We use this method because it is easy to decrease the selective information. Then, for the remaining learning steps, we have the same assimilation rule

$$w_{jk}^{(2,3)}(t+1) = \beta_1 \, w_{jk}^{(2,3)}(t) \tag{18}$$

However, the parameter β_1 should be between zero and one to reduce the cost and correspondingly to increase the selective information.

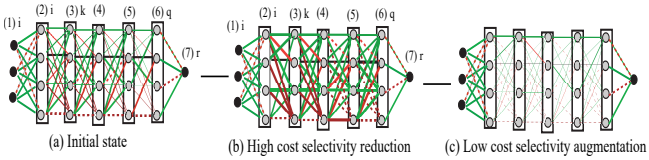


Fig. 5. Selective information augmentation (c) through higher cost selective information reduction (b).

3) Assimilation: Depending on their strength, weights are controlled to be smaller or larger. However, when the weights are controlled by the parameter β, we need to re-train a neural network to assimilate learning processes. We try to repeat this process of assimilation many times. One of the possible ways to do so is to use the dth sub-epoch t_d of the tth learning step, and it can be computed by

$$t_d = \theta_1 \left(\frac{t}{t_{max}} \right)^{\theta_2} + \theta_3 \tag{19}$$

where d is the dth sub-epoch of step of the tth learning step and t_{max} is the maximum number of learning steps with three parameters, $\theta_1, \theta_2, \theta_3$, to control the effect of assimilation.

Figure 6 shows a process of assimilation for a learning step. First, weights in an initial state in Figure 6(a) are multiplied by the parameter β (for example, smaller values) in Figure 6(b), and the strength of weights is reduced in proportion to the parameter β in Figure 6(c). Then, we repeat the assimilation steps for the learning step several times in Figure 6. Because the effect of the parameter β is weakened in this process of assimilation to reduce training errors, we must have weakened weights less than those at the initial stage of assimilation due to the effect of error minimization in Figure 6. Then, we repeat this process of assimilation for each learning step to obtain the final reduced weights. One of the important features of this assimilation method is that the assimilation (error minimization) and potentiality assignment (application of the parameter β) are performed separately. First, the strength of weights is reduced, and then, the effect of the parameter is assimilated (error minimization). This method, thus, can resolve the contradiction between error minimization and regularization, which are usually simultaneously performed.

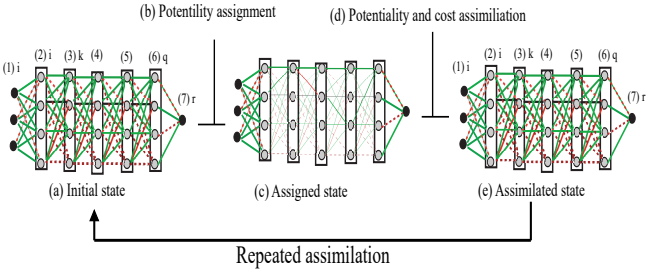


Fig. 6. A computational method to assimilate the effect of cost reduction.

III. Results and Discussion

We present here experimental results on three data sets: traffic, facility for the elderly, and wine. In the first two data sets, we used the simple cost reduction method to increase the selective information. For the third data set, the simple cost reduction could not produce reasonable results, so we first augmented the cost, and the usual cost reduction to increase the selective information was applied. With those three methods, we tried to show that we could compress networks syntagmatically and paradigmatically with the aid of cost or selective information control into simpler and clearer networks, whose connection weights could be closer to the correlation coefficients between inputs and targets of the original data sets. In addition, we could extract some properties due to the non-linear processing of neural networks.

A. Traffic Data Set

1) Experimental Outline: The database was created with records of behavior in urban traffic in the city of Sao Paulo in Brazil [63]. The number of inputs was 17, and the number of patterns was 135. Seventy percent of the data set was used for training, and the remainder was for testing. To make the reproduction of the present results easier, we tried to use the scikit-learning package with all default values except for the tangent-hyperbolic activation function and the number of epochs, which was changed according to the equation described above. Table I shows the parameter values for the experiments. In the following sections on experimental results, we used the same parameter values for the easy

TABLE I
SUMMARY OF PARAMETER VALUES FOR THE TRAFFIC DATA SET.

Parameters	Values
β_1	0.85
θ_1	5
θ_2	1
θ_3	5

reproduction of all results except for the third results, where a new parameter $\beta_2 = 1.3$ for augmentation of potentiality or information minimization was introduced.

2) Syntagmatic and Paradigmatic Compression : We compared compressed weights with correlation coefficients between inputs and targets of the original data set, supposing that the correlation coefficients were meaningful for describing the relations between inputs and outputs. The results show that the present method could produce syntagmatically and paradigmatically compressed weights close to the correlation coefficients between inputs and targets of the original data set. Though the weight decay and conventional method could produce reasonably high correlations, they were still lower and behind the correlations by the present method.

Figure 7 shows the syntagmatic (left) and paradigmatic (right) compression for the traffic data set for 100 different initial conditions and 100 different subsets of the data set. One of the main characteristics is that, when the parameter β_1 was 0.85 for the cost reduction, correlation coefficients between syntagmatically compressed weights and original correlations between inputs and targets of the original data set were much higher than those by any other method, and close to one (perfect correlation) in the box on the left-hand side of Figure 7(a). The box on the right-hand side of Figure 7 (a) shows the results of paradigmatic compression, and we could see that when the number of different initial conditions and different subsets of the data set increased, the correlation coefficients became close to the maximum of one. When the parameter α for the weight decay was set to 0.1 in Figure 7(b), the correlation coefficients for the syntagmatic compression became lower than those by the cost reduction in Figure 7(left, a). For the paradigmatic compression in Figure 7(right, b), the correlations became larger gradually, but the final correlations were lower than those by the present method in Figure 7(a). Finally, even without weight decay, the final results were quite similar to those with the weight decay in Figure 7(c).

The results confirmed that the collective interpretation could extract relations between inputs and outputs that were close to the original correlation coefficients between inputs and targets. Thus, neural networks, in particular, with the cost reduction, could disentangle connection weights that could be compressed to represent simple relations between inputs and outputs.

3) Selective Information, Cost, and Ratio: The results show that, though the new method could not increase selective information in the later stages of learning, the cost was reduced sufficiently to increase the ratio of information to its cost. On

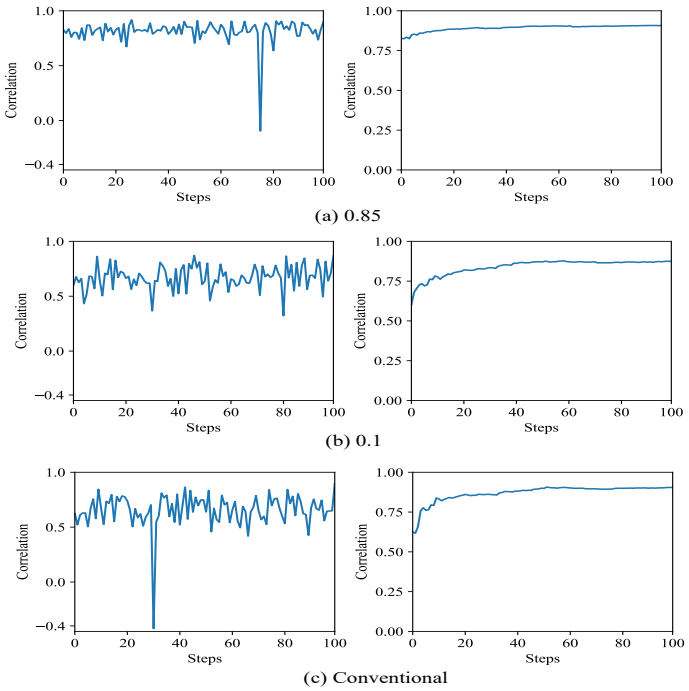


Fig. 7. Correlation coefficients between weights and original correlation coefficients by the syntagmatic compression (left) and by the paradigmatic compression (right), when the parameter β_1 was 0.85 (a), α was 0.1 for weight decay (b), and by the conventional method without weight decay (c) for the traffic data set.

the contrary, the weight decay and conventional method could not increase the ratio of information to its cost.

Figure 8 shows selective information (left), cost (middle), and the ratio of information to its cost (right). When the parameter for the cost reduction was 0.85 in Figure 8(a), the information first increased gradually, and then it decreased. On the other hand, the cost decreased gradually, and remained almost a constant in the later stages of learning. Naturally, the ratio of information to its cost increased and then decreased slightly in the end. When the weight decay was used and the parameter α was set to 0.1 in Figure 8(b), the information constantly increased, and the cost gradually decreased, though it did not decrease to the lower point attained by the present method. The ratio was much lower than that by the present method in Figure 8(right, b). Finally, when we used the conventional method without the weight decay in Figure 8(c), the information did not change, the cost remained higher, and finally, the ratio remained lower.

The results confirmed that the cost reduction could inhibit the generation of supposedly important connection weights. On the contrary, the weight decay constantly increased the selectivity of connection weights.

4) Weights: The results showed that the present method could produce weights where a small number of them became stronger, and we could also see that some groups of connection weights were identified. On the contrary, the weight decay and conventional method could not produce a similar result.

Figure 9(a) shows connection weights (1) and their individual potentiality (2). As can be seen in the figure, a

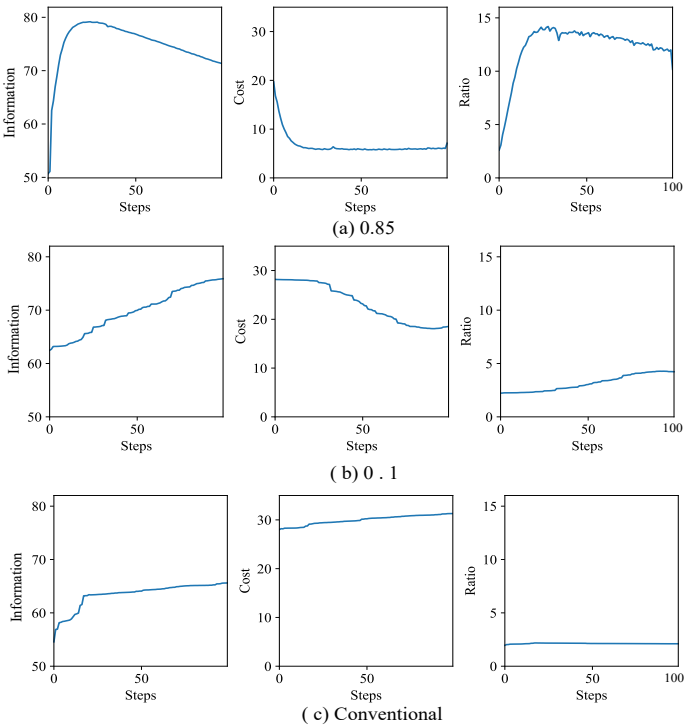


Fig. 8. Information (left), cost (middle), and ratio (right) when the parameter β_1 was 0.85 (a), α was 0.1 (b), and by the conventional method (c) for the traffic data set

small number of connection weights became stronger, and they responded to inputs in the precedent layers with clear regularity. This tendency was further enhanced over individual potentialities in Figure 9(2). Figure 9(b) shows weights and individual potentiality by the weight decay ($\alpha = 0.1$). Though we could not see any strong weights, a small number of weights could be seen by using the individual potentiality. Finally, when the conventional method was used in Figure 9(c), weights seemed to become randomly activated, though we could see a smaller number of individual potentiality.

5) Partial Compression: The results show that the present method tried to extract information from inputs, while the weight decay and conventional method tried to extract information from outputs.

Figure 10(a) shows partially compressed weights when the parameter β_1 was 0.85. As can be seen in the figure, only the initial partially compressed weights had higher connection weights, and the strength of weights remained small. This means that at the beginning the present method tried to acquire information on inputs and that it seemed to try to extract information on inputs as much as possible. Figures 10(b) and (c) show partially compressed weights by the weight decay and by the conventional method. Clear compressed weights could not be seen until the final compression was performed. This means that information from outputs played a critical role in creating the final connection weights.

6) Full Compression: The results show that the present method could produce compressed weights whose correlations with the original correlations between inputs and targets were

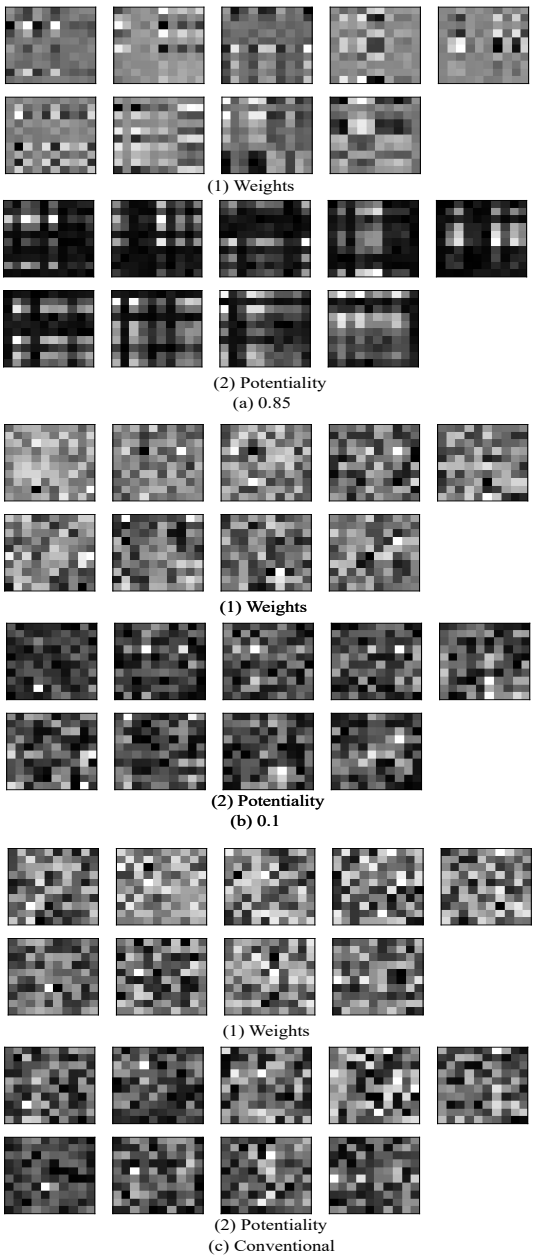


Fig. 9. Weights (1) and potentiality (2) when the parameter β_1 was 0.85 (a), α was 0.1 (b), and by the conventional method (c) for the traffic data set.

high and close to those by the logistic regression. In addition, the present method produced higher generalization accuracy.

Figure 11(1) shows the correlation coefficients between inputs and targets, and we could see that the first input (hour) played the most important role in traffic behavior. Figure 11(2) shows fully compressed weights by the paradigmatic compression when the parameter β_1 was 0.85. As can be seen in the figure, the correlation was 0.908, the second largest one behind the logistic regression analysis, and generalization accuracy was the highest at 0.812. When the weights decay was introduced, the correlation decreased to 0.875, and the accuracy also decreased to 0.808 in Figure 11(3). When the conventional method was used in Figure 11(4), the correlation

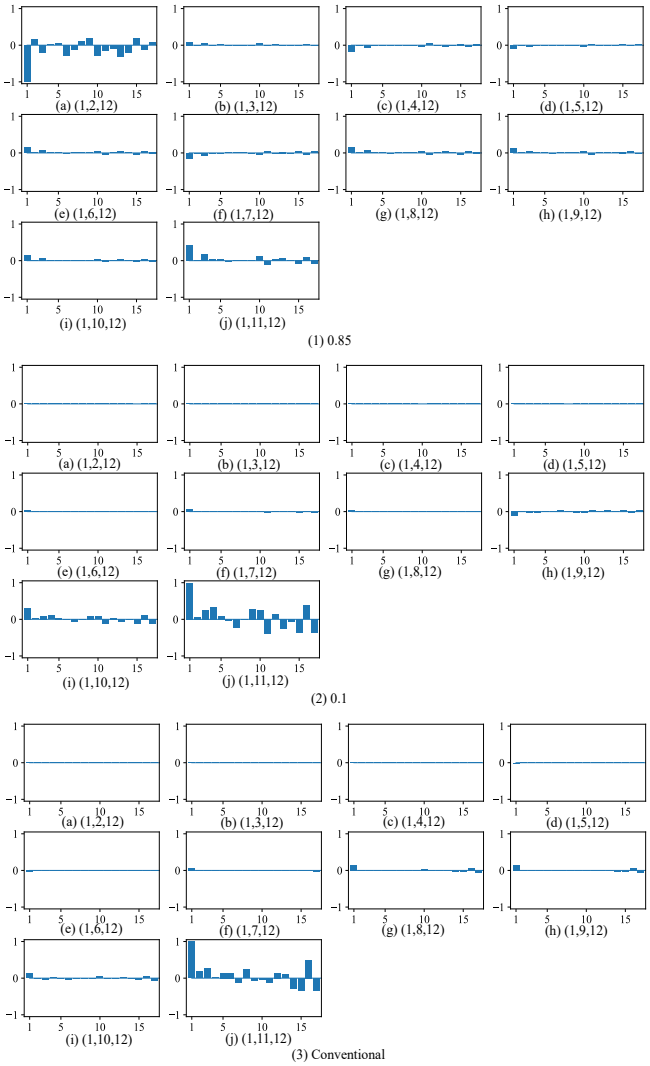


Fig. 10. Partially compressed weights when the parameter β_1 was 0.85 (1), α was 0.1 for weight decay (2), and by the conventional method (3) for the traffic data set.

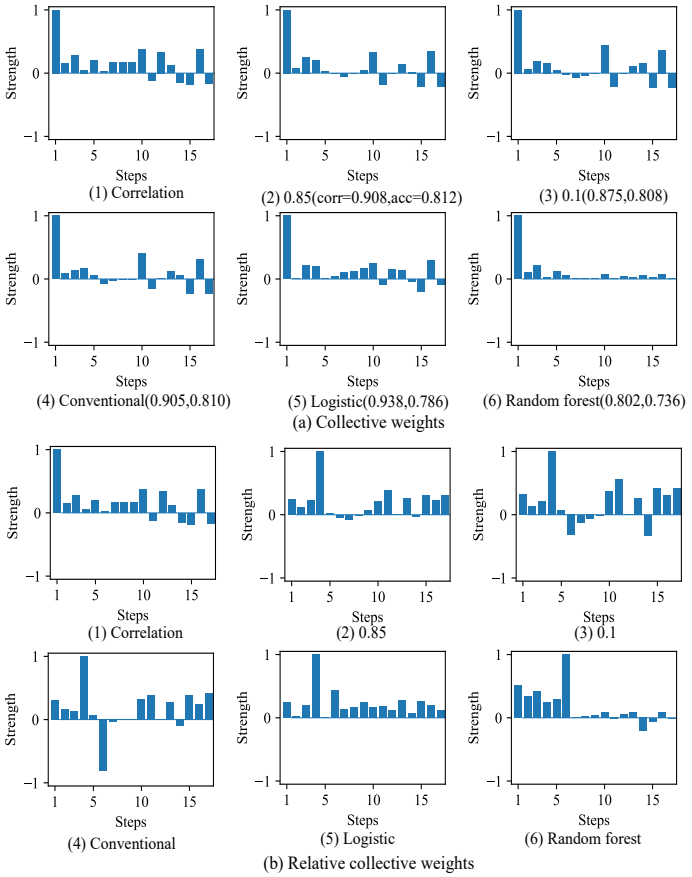


Fig. 11. Collective weights and related importance measures (a) and relative collective ones (b) for the present method (1) to the random forest (6) for the traffic data set. The numbers in the figure show the correlation coefficients (left) and generalization accuracy (right).

and accuracy were slightly larger than those by the weight decay. The logistic regression analysis in Figure 11(5) produced the largest correlation of 0.938, but the accuracy was lower, with the second worst value of 0.786, slightly better than the 0.736 of the random forest. Finally, when the random forest was used, the correlation and accuracy were the lowest in Figure 11(6).

When we used the relative correlation coefficients relative to the absolute original correlations between inputs and targets in Figure 11(b1)-(b5), the fourth input (vehicle excess) showed higher values for the cost reduction, weight decay, conventional method, and logistic regression analysis. This suggests that, in addition to the first input, the fourth input could play an important role in traffic behavior.

B. Facility for the Elderly Data Set

1) Experimental Outline: The second experiment used the data set of the facility for the elderly [64], in which we tried to distinguish between male and female residents and to identify the essential needs of residents for the facility. The objective of the experiment aimed to improve the services provided by the facility. The number of input variables was seven, and the number of patterns was 1,000. We used the same parameter values presented in the first experimental results on the traffic data set for easy reproduction of the results.

2) Syntagmatic and Paradigmatic Compression : The results show that the present method produced very high correlation coefficients, with almost perfect correlations with the original correlation coefficients between inputs and targets. The weight decay and conventional method could produce weights with higher correlations, but they were lower than those by the present method.

Figure 12(a) shows the syntagmatic (left) and paradigmatic (right) compression when 100 different initial conditions and 100 different subsets of data were used, where the parameter β_1 was set to 0.85 for cost reduction. As can be seen in the left-hand box on the syntagmatic compression, except for five low correlations between compressed weights and original correlations, the correlations became close to one. For the paradigmatic compression on the right-hand side, the correlations became immediately close to one, meaning that paradigmatic compression produced original correlations

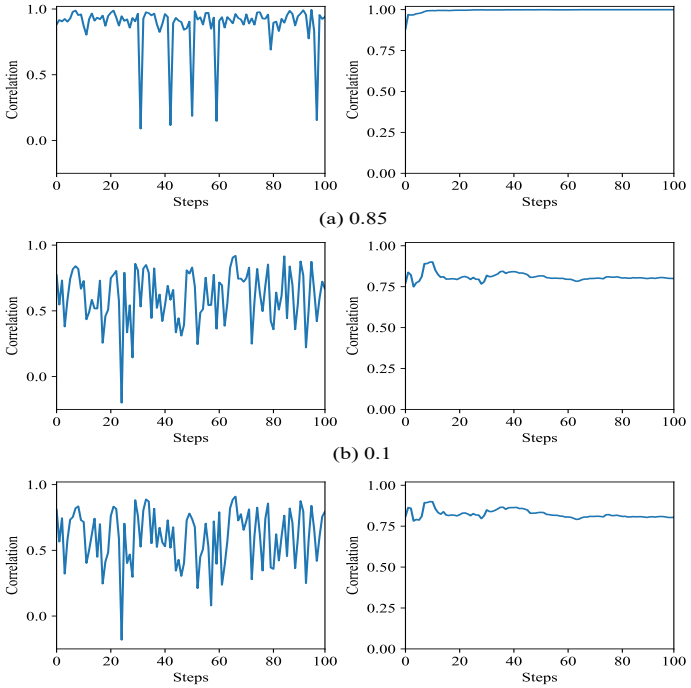


Fig. 12. Correlations by the syntagmatic compression (left) and by the paradigmatic compression (right) when the parameter β_1 was 0.85 (a), α was 0.1 for the weight decay (b), and by conventional method (c) for the facility for the elderly data set.

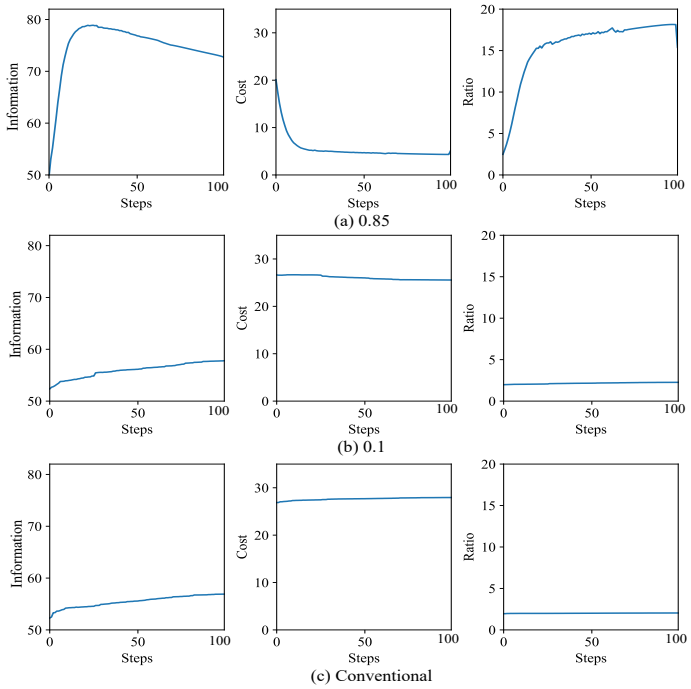


Fig. 13. Selective information (left), cost (middle), and ratio (right) when the parameter β_1 was 0.85 (a), α was 0.1 for weight decay (b), and by the conventional method (c) for the facility for the elderly data set

between inputs and targets over compressed weights. Figures 12(b) and (c) show syntagmatic (left) and paradigmatic compression (right) by the weight decay ($\alpha = 0.1$) and by the conventional method without weight decay. The two methods produced quite similar results for both types of compression. However, the correlations were lower than those by the present method. In particular, correlations with syntagmatically compressed weights fluctuated extensively.

These results showed that the present method could produce collective weights close to the original correlation coefficients between inputs and targets. We could obtain those results almost independently of different initial conditions and different inputs. On the contrary, the conventional methods produced lower correlations, and they fluctuated considerably.

3) Selective Information, Cost, and Ratio: The results show that the selective information increased up to a certain point, and then it decreased in the end. However, due to the smaller cost, the ratio increased gradually for all the learning steps. On the contrary, the weight decay and conventional method could not sufficiently increase selective information, and in addition, they could not decrease the cost. Then, ratios became smaller almost over all different runs.

Figure 13(a) shows selective information (left), cost (middle), and ratio of information to its cost (right) when the parameter β_1 was 0.85. The selective information increased, and then decreased gradually. Because information was not forced to be increased, the information could not naturally continue to be sufficiently increased. However, the cost constantly decreased when the number of learning steps increased. Then, the ratio of information to its cost increased rapidly. On the contrary, Figure 13(b) and (c) show the results by the weight decay and conventional method. The selective information slightly increased, but the cost remained large, and the ratios remained small for all the learning steps. The results confirmed that the present method could decrease the cost sufficiently to increase the selective information. Then, the ratio of information and cost increased gradually.

C. Weights and Individual Potentiality

The results show that the number of strong weights became smaller when the hidden layers became higher. On the contrary, the weight decay and conventional method could not produce explicit regularity over connection weights.

Figure 14(a) shows weights (1) and corresponding individual potentiality (2) when the parameter β_1 was 0.85. As can be seen in the figure, the number of strong connection weights gradually decreased when the hidden layers became higher. In addition, for the individual potentiality, we could see several groups of connection weights responding to the inputs in the same way. On the contrary, by the weight decay (b) and conventional method without weight decay (c), no regularity over connection weights and individual potentiality could be seen.

The results showed that the present method could decrease the number of strong connection weights, and connection weights cooperated with each other as several groups to transmit the information.

1) Partial Compression: The results show that the present method could extract information on inputs in the lower hidden

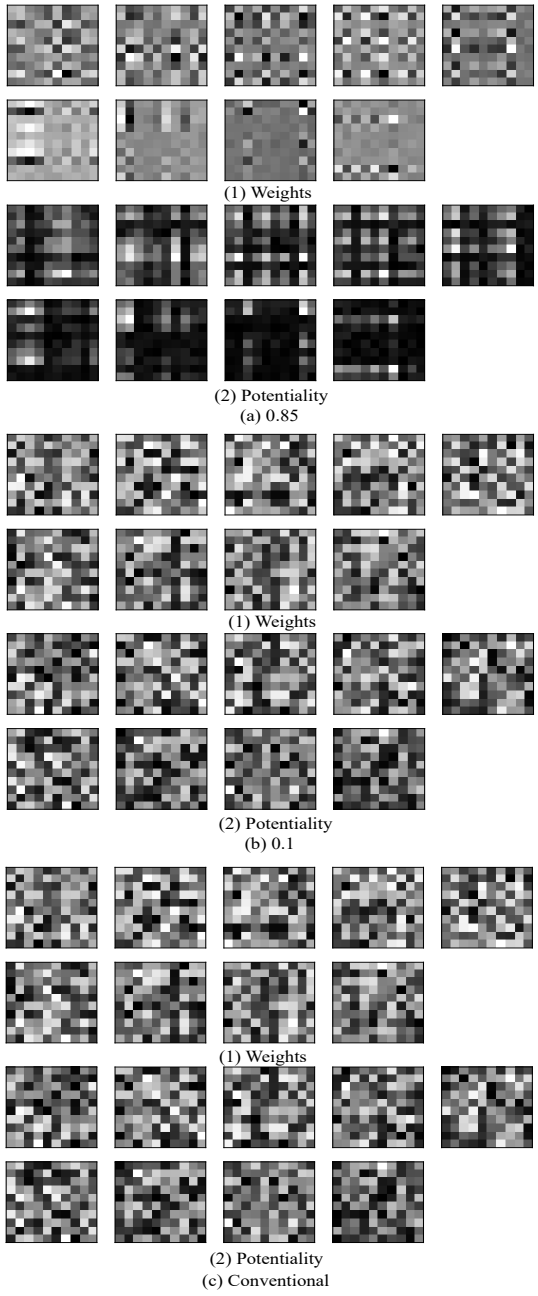


Fig. 14. Weights (a) and individual potentiality (b), when the parameter β_1 was 0.85 (a), α was 0.1 for the weight decay (b), and by the conventional method (c) for the facility for the elderly data set.

layers. On the contrary, the weight decay and conventional method could not extract the information in the hidden layers.

Figure 15 shows partially compressed weights by the present method (a), weight decay (b), and conventional method (c). The present method in Figure 15(a) produced strong partially compressed weights in the beginning, and the strength of compressed weights became smaller when the layers became higher. On the contrary, by the weight decay in Figure 15(b) and conventional method in Figure 15(c), the strength of partially compressed weights remained small until the final compression was applied.

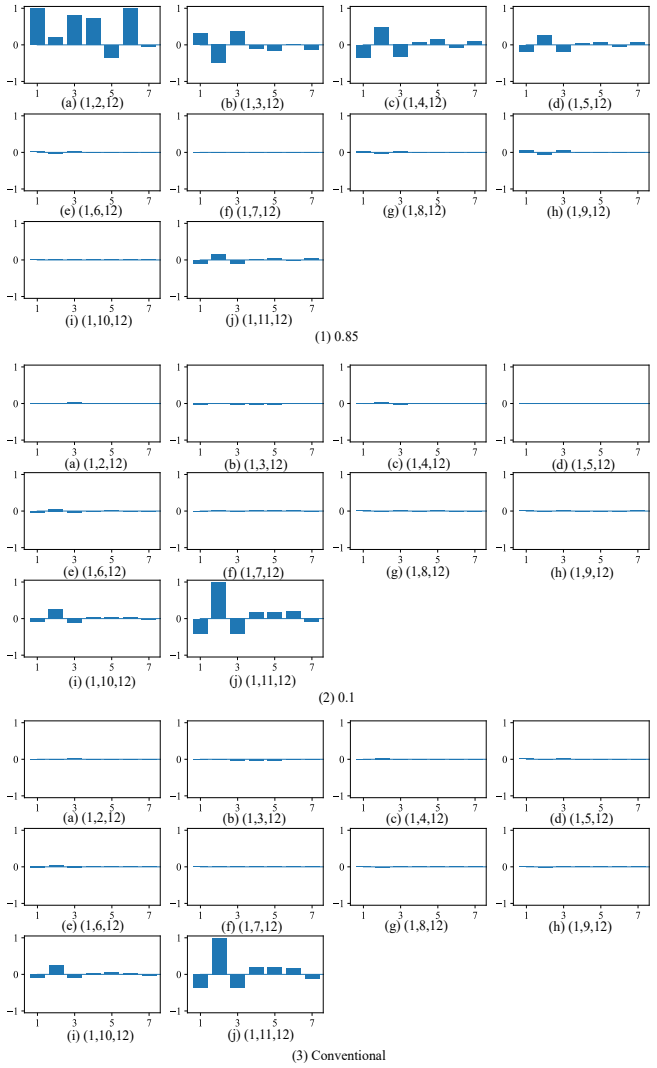


Fig. 15. Partially compressed weights, when the parameter β_1 was 0.85 (1), α was 0.1 for the weight decay (2), and by the conventional method (3) for the facility for the elderly data set.

These results show that the present method tried to acquire information content from inputs, and this information gradually decreased when going through many layers. On the contrary, the other conventional methods could not acquire enough information until we reached the final layer.

2) Full Compression: The results show that the present method could extract almost perfect correlations with higher generalization accuracy, compared with the weight decay and conventional method. The correlation coefficient was still higher than that obtained by the logistic regression.

Figure 16(a) shows correlation coefficients between inputs and targets of the original data set (1); collective weights by the present method, with the highest correlation coefficient (2), weight decay (3), and conventional method (4); regression coefficients by the logistic regression analysis (5); and prediction importance by the random forest (6). As can be seen in Figure 16(a2), the correlation was rounded to one (perfect correlation), and the generalization accuracy of 0.566 was the

second best, behind the 0.568 by the weight decay. Figure 16(a3) shows a case with the best generalization accuracy of 0.568 by the weight decay. The correlation coefficient decreased to 0.8. The conventional method in Figure 16(a4) produced the correlation of 0.804, and the accuracy was 0.566. The logistic regression in Figure 16(a5) produced a high correlation of 0.985, but the accuracy decreased to 0.551. Finally, the random forest produced the worst accuracy of 0.541 and the worst correlation of -0.294.

Figure 16(b) shows the relative collective weights. As shown in Figure 16(b2), the present method with the best correlation coefficient produced an almost even score over all inputs. On the contrary, the weight decay and conventional method in Figure 16(b3) and (b4) produced negative values for the latter three inputs. The logistic regression analysis in Figure 16(b5) produced evenly distributed and positive relative weights, but the strength varied considerably. Finally, the prediction importance in Figure 16(b6) by the random forest produced importance values completely different from other measures.

The results show that the present method with many hidden layers could produce connection weights close to the original correlation coefficients, keeping generalization sufficiently good. These results demonstrate that multi-layered neural networks could be transformed to identify individual correlation coefficients, and if differences between them and their original correlations were considerably large, neural networks tried to use non-linear and complicated connection weights.

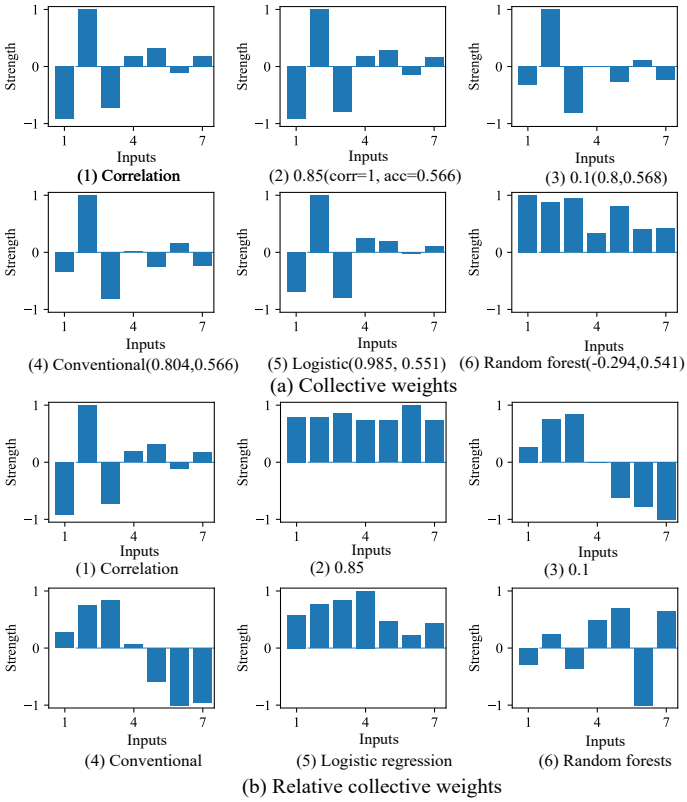


Fig. 16. Collective weights (a) and relative collective weights (b) for the facility for the elderly data set. Figures 1 to 6 denote the original correlation, compressed weights by the present method with best correlation, weight decay, conventional method, logistic regression, and random forest method.

D. Wine Data Set

1) Experimental Outline: The data set was composed of red and white wine samples from the north of Portugal, where we tried to distinguish between red and white ones based on 12 variables [65]. The number of samples was 6,497. Because the resultant correlation coefficients were lower than those in the above sections by the simple cost reduction, we tried to use the two-steps selectivity or cost control method. All the parameters used in this experiment were forced to be set to the same values as those in the above two experiments, except for the parameter β for the initial learning stage. The parameter β_2 was larger than one, actually, 1.3, in the beginning of learning (until one third of the total learning steps was reached). Then, the parameter was reduced to the normal 0.85. Thus, this method lay in cost augmentation in the first place, and then the cost was reduced.

2) Correlation Coefficients : The correlation coefficients between compressed weights and the original correlations computed by the data set were relatively high for all methods. However, the present method could produce higher correlations for all different runs.

Figure 17 shows correlation coefficients between compressed weights and the original correlations when the parameter β_2 was 1.3 (first part) and when β_1 was 0.85 (remaining part) (a), when the decay parameter α was 0.1 (b), and by the conventional method without the weight decay and selectivity control (c). As shown in the left-hand box in Figure 17(a), the correlation coefficients by the present method fluctuated in the processes of syntagmatic compression. However, in the processes of paradigmatic processing in the right-hand box in Figure 17(a), the correlation coefficients were very stable and close to those from the beginning. On the contrary, the correlation coefficients by the weight decay in Figure 17(b) and by the conventional method in Figure 17(c) tended to decrease gradually when the number of different runs increased. In addition, the correlation coefficients by the syntagmatic and paradigmatic compression were smaller than those by the present method.

The results show that the simplified two-step method could produce higher correlation coefficients for syntagmatic and paradigmatic compression.

3) Selective Information, Cost, and Ratio: The initial steps of learning by the simplified method could increase the cost considerably, keeping the selective information smaller. Then, in the subsequent steps, the selective information increased rapidly and, at the same time, the cost decreased considerably. Finally, the ratio of selective information to its cost increased in the subsequent steps. On the contrary, the other methods could not increase the selective information and decrease the cost.

Figure 18 shows the selective information (left), cost (middle), and the ratio of the information to the cost (right) by the present method (a) and the weight decay (b), and the ratio (c). Figure 18(a) shows the results by the present method when the

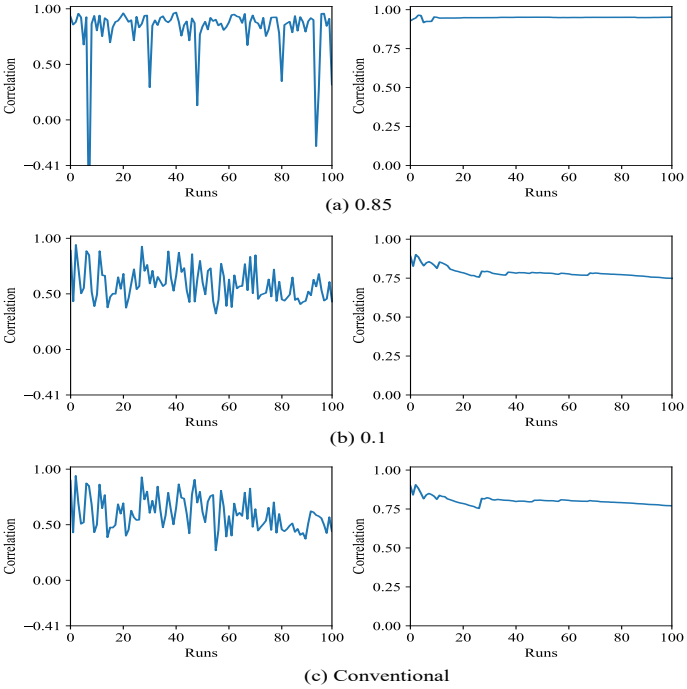


Fig. 17. Correlations between compressed weights and correlations of the original data set by the syntagmatic compression (left) and by the paradigmatic compression (right) when the parameter β_2 was 1.3 and β_1 was 0.85 (a), α was 0.1 for the weight decay (b), and by the conventional method (c) for the wine data set.

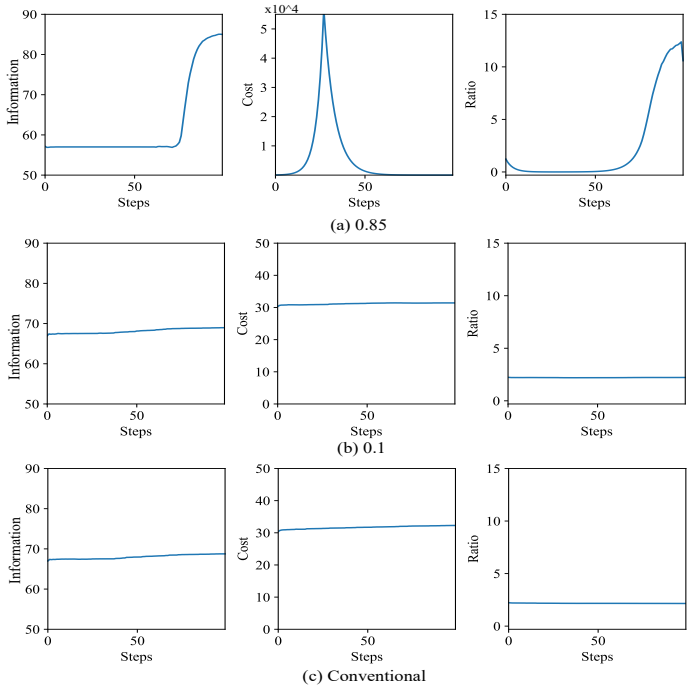


Fig. 18. Selective information (left), cost (middle), and ratio (right) when the parameter β_2 was 1.3 and β_1 was 0.85 (a), α was 0.1 for the weight decay (b), and by the conventional method (c) for the wine data set

parameter β_2 was 1.3 (initial) and β_1 was 0.85 (remaining). As can be seen in the figure, selective information was kept small in the initial steps of learning. Then, the selective information increased considerably in the remaining learning steps. The cost (middle) was forced to be increased up to a point where a further increase in the cost degraded the performance, and the cost was forced to be decreased considerably in the end. On the contrary, by using the weight decay in Figure 18(b), and the conventional method in Figure 18(c), the selective information had relatively high values without changes. The costs, shown in the figures in the middle, were larger than those by the present method. Finally, the ratio of selective information and its cost remained small for all learning steps.

The experimental results show that the present method could increase and then decrease the cost and correspondingly decrease and increase the selective information. On the other hand, the weight decay and conventional method could not well control the selective information and its cost.

4) Weights and Individual Potentiality: The weights for all hidden layers became relatively sparse by the present method, and in particular, the individual potentiality showed this sparsity tendency. However, the property of sparsity of the present method was not so different from that by the conventional methods.

Figure 19(a) shows connection weights (1) and the corresponding individual potentialities (2) by the present method. As can be seen in the figure, in particular, by seeing the individual potentialities, the number of stronger weights tended to be smaller, and weights became more selective by the present method. In the same way, by using the weights decay in Figure 19(b) and conventional method in Figure 19(c), the number of stronger weights seemed to be smaller. In particular, when we examined the individual potentialities, the sparse properties could be seen. However, we could not see large differences among the three methods. The results show that the final weights by the three methods seemed to be approximately the same in terms of their sparseness, though the present method could produce slightly more selective weights. This is due to the large parameter value β for the present method, and this large parameter value, accompanied by the large cost, prevented the present method from producing more selective states.

5) Partial Compression: The partially compressed weights produced a similar tendency for all three methods. The compressed weights by all the methods could not show explicit characteristics until the final and output layer was considered.

Figures 20(a), (b) and (c) show partially compressed weights by the present method, the weight decay, and the conventional method, respectively. Though the final compressed weights were different, all partially compressed weights were kept small. Only in the final compression step did compressed weights tend to be reasonably large. This can be explained by the fact that the selective information was forced to be smaller by increasing the cost. Then, selective information on inputs tended to disappear by the present method. This means that the information content in inputs could not be used to relate inputs and outputs.

6) Full Compression: The results by the full compression show that the present method could produce collective weights

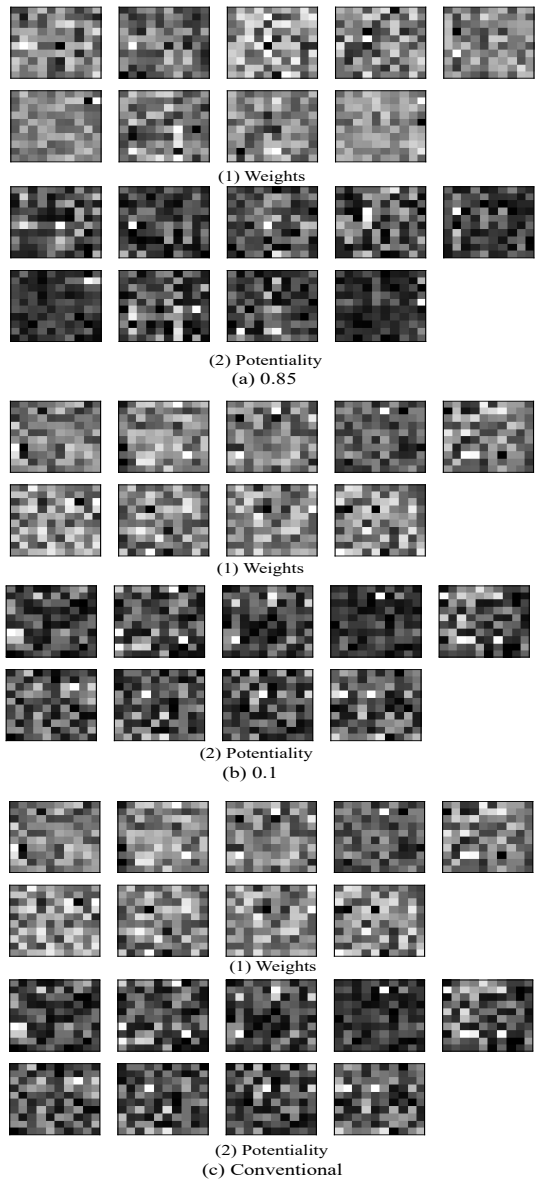


Fig. 19. Weights (a) and individual potentiality (b), when the parameter β_2 was 1.3 and β_1 was 0.85 (a), α was 0.1 for the weight decay (b), and by the conventional method (c) for the wine data set.

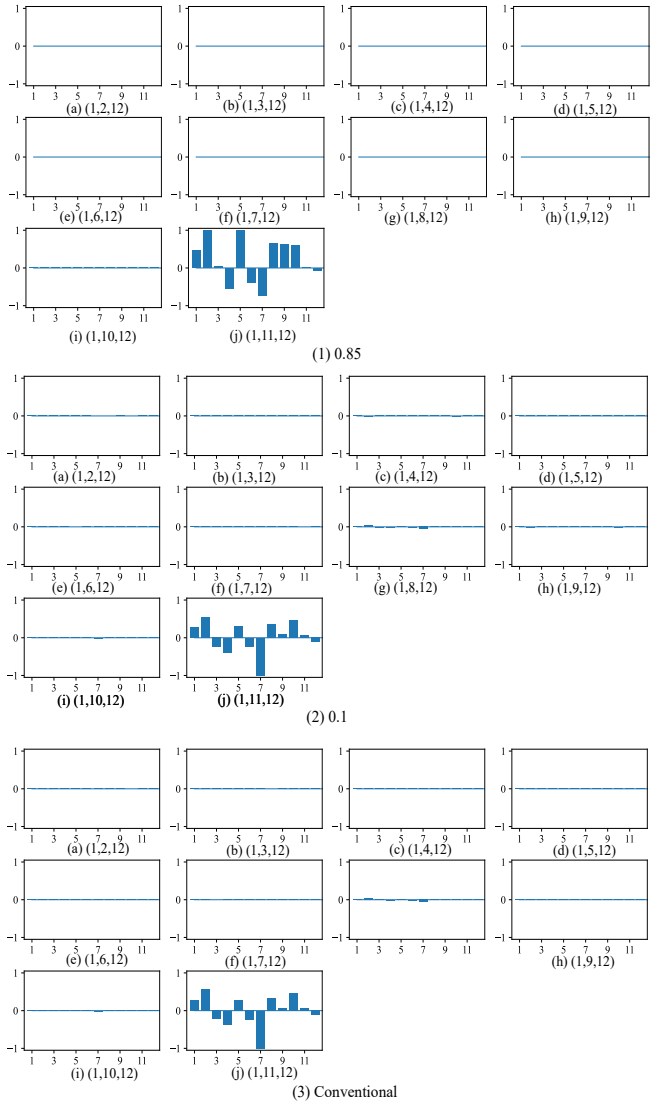


Fig. 20. Partially compressed weights when the parameter β_2 was 1.3 and β_1 was 0.85 (1), α was 0.1 for the weight decay (2), and by the conventional method (3) for the wine data set.

close to the original correlation coefficients. Though the conventional logistic regression analysis could produce similar correlation coefficients, its accuracy rate was smaller than that by the present method.

Figure 21 shows the correlation coefficients and the fully compressed weights by five methods. By using the present method, the correlation coefficient became 0.952, and the accuracy was 0.952 in Figure 21(a2). In addition, the similarity between the original correlation and compressed weights was observed in the positive relative weights for all inputs in Figure 21(b2). By using the weight decay, the correlation decreased to 0.749, and the accuracy rate was the highest one of 0.996 in Figure 21(a3). The conventional method could also produce the highest accuracy of 0.996, but the correlation coefficient decreased to 0.771 in Figure 21(a4). Though those methods produced lower correlation coefficients than those by the present method, the relative collective weights were positive except for input No.11 in Figure 21(b3) and (b4). Then, by the logistic regression analysis in Figure 21(a5), the correlation was 0.937, which was lower than the 0.952 by the present method. In addition, the accuracy by the present method was 0.995, larger than the 0.989 by the logistic regression analysis. Finally, the random forest in Figure 21(a6) produced the lowest correlation of -0.075, though the accuracy was the highest at 0.996. The random forest produced importance measures quite different from those by the other methods. The results show that the present method could produce the highest correlation coefficient, keeping high accuracy rates.

IV. Conclusion

The present paper aimed to propose a new type of interpretation method for multi-layered neural networks. The method

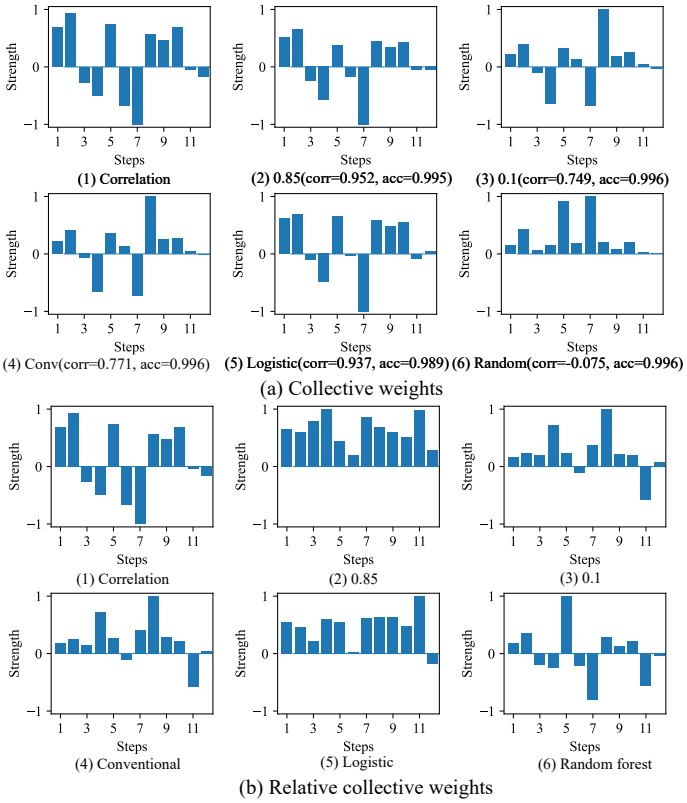


Fig. 21. Collective weights (a) and relative collective weights (b) by five methods for the wine data set. The numbers in the figure represent the correlation coefficients (left) and accuracy (right).

lies in considering all possible internal representations generated by multi-layered neural networks, in which we suppose that all representations by multi-layered neural network have the same status, meaning that many representations should be created by seeing a data set from different points of view.

One of the main shortcomings of interpretation methods of neural networks is that they try to understand only one aspect of representations. For example, they tried to show what components in a neural network can be responsible for a specific input. This type of individual interpretation has been extensively used in the present state of neural networks. In particular, in the CNN, dealing with image data sets, it has been extensively used, because it is easy to understand the specific input and the corresponding component intuitively. However, those corresponding components should be changed, sometimes drastically, by using different initial conditions, which is one of the main problems of neural networks. In our approach, we suppose that different representations created by different initial conditions can be used to explain the inference mechanism of neural networks. This means that we can interpret the representations from different points of view.

In actual learning, we have different internal representations in the course of learning. In addition, by different initial conditions and inputs, we have also different internal representations. We first take into account different representations in the course of learning, which can be called "syntagmatic compression." In the syntagmatic compression, all weights created in the course of learning with a specific initial condition are averaged and compressed. Then, all syntagmatically compressed weights are again averaged, which is called "paradigmatic compression." With syntagmatically and paradigmatically compressed representations, we can interpret neural networks in terms of collective interpretation, namely, from many viewpoints.

The collective compression was flexibly controlled by controlling the selective information. However, we proposed a more simplified method to control the selective information, that of controlling the cost in terms of weight strength. This means that the selective information control eventually corresponds to the cost control, which is much simpler to be implemented in actual learning. In addition, we proposed a new method to control the selective information by its cost, where the cost is first increased, and then it is decreased. This increase in the cost, corresponding to a decrease in selective information aimed to eliminate information on input patterns as much as possible. This is because the information represented by the inputs of neural networks cannot be used to relate the inputs to the corresponding targets.

The method was applied to three real data sets: the traffic, facility for the elderly, and wine data sets. In the first two cases, we could see that the selective information could be increased and, at the same time, the cost could be decreased in terms of the sum of absolute weights. The final collective weights by the present method were very close to the correlation coefficients between inputs and targets of the original data set. This could be explained by the fact that the present method could extract much information from inputs; on the contrary, the other conventional methods could not extract sufficient information from the inputs, but they were dependent exclusively on the outputs. In the experimental results of the third data set, the wine data set, the original information by the corresponding inputs was forced to be eliminated by the cost augmentation in the initial learning steps. This means that the input variables cannot represent well the information on the relations between inputs and outputs. This could be observed in the results of partial compression, where partially compressed weights could not show any regularity until the final output layer was included.

We should point out here two problems with the present method: extraction of features specific to input patterns, and how to control selective information. One of the main problems is how to identify differences between the original correlations and specific ones by the present method. We proposed a method to extract relative differences between them. However, we should develop a more refined method to distinguish between the original and new features dealt with by the present method. Second, we proposed a method to eliminate the selective information in the initial learning steps and applied it to the third data set. However, we did not know to what level we should eliminate the selective information. Thus, we need to examine more closely the exact effect of information reduction over information augmentation.

Finally, we should mention briefly some future work to be done on robustness and its relation to the selective information. First, while we focused on the interpretation in this paper, the collective concept described in this paper can be naturally applied to generalization accuracy. This is because the collective interpretation tries to interpret the inference mechanism, considering as many different internal representations as possible, including ones with higher and lower robustness. Our objective is to find some transformation rules from the collective and core ones to more concrete networks with different types of robustness [66], [67], [68]. We think that, for these transformation rules, the selective information control presented here can be of some use.

Though several problems should be solved for the present method to be applied to more practical data sets, the present study surely contributes to the problem of interpretation as well as the relations between selectivity and network performance.

V. Acknowledgments

We would like to thank the reviewers and editors for taking their time to read the drafts of the paper and give valuable comments on how to improve it. In addition, special thanks go to Mitali Das for reading and correcting the paper. Finally, this paper has been written, based on two papers: "Controlling Individual and Collective Information for Generating Interpretable Models of Multi-Layered Neural Networks" [2], presented in INTELLI2021 and "Selective Information-Driven Learning for Producing Interpretable Internal Representations in Multi-Layered Neural Networks" [1] in COGNITIVE2021.

References

[1] R. Kamimura, "Selective information-driven learning for producing interpretable internal representations in multi-layered neural networks," in *COGNITIVE 2021, The Thirteenth International Conference on Advanced Cognitive Technologies and Applications*, pp. 20–27, IARIA, 2021.

[2] R. Kamimura, "Controlling individual and collective information for generating interpretable models of multi-layered neural networks," in *INTELLI 2021, The Tenth International Conference on Intelligent Systems and Applications*, pp. 27–35, IARIA, 2021.

[3] D. E. Rumelhart, G. E. Hinton, and R. Williams, "Learning internal representations by error propagation," in *Parallel Distributed Processing* (D. E. Rumelhart and G. E. H. et al., eds.), vol. 1, pp. 318–362, Cambridge: MIT Press, 1986.

[4] R. Andrews, J. Diederich, and A. B. Tickle, "Survey and critique of techniques for extracting rules from trained artificial neural networks," *Knowledge-based systems*, vol. 8, no. 6, pp. 373–389, 1995.

[5] M. Ishikawa, "Structural learning with forgetting," *Neural Networks*, vol. 9, no. 3, pp. 509–521, 1996.

[6] J. A. Alexander and M. C. Mozer, "Template-based procedures for neural network interpretation," *Neural Networks*, vol. 12, pp. 479–498, 1999.

[7] M. Ishikawa, "Rule extraction by successive regularization," *Neural Networks*, vol. 13, no. 10, pp. 1171–1183, 2000.

[8] D. E. Rumelhart and D. Zipser, "Feature discovery by competitive learning," in *Parallel Distributed Processing* (D. E. Rumelhart and G. E. H. et al., eds.), vol. 1, pp. 151–193, Cambridge: MIT Press, 1986.

[9] D. E. Rumelhart and J. L. McClelland, "On learning the past tenses of English verbs," in *Parallel Distributed Processing* (D. E. Rumelhart, G. E. Hinton, and R. J. Williams, eds.), vol. 2, pp. 216–271, Cambrige: MIT Press, 1986.

[10] B. Goodman and S. Flaxman, "European union regulations on algorithmic decision-making and a right to explanation," *arXiv preprint arXiv:1606.08813*, 2016.

[11] J. L. Castro, C. J. Mantas, and J. M. Benítez, "Interpretation of artificial neural networks by means of fuzzy rules," *IEEE Transactions on Neural Networks*, vol. 13, no. 1, pp. 101–116, 2002.

[12] T. Q. Huynh and J. A. Reggia, "Guiding hidden layer representations for improved rule extraction from neural networks," *IEEE Transactions on Neural Networks*, vol. 22, no. 2, pp. 264–275, 2011.

[13] F. Wang and C. Rudin, "Falling rule lists," in *Artificial Intelligence and Statistics*, pp. 1013–1022, 2015.

[14] B. Letham, C. Rudin, T. H. McCormick, D. Madigan, *et al.*, "Interpretable classifiers using rules and bayesian analysis: Building a better stroke prediction model," *The Annals of Applied Statistics*, vol. 9, no. 3, pp. 1350–1371, 2015.

[15] A. Nguyen, J. Yosinski, and J. Clune, "Understanding neural networks via feature visualization: A survey," in *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning*, pp. 55–76, Springer, 2019.

[16] D. Erhan, Y. Bengio, A. Courville, and P. Vincent, "Visualizing higher-layer features of a deep network," *University of Montreal*, vol. 1341, 2009.

[17] A. Nguyen, J. Clune, Y. Bengio, A. Dosovitskiy, and J. Yosinski, "Plug & play generative networks: Conditional iterative generation of images in latent space," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4467–4477, 2017.

[18] A. Mahendran and A. Vedaldi, "Understanding deep image representations by inverting them," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 5188–5196, 2015.

[19] A. Nguyen, A. Dosovitskiy, J. Yosinski, T. Brox, and J. Clune, "Synthesizing the preferred inputs for neurons in neural networks via deep generator networks," in *Advances in neural information processing systems*, pp. 3387–3395, 2016.

[20] A. v. d. Oord, N. Kalchbrenner, and K. Kavukcuoglu, "Pixel recurrent neural networks," *arXiv preprint arXiv:1601.06759*, 2016.

[21] K. Simonyan, A. Vedaldi, and A. Zisserman, "Deep inside convolutional networks: Visualising image classification models and saliency maps," *arXiv preprint arXiv:1312.6034*, 2013.

[22] J. Khan, J. S. Wei, M. Ringner, L. H. Saal, M. Ladanyi, F. Westermann, F. Berthold, M. Schwab, C. R. Antonescu, C. Peterson, *et al.*, "Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks," *Nature medicine*, vol. 7, no. 6, pp. 673–679, 2001.

[23] D. Baehrens, T. Schroeter, S. Harmeling, M. Kawanabe, K. Hansen, and K.-R. MÃžller, "How to explain individual classification decisions," *Journal of Machine Learning Research*, vol. 11, no. Jun, pp. 1803–1831, 2010.

[24] D. Smilkov, N. Thorat, B. Kim, F. Viégas, and M. Wattenberg, "Smoothgrad: removing noise by adding noise," *arXiv preprint arXiv:1706.03825*, 2017.

[25] M. Sundararajan, A. Taly, and Q. Yan, "Axiomatic attribution for deep networks," *arXiv preprint arXiv:1703.01365*, 2017.

[26] G. Montavon, A. Binder, S. Lapuschkin, W. Samek, and K.-R. Müller, "Layer-wise relevance propagation: an overview," in *Explainable AI: interpreting, explaining and visualizing deep learning*, pp. 193–209, Springer, 2019.

[27] S. Bach, A. Binder, G. Montavon, F. Klauschen, K.-R. Müller, and W. Samek, "On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation," *PloS one*, vol. 10, no. 7, p. e0130140, 2015.

[28] S. Lapuschkin, A. Binder, G. Montavon, K.-R. Muller, and W. Samek, "Analyzing classifiers: Fisher vectors and deep neural networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2912–2920, 2016.

[29] F. Arbabzadah, G. Montavon, K.-R. Müller, and W. Samek, "Identifying individual facial expressions by deconstructing a neural network," in *German Conference on Pattern Recognition*, pp. 344–354, Springer, 2016.

[30] I. Sturm, S. Lapuschkin, W. Samek, and K.-R. Müller, "Interpretable deep neural networks for single-trial eeg classification," *Journal of neuroscience methods*, vol. 274, pp. 141–145, 2016.

[31] A. Binder, G. Montavon, S. Lapuschkin, K.-R. Müller, and W. Samek, "Layer-wise relevance propagation for neural networks with local renormalization layers," in *International Conference on Artificial Neural Networks*, pp. 63–71, Springer, 2016.

[32] M. Polanyi, *The tacit dimension*. University of Chicago press, 2009.

[33] E. T. Hall, "Beyond culture. garden city, ny: Anchor," 1976.

[34] N. Carlini and D. Wagner, "Adversarial examples are not easily detected: Bypassing ten detection methods," in *Proceedings of the 10th ACM Workshop on Artificial Intelligence and Security*, pp. 3–14, 2017.

[35] A. Ilyas, S. Santurkar, D. Tsipras, L. Engstrom, B. Tran, and A. Madry, "Adversarial examples are not bugs, they are features," *arXiv preprint arXiv:1905.02175*, 2019.

[36] A. Kurakin, I. Goodfellow, and S. Bengio, "Adversarial machine learning at scale," *arXiv preprint arXiv:1611.01236*, 2016.

[37] C. Bucilu, R. Caruana, and A. Niculescu-Mizil, "Model compression," in *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 535–541, ACM, 2006.

[38] J. Ba and R. Caruana, "Do deep nets really need to be deep?," in *Advances in neural information processing systems*, pp. 2654–2662, 2014.

[39] G. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," *arXiv preprint arXiv:1503.02531*, 2015.

[40] R. Adriana, B. Nicolas, K. S. Ebrahimi, C. Antoine, G. Carlo, and B. Yoshua, "Fitnets: Hints for thin deep nets," *Proc. ICLR*, 2015.

[41] P. Luo, Z. Zhu, Z. Liu, X. Wang, and X. Tang, "Face model compression by distilling knowledge from neurons," in *Thirtieth AAAI Conference on Artificial Intelligence*, 2016.

[42] J. O. Neill, "An overview of neural network compression," *arXiv preprint arXiv:2006.03669*, 2020.

[43] J. Gou, B. Yu, S. J. Maybank, and D. Tao, "Knowledge distillation: A survey," 2020.

[44] Y. Cheng, D. Wang, P. Zhou, and T. Zhang, "A survey of model compression and acceleration for deep neural networks," 2020.

[45] R. Kamimura, "Neural self-compressor: Collective interpretation by compressing multi-layered neural networks into non-layered networks," *Neurocomputing*, vol. 323, pp. 12–36, 2019.

[46] R. Linsker, "Self-organization in a perceptual network," *Computer*, vol. 21, no. 3, pp. 105–117, 1988.

[47] R. Linsker, "How to generate ordered maps by maximizing the mutual information between input and output signals," *Neural computation*, vol. 1, no. 3, pp. 402–411, 1989.

[48] R. Linsker, "Local synaptic learning rules suffice to maximize mutual information in a linear network," *Neural Computation*, vol. 4, no. 5, pp. 691–702, 1992.

[49] R. Linsker, "Improved local learning rule for information maximization and related applications," *Neural networks*, vol. 18, no. 3, pp. 261–265, 2005.

[50] K. Torkkola, "Nonlinear feature transform using maximum mutual information," in *Proceedings of International Joint Conference on Neural Networks*, pp. 2756–2761, 2001.

[51] K. Torkkola, "Feature extraction by non-parametric mutual information maximization," *Journal of Machine Learning Research*, vol. 3, pp. 1415–1438, 2003.

[52] J. M. Leiva-Murillo and A. Artés-Rodríguez, "Maximization of mutual information for supervised linear feature extraction," *Neural Networks, IEEE Transactions on*, vol. 18, no. 5, pp. 1433–1441, 2007.

[53] M. M. Van Hulle, "The formation of topographic maps that maximize the average mutual information of the output responses to noiseless input signals," *Neural Computation*, vol. 9, no. 3, pp. 595–606, 1997.

[54] J. C. Principe, D. Xu, and J. Fisher, "Information theoretic learning," *Unsupervised adaptive filtering*, vol. 1, pp. 265–319, 2000.

[55] J. C. Principe, *Information theoretic learning: Renyi's entropy and kernel perspectives*. Springer Science & Business Media, 2010.

[56] A. S. Morcos, D. G. Barrett, N. C. Rabinowitz, and M. Botvinick, "On the importance of single directions for generalization," *stat*, vol. 1050, p. 15, 2018.

[57] I. Rafegas, M. Vanrell, L. A. Alexandre, and G. Arias, "Understanding trained cnns by indexing neuron selectivity," *Pattern Recognition Letters*, vol. 136, pp. 318–325, 2020.

[58] J. Ukita, "Causal importance of low-level feature selectivity for generalization in image recognition," *Neural Networks*, vol. 125, pp. 185–193, 2020.

[59] M. L. Leavitt and A. Morcos, "Selectivity considered harmful: evaluating the causal impact of class selectivity in dnns," *arXiv preprint arXiv:2003.01262*, 2020.

[60] W. J. Johnston, S. E. Palmer, and D. J. Freedman, "Nonlinear mixed selectivity supports reliable neural computation," *PLoS computational biology*, vol. 16, no. 2, p. e1007544, 2020.

[61] M. L. Leavitt and A. S. Morcos, "On the relationship between class selectivity, dimensionality, and robustness," *arXiv preprint arXiv:2007.04440*, 2020.

[62] P. Lennie, "The cost of cortical computation," *Current biology*, vol. 13, no. 6, pp. 493–497, 2003.

[63] C. Affonso, R. J. Sassi, and R. P. Ferreira, "Traffic flow breakdown prediction using feature reduction through rough-neuro fuzzy networks," in *The 2011 International Joint Conference on Neural Networks*, pp. 1943–1947, IEEE, 2011.

[64] U. Kenji, *Text mining (in Japanese)*. Asakura-shoten, 2021.

[65] P. Cortez, A. Cerdeira, F. Almeida, T. Matos, and J. Reis, "Modeling wine preferences by data mining from physicochemical properties," *Decision Support Systems*, vol. 47, no. 4, pp. 547–553, 2009.

[66] S. Zheng, Y. Song, T. Leung, and I. Goodfellow, "Improving the robustness of deep neural networks via stability training," in *Proceedings of the ieee conference on computer vision and pattern recognition*, pp. 4480–4488, 2016.

[67] N. Carlini and D. Wagner, "Towards evaluating the robustness of neural networks," in *2017 ieee symposium on security and privacy (sp)*, pp. 39–57, IEEE, 2017.

[68] B. Liu, C. Malon, L. Xue, and E. Kruus, "Improving neural network robustness through neighborhood preserving layers," in *25th International Conference on Pattern Recognition Workshops, ICPR 2020*, pp. 179–195, Springer Science and Business Media Deutschland GmbH, 2021.

Time-Efficient Techniques for Improving Student and Instructor Success in Online Courses

Julie R. Newell
Radow College of Humanities and Social Sciences
Kennesaw State University
Kennesaw, GA, USA
jnewell2@kennesaw.edu

Stephen Bartlett
Radow College of Humanities and Social Sciences
Kennesaw State University
Kennesaw, GA, USA
sbartlet@kennesaw.edu

Deborah Mixson-Brookshire
Coles College of Business
Kennesaw State University
Kennesaw, GA, USA
dmixson@kennesaw.edu

Julie Moore
Bagwell College of Education
Kennesaw State University
Kennesaw, GA, USA
jmoor151@kennesaw.edu

Tamara Powell
Radow College of Humanities and Social Sciences
Kennesaw State University
Kennesaw, GA, USA
tpowel25@kennesaw.edu

Sam Lee
Radow College of Humanities and Social Sciences
Kennesaw State University
Kennesaw, GA, USA
slee229@students.kennesaw.edu

Tiffani Reardon Tijerina
Affordable Learning GA Program Manager
University System of Georgia
Athens, GA, USA
tiffani.tijerina@usg.edu

Brayden Milam
Radow College of Humanities and Social Sciences
Kennesaw State University
Kennesaw, GA, USA
bmilam3@kennesaw.edu

Lauren Snider
Radow College of Humanities and Social Sciences
Kennesaw State University
Kennesaw, GA, USA
lsnider1@students.kennesaw.edu

Justin Cochran
Coles College of Business
Kennesaw State University
Kennesaw, GA, USA
jcochr48@kennesaw.edu

***Abstract*—Researchers at a public (state-funded) institution in the United States seek to increase student success rates in online courses by encouraging faculty implementation of research-based strategies in their online courses without significantly increasing faculty workloads. This goal was identified partially in response to a survey of faculty at the institution and partially in response to funding priorities. In the first phase of the faculty development project, the researchers created a training program that provided short, research-based, student-success strategy segments to faculty already enrolled in faculty development. These teaching tools were largely based on pedagogical research and methods long understood within traditional education disciplines but not as obviously applied to online course delivery. In this sense, the professional development modules are innovations to traditional online training. After the training program, the researchers analyzed faculty response to the training to improve design principles and delivery for future development of eLearning materials. In the second phase of the project, the short segments were offered as standalone training modules to anyone who wished to view them. Users were then surveyed regarding their perceptions of the strategies. While the impact of the innovations developed in this student success endeavor are still largely to be determined, preliminary results indicate that faculty find the professional development modules helpful and will be implementing them in their courses.**

Keywords-student success rates; innovation; feedback; open educational resources; training transfer; social media

I. INTRODUCTION

This research project, first presented at eLmL 2021: The Thirteenth International Conference on Mobile, Hybrid, and

On-line Learning [1], aims to increase student success in online courses while being mindful of faculty workload as well as lack of time for faculty development and course redesign. Initially focused on courses offered in RCHSS (Radow College of Humanities and Social Sciences) at KSU (Kennesaw State University), the project has moved beyond that to create an OER (Open Educational Resource) available to any interested individual with internet access.

This paper moves beyond that original conference presentation [1] to examine a wide range of published research that both provides context and underpins the resources we continue to develop. It then analyzes previously unpublished data from a fall 2019 faculty survey examining faculty motivations and expectations. Further, the paper describes the first and second phases of a research project seeking to develop faculty development resources that require minimal investments of institutional resources and faculty time and effort while facilitating the implementation of research-based techniques to improve student success in online instruction.

II. Overview and Rationale

In the United States, education is supported financially by a complicated combination of federal and state funding. In fact, state by state comparisons reveal huge differences in how much a state contributes to its higher education coffers. Government funding of higher education has dropped substantially in recent decades [2]. For example, overall, higher education state funding per student dropped 27% from 2000-2014. State by state, the numbers vary widely. The state of Michigan cut funding by 53% overall during that time while North Dakota increased funding by 31%. Our own state of Georgia cut funding by 17% [3].

When cuts are substantial, the difference is made up in budget cuts at the institutional level (such as reduction in library holdings and elimination of staff and programs) and tuition increases, among other strategies. But, in the United States' political system, the same politicians who strive to cut funding to education also strive to claim that they keep taxes and other expenses low. Therefore, some states rarely allow public (that is, state-funded) institutions to raise tuition to make up for these budget cuts.

With funding so tight, opportunities to gain additional funding to support faculty and students is highly prized, and competition is fierce when such opportunities are announced. Opportunity sometimes comes in the form of "student success dollars," which is funding that can be awarded for initiatives with the intent of bolstering student success. In this case, student success is defined as decreased DFWI rates (students earning Ds or Fs, withdrawing from courses, or taking incomplete grades) and increased retention (the student stays in individual courses and in the university as a whole), progression (the student progresses through a degree program), and graduation (within a proscribed number of years). This definition is often abbreviated as RPG (Retention, Progression, and Graduation). While student success dollars are not tied directly to RPG, our Executive Director for Academic & Fiscal Operations at KSU, Dr. Michael Rothlisberger, explained, "Student success dollars are a systemic example of tying resources to strategy" because meeting RPG targets is seen as "a moral imperative" [4].

To compete for these highly prized student success dollars, our college wants to stand ready with research-based support to facilitate faculty implementation of techniques that foster student success. But just as there is a balancing act that goes along with cutting state funding to higher education and refusing to allow tuition to rise, there is also a balancing act with innovating to improve student success and being mindful of innovations that might challenge academic freedom or increase already strained faculty workloads. For example, for the past ten years, the RCHSS ODE (Office of Digital Education) has offered an award-winning "Build a Web Course Workshop" to support RCHSS faculty in creating and teaching online courses using research-based best practices. This workshop is time intensive, moving faculty through at least eight hours of in-person or synchronous virtual training and then at least another eight hours of training in a learning management system. In preparing to apply for student success funding, college administrators recently looked at the DFWI rates of online courses offered pre-pandemic. Surprisingly, it was determined that there was no significant difference in DFWI rates between classes where faculty had been trained to teach online using best practices versus online courses created and taught by faculty who had not received training.

Our administration theorized the lack of discernable difference may stem from the fact that the ODE delivers training focused on best practices in online and hybrid teaching and not specifically on student success. That is, the courses created by trained faculty may have been better designed because the faculty who created and taught them had been trained in research-based best practices, but the courses may not have specifically implemented student success strategies.

III. Theory and Current Practice

In 1970, Paulo Freire first published *Pedagogy of the Oppressed*, taking issue with what he called the "banking method" of education, where a teacher deposits knowledge into the student, as if the student were a bank [5]. Freire called for a partnership between teacher and student to liberate students and recommended "[p]roblem-posing education" [5]. While Freire did not envision the revolution in education that would be digital learning, his ideas are still the foundation of many of the current student success strategies that are modality agnostic, including transparent pedagogy and HIPs (High-Impact Practices).

A. High-Impact Practices

HIPs were identified by George Kuh in 2008 [6] and lauded throughout academia. Since then, many institutions have made concerted efforts to increase these practices throughout their educational offerings. The practices include "First-Year Seminars and Experiences," "Common

Intellectual Experiences," "Learning Communities," "Writing-Intensive Courses," "Collaborative Assignments and Projects," "Undergraduate Research," "Diversity/Global Learning," "ePortfolios," "Internships," and "Capstone Courses and Projects" [6].

Over a decade later, Indiana University's Center for Postsecondary Research surveyed students regarding HIPs and recommended "three ways educators can assess high-impact practices and ensure high-quality experiences for all students" [7]. The proposed strategies look at which HIPs practices students are being exposed to most and redesign current practices to include more HIPs, evaluate what "high quality" means and ensure that that measure is communicated and upheld, and evaluate student satisfaction with an awareness of critiques of HIPs that "HIPs are centered in the ideology of Whiteness" [7]. It is important to note that the researchers found that satisfaction levels with HIPs practices among students involved in them did not vary with regard to race and ethnic group identification [7].

In a separate study, Kinzie and Kuh specifically analyzed over 15 major contributions to student success literature that covered "college impact, student effort and engagement and importance of the first college year for student success" [8]. Their findings showed that while the information about best practices for student success is widely available, implementation rates are still unacceptably low. "Institutions for various reasons do not faithfully and effectively implement the kinds of promising policies and practices that seem to work elsewhere" [8]. Kinzie and Kuh believe that the failures to effectively implement the strategies and changes to increase student success are due to an approach that is too broad and overwhelming. Schools use the smorgasbord approach for strategies instead of narrowing the options down to the specific strategies that would benefit their particular institutions best. Kinzie and Kuh's suggestion is for institutions to implement Driver diagrams, to streamline goals as well as "to build and test theories for improvement and to clarify what is needed to achieve student success goals" [8].

In addition, Stewart and Nicolazzo take issue with HIPs practices, pointing out that activities like study abroad and internships may pose dangers to trans students. Overall, HIPs practices are more beneficial to students when those students experience fewer levels of oppression and higher levels of privilege. Stewart and Nicolazzo recommend instead that educators implement "trickle up high impact practices (TUHIPs)" creating practices that "recogniz[e] the central importance of working alongside multiply marginalized populations in higher education praxis" and see TUHIPs "as a process through which educators and redistribute human and financial resources toward those who are most vulnerable" [9].

HIPs practices are well-known and widely recognized, and many of them do rely on institutional support (and funding) and financial resources and physical security on the part of students. But not all strategies for student success are dependent upon such resources.

B. Low Resource Strategies

Saundra Yancy McGuire's revolutionary *Teach Students How to Learn* describes strategies to improve student metacognition, which, in turn, improves student learning across many areas [10]. McGuire's strategies and examples are accessible to teachers and students and include fostering growth mindsets, providing clear expectations, and increasing student confidence and self-esteem [10]. As one can see, such strategies do not rely heavily on institutional support and student resources.

Moving beyond metacognition, Mary-Ann Winkelmes identified factors "to promote equitable learning experiences" in higher education [11]. Dubbed "transparent teaching," these practices do not assume that metacognition alone supports increased student learning. Winkelmes concludes practices such as low stakes projects, coaching models, and clear goals and criteria for courses and assignments foster student success [11].

While McGuire and Winklemes do not address modality directly, Flower Darby looks directly at success in online learning in *Small Teaching Online* [12]. Darby, McGuire, and Winkelmes all recommend smaller changes that individual instructors can implement in courses regardless of modality to support student learning and success. Darby's explanation of scaffolding content in courses is not original to her, but her explanation and examples and application to online learning are innovative.

In 2010, Anya Kamenetz predicted the disruption that online learning would bring to education. She challenged educators to recognize the necessary changes that must be made for higher education to be relevant. She foresaw institutions needing to be clear about "meaningful objectives" [13], supporting diverse learners, embracing the benefits that technology can bring to education, and moving toward open educational resources [13].

By the end of the decade, however, when Hoyert and O'Dell examined educational practices at universities across the country, they found that most faculty still used the traditional lecture and textbook system [14]. This data shows that modern alternative pedagogical techniques are not being implemented to meet the needs of a wider range of students. "One of the factors limiting the expansion of the techniques is simply a lack of knowledge of the mechanics and the advantages of particular pedagogies on the part of college faculty" [14]. The authors developed a series of faculty communities that they refer to as pedagogical interest groups, and each group was an interdisciplinary team of six to eight faculty. Each of these groups analyzed and implemented different pedagogical techniques in their classes and shared the results. If a technique was found successful, the instructors

redesigned the courses using that technique. "One of the most impressive outcomes of this project is that all the techniques improved aspects of student learning and with some techniques, the change was immediate" [14]. The authors suggest that "linking measures of teaching effectiveness and recognition for innovative teaching are needed to sustain pedagogical transformation" [14]. Also, it seems clear that small changes can support student learning.

C. Student Expectations

Other factors also challenge efforts toward increased student success. Arum and Roksa studied 2,322 college students "enrolled across a diverse range of campuses" [15]. The researchers found that students don't have a clear idea of why they are in college, how it might benefit them, or how they might succeed in college. While teaching students to think critically is held as the main goal of college education, "[t]hree semesters of college education . . . have a barely noticeable impact on students' skills in critical thinking, complex, reasoning, and writing" [15]. This statistic is measured without regard to whether students are succeeding or failing in their courses, meaning that strategies that improve DFWI rates need to also positively impact student learning and critical thinking skills in order to truly make a difference.

What does the research say online students want? Two separate studies, one by Toufaily, Zalan and Lee [16] and one by Magda and Aslanian [17] found that students choose an online program because 1) it is the program they want, 2) it is the least expensive, and 3) it has a good reputation. During their online study, students surveyed in the UAE responded that they value an instructor "who possesses good interpersonal skills, who is a good leader, who is prepared for class, is precise and teaches so that students can understand" [16]. They want responsive instructors. They want functional elearning platforms, and they value the use of social media platforms to gain a sense of belonging [16]. In a 2010 survey conducted by Penn State University, students were asked how they felt about the Quality Matters criteria. The results revealed that students wanted appropriate assessments, clear guidance on how to access resources in the course, and web-based course components or course components that are easy to use offline [18].

Other studies specifically focused research on majors, groups of students, or interventions. One large study conducted by The Learning House, Inc. and Aslanian Market Research in 2018 surveyed 1500 online students to determine what students want once they are enrolled in online courses. In 2018, students' first concern was for mobile-friendly course materials. Their second priority was for asynchronous, interactive course materials (videos and PowerPoints from the instructor, textbooks, and written assignments). Students were not enthusiastic about synchronous sessions and third-party videos [17]. But, students were very attracted to "textbook free" courses or courses that use OERs (not courses without course materials) [17]. Finally, while a big advantage of online programs is that they are available to anyone, anywhere, surprisingly, students want their online programs close by. In fact, 78% of students surveyed lived within 100 miles of their campus, with 44% living under 25 miles away. Students may study online, but they prefer the institution to be close [17].

It does seem that one thing that has changed since the 2010 Penn State survey is that now students are more desirous of mobile materials rather than downloadable ones; however, students still value knowledgeable instructors who have a clear presence in their courses.

Muljana and Luo produced a systematic literature review of 40 studies published between 2010 and 2018 related to "the underlying factors that influence the gap between the popularity of online learning and its completion rate" [19]. Student success strategies identified in the study were grouped under common headings referencing early intervention, engagement, course design, and synergy of stakeholders. While the reviewers found significant discussion of "[p]rofessional development, training, and workshops to inform faculty practices associated with online learning theories, student engagement, students' needs, dynamic dialogue, high quality feedback, appropriate delivery methods and technology," they also noticed a lack of discussion on efficacy of faculty support, "such as professional development opportunities such as a summer institute, training, and workshop" [19]. Critically, the reviewers noted "while student characteristics are among determinants of student retention in online learning, the results of this study do not include a detailed discussion on suitable instructional strategies for fostering behaviors associated with academic success" [19].

D. The Criticality of Faculty Development

All of these problems, goals, strategies, interventions, etc. require faculty development if they are to be addressed or implemented effectively. Faculty success in the in-person, online, and hybrid classroom is necessary, although it is not sufficient, to achieve student success. Brinkley conducted research focusing on how participation in a professional development program impacted faculty teaching effectiveness in the online environment and attitudes toward the effectiveness of the training [20]. Brinkley concluded "that instructors demonstrated (a) statistically significant changes in the incorporation of elements into the redesign of their syllabi, and (b) improvements in their teaching abilities, [but] there were no statistically significant differences in student evaluation scores of teaching pre- and post-training. Overall, the findings to the first research question revealed only modest improvements to the instructors' teaching effectiveness." As to the faculty attitude toward the training, "prior to the training, instructors were highly optimistic about their course redesign plans and the skills and knowledge they would develop in the training" and were "generally satisfied with the program" after the training. "However, after delivering their newly redesigned course online, participants were less optimistic and satisfied with their training experience than they had been prior to and following it, and multiple

instructors cited a need for additional or continued training and support" [20]. Brinkley notes it is important we use multiple data sources, spanning greater periods of time to gain a better understanding of the impact of professional development.

Daly analyzes grant-funded faculty learning communities at seven different institutions of higher education after determining a need for scholarship that analyzed why faculty learning communities are generally successful [21]. The research is grounded in social cognitive theory, which was chosen to "examine the relationship between faculty needs and the conditions for learning that are provided by the colleges and universities in which they work" [21]. Each learning community met weekly to "engage in professional reflection and initiate changes" over one semester, then met weekly over a second semester to implement projects that would address campus-wide diversity needs [21]. The learning communities were considered a success by the researchers, and the exit interviews of the participants aligned with "Deci and Ryan's (2000) self-determination theory, which focuses on the needs of individuals for autonomy, competence, and relatedness," and boosted self-confidence [21]. Ultimately, Daly indicated that topic-based learning communities "promote[d] specific types of pedagogical change" [21].

After analyzing 47 published studies on best practices in online learning and studying the rapid advancements of technology in the past century, Sun and Chen determined "that most online faulty have not received adequate training and support from their institutions" [22]. They define adequate training to include the following topics: "how to promote effective online collaboration for students, how to set high expectations, how to adjust instructors' teaching to conform to the online environment, and how to create proper online teaching strategies,… [along with] adequate training in the technologies applicable to online teaching" [22]. Sun and Chen, like Daly, stressed the success of a learning community approach for both student learning and faculty development [22].

The Canadian Digital Learning Research Association (CLDRA) found that the most reported challenges in provision of digital education professional development were "culture change, work security, and unclear expectations" [23]. That is, there may be underlying factors beyond lack of training or lack of understanding of technology that hamper professional development efforts at the institutional level. In addition to analyzing faculty development programs themselves, it is important to realize that when we discuss faculty development, we may not mean "all faculty" at a particular institution. Brady studied the impact of professional development for adjuncts on student success, noting "adjunct instructors have not always been afforded the same training and development opportunities" as full-time faculty. [24]. It is also important to note that in the current teaching climate, with the popularity of online courses, faculty may now be asked to teach courses designed by others—colleagues, subject matter experts, and/or instructional designers. Implementing student success strategies into a course that one did not initially create can pose its own challenges.

Another strategy regarding faculty professional development is rooted in asking the faculty member to share the student's experience. Utah Valley University created professional development for the teaching of its online English language program courses, based on parallel design, to mirror the educational experience of online students. This design incorporated three theories for developing learner autonomy – transactional distance, self-regulated learning, and collaborative control. Through the professional development, faculty learned the importance of decreasing transaction distance through the establishment of effective and timely communication and positive student-teacher relationships. Faculty learned to approach the course through phases of self-regulated learning - including forethought, performance, and self-reflection – to understand the factors affecting learning online. Lastly, faculty practiced encouragement of community building and learner autonomy through giving students more collaborative control of discussion boards. A key component in the training was helping instructors "recognize how they can incorporate their own voice through response to learners in order to make a course that may have been authored by someone else their own." [25]. Results highlight the importance of "implementing the elements of goal-setting, learning and applying new teaching strategies or adapting known strategies, and reflection on the effectiveness of these strategies parallels effective student learning processes based on the theory of self-regulated learning," and that "online learning is not an isolated activity. Socialization, support, team-building, and problem-solving can be developed through well-designed online course activities. These can result in ownership of learning, self-direction, and autonomy" [25].

In Southern Oregon University's 2015-2016 Faculty Writing Fellows Seminar, eight instructors of first-year foundational courses across diverse disciplines learned about methods and implementation of various pedagogical techniques to strengthen their students' writing abilities [26]. The seminar, structured similarly to a professional learning community, aimed to remedy the lack of faculty expertise in teaching writing skills to students by assigning faculty with readings and discussions, encouraging them to position themselves as learners. Researchers then compared fifty student compositions written in five participating instructors' subsequent courses with those of students in courses taught by five non-participating instructors, noting that the former substantially outscored the latter. In addition, researchers surveyed participating instructors and observed an increase in "confidence as a writing instructor, … empathy for students, … knowledge about writing instruction, and … instructional practices that support students' success" [26]. As the research shows, there is a thick web of complication surrounding the

relationships among student learning, student success, and faculty development,

A cornerstone to the effort to support student success is faculty development. A common refrain is "[s]tudent success is faculty success." The work to increase student retention in college, progression through an academic program, and graduation from an academic program must include both effort toward pieces that support student success and pieces that provide faculty the support they need to support student success. This axiom is true in online courses as well as face to face and hybrid courses. The Student Success Minutes training was created to help gently nudge faculty in the direction of the research and take into account faculty conceptions and needs as well as research-based student success findings.

IV. Faculty Survey

Our Build a Web Course training program is peppered with student success research from well-known experts like Saundra McGuire [10], Jessamyn Neuhaus [27], Flower Darby [12], Anya Kamenetz [13], and Richard Arum and Josipa Roksa [15], which we couple with advice and examples of successful strategies employed by our own faculty. However, research conducted by Karen Brinkley on the effectiveness of faculty development training [20] caused us to consider the effectiveness of our own training. Brinkley found that "prior to the [faculty development] training, instructors were highly optimistic about their course redesign plans and the skills and knowledge they would develop in the training" and were "generally satisfied with the program" after the training. "However, after delivering their newly redesigned course online, participants were less optimistic and satisfied with their training experience than they had been prior to and following it, and multiple instructors cited a need for additional or continued training and support" [20]. With this need for more training and support in mind, we conducted an informal survey in the spring 2021 of former Build a Web Course workshop participants and found that none of them remembered those aspects of the workshop that addressed student success in online courses. It seems that the focus of the faculty had been on the technology and general design of their courses rather than the details related to student success.

Next, we combined our informal survey findings with those of a formal survey of 177 Kennesaw State University faculty conducted at the end of 2019. A team of college-level online coordinators at KSU surveyed faculty regarding aspects of online teaching valued by higher education faculty in an effort to ascertain what students found to be the valuable part of an online course vs. what faculty found to be the valuable part of an online course. The purpose of the survey was to determine the similarities between what the research said students valued and what faculty valued. We then used that information to shape faculty development in a way that better responded to faculty assumptions and current faculty practices and preferences.

KSU is made up of 11 different colleges, but over 80% of the survey respondents came from only three colleges: Coles College of Business (13%), Bagwell College of Education (16%), and RCHSS (55%). The online teaching experience of the survey's participants varied, with 22.35% having no online teaching experience in the past two years, and 22.35% having taught nine or more online sections in the past two years. The majority of respondents, 53.53%, had taught no blended or hybrid sections in the past two years, with the next largest group, 28.82%, having taught 1-3 blended/hybrid sections in the past two years. Thirty percent of respondents had taught 1-3 years online, and 58.82% had developed 1-3 online/hybrid/blended courses to teach.

We asked faculty in the survey to identify the five items that they believe are most valuable in their online courses with regard to making a class better for them as the instructors (i.e., easier to manage, easier to teach) and for the students (i.e., learning effectiveness).

The top five items that faculty felt made the course easier to teach and manage were

1. Peer reviews of the online course by colleagues or instructional designers (90.32%)
2. A course quality rubric such as Quality Matters (86.67%)
3. Publisher course packs (76.74%)
4. Proctoring tools such as Respondus and/or proctoring services such as the KSU Testing Center and ProctorU (76.47%)
5. Tools such as SoftChalk and Kaltura/MediaSpace (71.76%)

The top five items faulty felt were important for student learning were

1. Clear guidance on how to access resources in a course (78.95%)
2. Clear "start here" information (75.86%)
3. Clear grading information (71.43%)
4. Quick response time to emails and grading (67.21%)
5. Mobile friendly (65.57%)

As one can see (Fig. 1), there is little overlap between the two groups from the faculty perspective, which means faculty feel they must choose, as they design their courses, whether to focus on things that make the course easier to teach (such as publisher packs) versus things that they believe are important for student learning (such as clear guidance on how to access resources in a course).

In addition, there is little overlap between what the research says students find valuable and what faculty believe students value (Fig. 2). As discussed above, students emphasize responsive instructors, functional e-learning platforms, use of social media to create a sense of belonging appropriate assessments, clear guidance on how to use resources in the course, instructor interaction, mobile friendly courses, asynchronous and interactive course materials, and open educational resources.

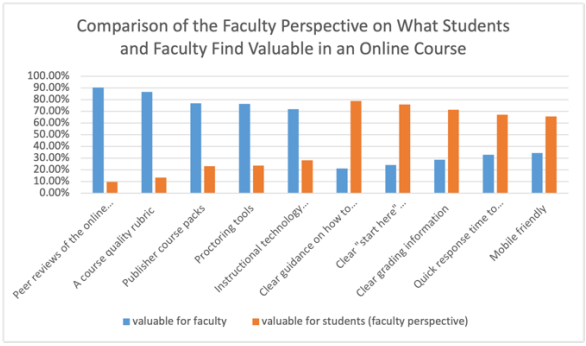


Figure 1. A comparison of the faculty perspective on what students and faculty find valuable in an online course. This information is from the 2019 survey of KSU faculty.

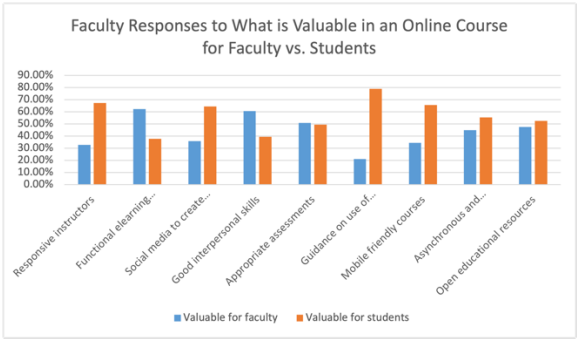


Figure 2. A comparison of faculty responses to what is valuable in an online course from 2019 survey of KSU faculty vs. Student responses to what is valuable in an online course [16] [17] [18] [19].

Of course, these differences in preferences do not mean that techniques that make teaching easier and those that support student success are mutually exclusive. For example, faculty do not appear to have made the connection that a course quality rubric such as Quality Matters supports faculty providing clear grading information in a course. It is worth noting that while faculty did not generally perceive items such as responsive instructors and guidance on how to access resources in the class as things that would be valuable to the instructor, clear guidance on how to use resources would cut down on email to instructors and explanations from instructors regarding how to use such resources. Additionally, while neither research on what students wanted nor the survey on what faculty value rated a clear course schedule very highly, such a tool would also both ease the burden on faculty regarding student emails and scheduling and support student success. Clearly, this discrepancy between research and perception merits further exploration.

Fig. 3 includes the full set of responses to the question "Identify the five items that you believe are most valuable in your online courses with regard to making a class better for you as the instructor (i.e., easier to manage, easier to teach) and for the student (i.e., learning effectiveness)" on the original survey.

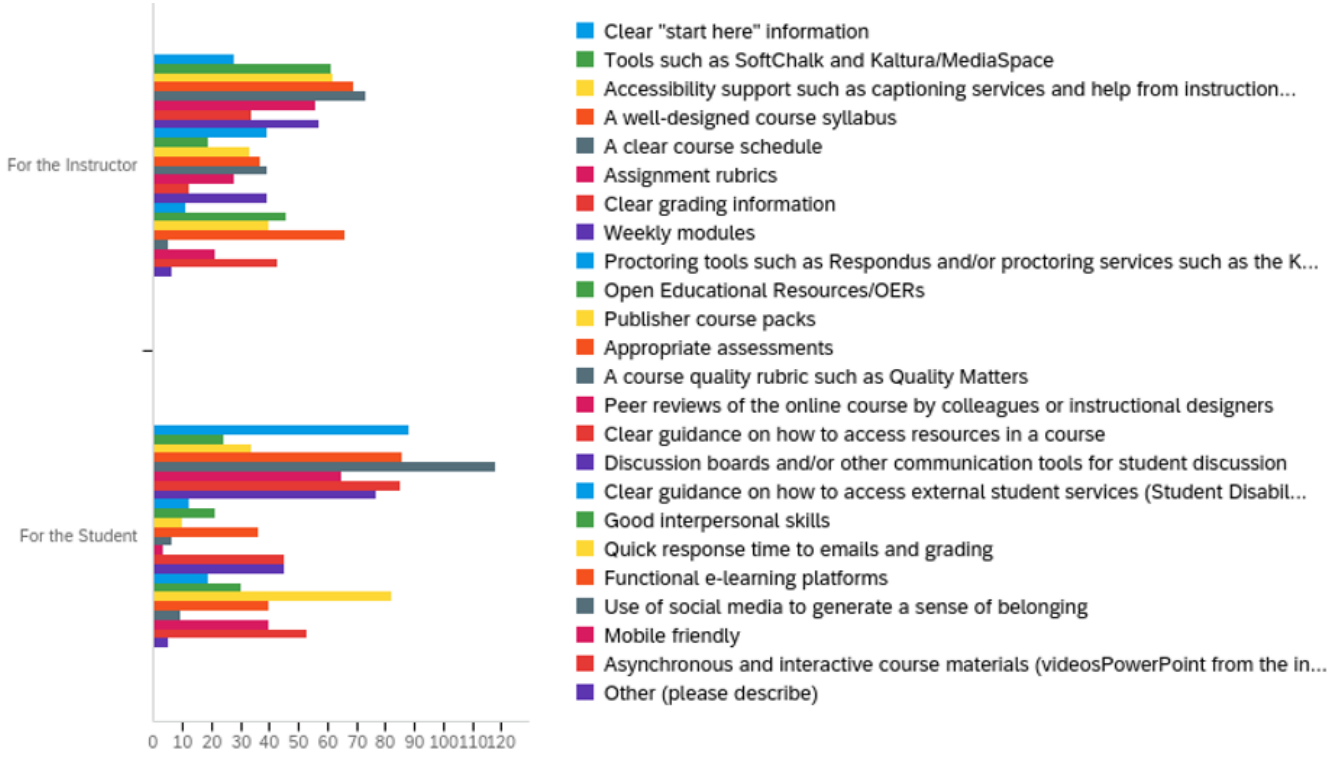


Figure 3. Overall results of the survey question asking faculty to identify five items that they believe are most valuable for them in their online courses (with regard to making a class better for the instructor) and five items that are most valuable to the student.

The survey also showed that students are the reason faculty take steps to improve their courses. One question in the study asked faculty, "What motivates you to make improvements to your online courses?" As shown in Table 1, the responses overwhelmingly stated students were the key motivator for course improvements along with learning and teaching. Feedback through student evaluations is also a motivator for improvement with course materials. The desire to improve the experience for students within the online environment can result in further professional development. The survey demonstrates there is no lack of desire on the part of faculty to improve online courses for students. Clearly, then, the fact that faculty assumptions regarding what students value in an online course do not equate to what students actually value is most likely the result of a lack of information. Furthermore, faculty were asked within the survey, "Is there anything preventing you from participating in distance education efforts (teaching, training, etc.) as a faculty member?" With the other obligations faculty are required to engage in on an annual basis, time is the number one reason preventing further training/education and/or teaching in the online environment. Faculty are concerned about the time it takes to train, create, and teach within the online environment, which causes them to discount the opportunity (Table 2). Additionally, faculty are concerned about the teaching workload and believe it takes a substantial amount of time to grade and offer continuous updating of materials etc. Faculty concern about time is reflected in Table II.

We concluded from this survey that, while their motivations for participating in faculty development and updating their online courses were inspiring, faculty ideas about what supported student learning were sometimes at odds with the research, and their ideas about what faculty need to create effective online courses were sometimes at odds with what students needed (for example, OERS) or not even on student radar (mobile-friendly—although perhaps students take this for granted). Instructors want to help students succeed, but there is an incongruous response pattern in the survey. What is important to the student and what is important to the instructor are not always in alignment. It seems clear, then, that providing faculty with information regarding what students value and what strategies support student success would help faculty achieve their goals. Additionally, training and opportunities to further faculty development and support faculty in planning online courses and reflecting on course design and delivery can be advantageous for the student and the instructor.

TABLE I. WORD FREQUENCY IN THE RESPONSES TO THE 2019 KSU FACULTY SURVEY QUESTION "WHAT MOTIVATES YOU TO MAKE IMPROVEMENTS IN YOUR ONLINE COURSE?"

Word	Frequency
Student(s)	188
Learn(ing)	77
Teaching (material updates)	33
Feedback (from students)	27
Experience (better for students)	21
Desire (to improve)	15
Professional (development)	15

TABLE II. WORD FREQUENCY IN THE RESPONSES TO THE 2019 KSU FACULTY SURVEY QUESTION "IS THERE ANYTHING PREVENTING YOU FROM PARTICIPATING IN DISTANCE EDUCATION EFFORTS (TEACHING, TRAINING, ETC.) AS A FACULTY MEMBER?"

Word	Frequency
Time (teaching, training, creating)	71
Online/distance (challenge)	37
Training/education (effort)	29
Teach(ing)	27

V. RESEARCH PROJECT

We realized that we needed a more focused strategy to supply faculty with information regarding implementing strategies for student success in their own online courses. We initially sought to emphasize student success information within the existing training without adding significant time and work for the faculty participants. We then realized that by embedding the information within a larger faculty development workshop, we were creating barriers to faculty adoption of the techniques and limiting the audience. That realization led us to develop phase two of the project.

We already included a wealth of student success strategies in the workshop. However, the information was provided along with information on research-based best practices in course design and technology tutorials for creating course materials. Student success strategy information was not prioritized or emphasized for faculty participants.

Especially for faculty new to online teaching, we could see how workshop participants would prioritize "how do I create the class," "how do I make it accessible to students who use screen readers or who need captioning," and "what software do I use to create course materials" over "how do I strategize for student success." The faculty participants had finite time and energy to complete the training and create the course. But could we also call attention to student success strategies in hopes of encouraging faculty participants to add a few of those to their courses, as well?

A. Research Project

As mentioned earlier, the chief impetus of this research was to prepare the college for successful request of student success funding. Beyond that, we wanted to be able to demonstrate that we had identified a way to increase and support student success. And of course, the heart of our motivation was to assist our students in achieving their academic goals.

The researchers designed a two-phase research project. Phase 1 (completed) involved creating stand-alone student success content, sharing it with faculty within an existing professional development workshop, and following up with a survey to measure their intent to adopt student success strategies into their courses. In phase two (ongoing), we extracted the student success modules from the faculty development workshop and created a standalone faculty development training available as an open educational resource. This training is available online and on demand to anyone who wishes to access it through Affordable Learning Georgia's *Open Educational Resources: ALG Repository*. Throughout this phase of the research, we are asking participants to complete a survey to measure their intent to adopt student success strategies into their courses.

B. Phase I: Lower Barriers to Adoption

To begin phase one, the researchers did three things: 1) isolated the research-based, student-success content from the general content of the faculty training modules and emphasized it in highlighted segments of the training called Student Success Minutes; 2) added an activity to each of the Student Success Minutes to support the faculty in remembering the content; 3) surveyed faculty at the end of the training to see if they recall and plan to use the Student Success Minutes information (intent to transfer) [28].

The researchers designed each Student Success Minutes segment to be less than 10 minutes, including the activity, so as not to overburden the faculty with more training content. In this initial, pilot phase of the project, our goals were to create the segments and present them to the faculty participating in the spring 2021 "Build a Web Course Workshop" and then survey faculty participants, as described above, regarding intent to transfer. We started with a small number of faculty participants (8). Because of low faculty enrollments, in this first phase of the project we were able to gather little more than a handful of initial reactions.

C. Phase II: Lower Barriers to Access

Phase two shifted the content to a second, shorter, asynchronous training using the Student Success Minutes segments. This training initially targeted faculty who had previously completed the "Build a Web Course" Workshop but did not receive the redesigned content on research-based strategies for student success.

The redesigned six-module, asynchronous, self-paced training takes participants less than two hours to complete. Other workshops focusing on student success strategies at the institution take more time and/or lack the flexibility and interaction of our Student Success Minutes training, which is hosted on the internet and freely available. The redesigned training and the accompanying survey of intent to transfer will be available indefinitely as we continue to refine the training content and collect data from users. The success of this project will be measured in the survey results and findings on intent to transfer techniques discussed in the training.

D. Next Steps

While measuring DFWI rates at our institution might also be helpful in determining impact of a singular initiative, the truth is that we have so many student success efforts ongoing that it would be impossible to tell which one or ones had what impact. In addition, the researchers are cognizant that students drop courses for many reasons that may have nothing to do with the professor or the course content. Use of DFWI rates is also sensitive due to the potential for professors to feel targeted by attention to such information. For this reason, we chose not to measure individual DFWI rates in this research. At the end of the project, we will gather aggregate data on DFWI rates as a measure of overall student success trends within the College.

After the two phases of the project, the researchers plan to use the information gathered to assess whether highlighting student success strategies in faculty development training can encourage faculty to implement these strategies. If we find we have a successful strategy, we will be able to use this information to better position our college to receive student success funding when future opportunities arise.

VI. Research-Based Modules on Student Success

In the first phase of this project, the research team created six Student Success Minutes segments. This section will describe each segment, provide the research it is based on, and describe the activity provided with it and faculty participant results, if available.

A. Student Success Minutes 1: Scaffolding

This Student Success Minutes segment was based on the work of Flower Darby (Fig. 4). Darby explains scaffolding through her experience teaching jazz dance. She writes,

> [B]eginning dancers get frustrated and demotivated if I constantly throw new things at them. Better to practice one new step for a while, get feedback from me on their progress, and build confidence and self-efficacy before introducing a slightly more complex step or one that requires greater skill. [12]

Darby extrapolates this idea to other academic realms. While scaffolding in college classes is not a brand-new idea, Darby provides an excellent explanation and rationale for the practice. For example, in a research paper assignment, instead of assigning a 10-page research essay, ask students to turn in a topic early in the course; a few weeks later, ask students to turn in an annotated bibliography with a tentative thesis; and two weeks before the paper is due, ask students to turn in (or share on a discussion board) a PowerPoint with the title and thesis on slide 1 and the topic sentence and paragraph supporting points for each paragraph in the paper.

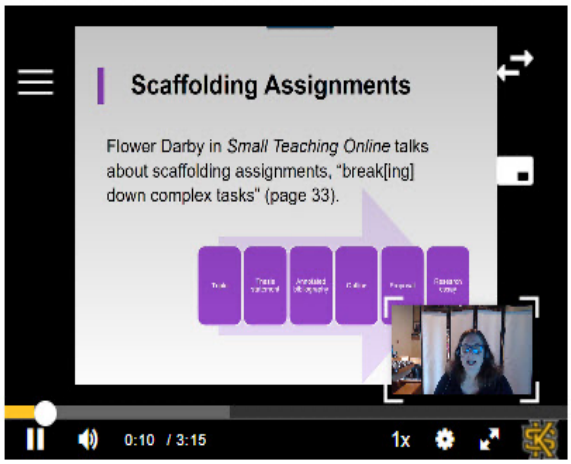


Figure 4. Student Success Minutes 1:a video explaining Flower Darby's approach to scaffolding.

Figure 5. Student Success Minutes 2: a short, self-paced, interactive presentation on the social media tool GroupMe.

Of course, the faculty member would be expected to provide timely and helpful feedback on each phase before the next phase is due. To introduce (or remind) faculty of this student success strategy, in a three-minute video, Tamara Powell, Director of the RCHSS ODE, explained the concept of scaffolding and asked participants to share a reflection on when they might use the strategy to support student success in a class. In the reflection assignment, 100% of faculty participants in phase one indicated that they already used scaffolding strategies in their courses to some degree.

B. Student Success Minutes 2: GroupMe

The second Student Success Minutes segment was based on a need within the institution. At Kennesaw State University, student culture results in the creation of a GroupMe (Fig. 5) for each class in which students are enrolled—bypassing the professor. This student-created classroom community harkens back to Kamenetz' prediction that faculty and institutions will need to change and adapt to the disruptions that technology brings. Kamenetz told us there would be "rough spots" [13], and the "do it yourself" approach to community building in online courses that students have taken with GroupMe certainly can be one of those rough spots. But faculty can work to lessen the negative impact and increase the positive through knowledge and deliberate action.

GroupMe is a social media application that allows a group to chat via mobile app or website without exchanging personal information [29]. On the one hand, GroupMe is excellent for creating community and support in an online course. On the other hand, some students with the best intentions have been tempted to use GroupMe to commit breaches of academic integrity.

In response to these problems, Sam Lee, a student at Kennesaw State University as well as a teaching assistant in the Spanish and French programs and an assistant instructional designer in the RCHSS ODE, created an interactive presentation using Articulate Storyline 360.

The presentation walked faculty participants through an overview of GroupMe and provided suggestions to faculty regarding how to minimize student cheating with it and how to use it with students to support student success.

This presentation concluded with a short quiz to support comprehension of the main ideas. Faculty participants were allowed to attempt the quiz multiple times, and all faculty participants in phase one scored 100% on their final attempts.

C. Student Success Minutes 3: Open Educational Resources and Creative Commons

In the past five years, a great deal of research has been done on the impact of OERs—or no-cost or low-cost course materials—upon student success efforts. In the United States, textbook prices have risen astronomically. In the last 10 years, the "average cost of college textbooks has risen four times faster than the rate of inflation," and "65 percent of students . . . skip buying required texts" to save money or simply because they cannot afford them [30].

As an alternative to expensive textbooks, many faculty members turn to OERs. Research into OERs has shown that OERs increase student participation, satisfaction, learning, retention, and course and program completion. They reduce student debt not only by lowering textbook costs in individual classes but also by allowing students to take more courses in a term, thereby graduating more quickly and accruing less student loan debt [31]. Kamentez predicted the rise of OERs in her 2010 work *DIY U: Edupunks, Edupreneurs, and the Coming Transformation of Higher Education* [13]. Kamenetz did not predict the power with which for-profit publishing houses would attempt to subvert OERs and even try to monetize them. It can be difficult, now, for some faculty to envision teaching without high-priced publisher supplements to their instructional materials. But for many students, OERs are one of the most important factors a faculty member can implement into a course. And, OERs become a social justice issue as textbook prices climb.

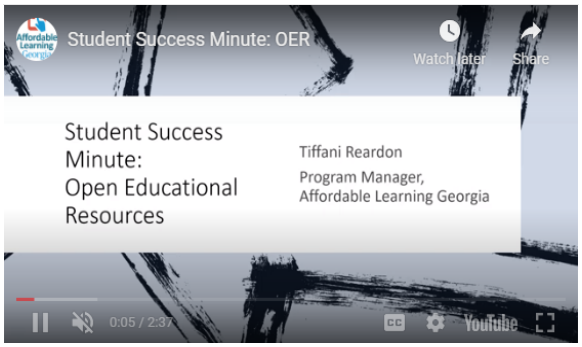


Figure 6. Student Success Minutes 3: a short video and quiz on Open Educational Resources.

Tiffani (Reardon) Tijerina (Fig. 6), the Program Director for the Affordable Learning Georgia initiative, created a Student Success Minutes segment on OERs for this project. In the two minute and 37 second video, Tijerina defines open educational resources and explains their benefits as well as Creative Commons licensing. The Creative Commons licensing explanation is provided to support understanding of the types of resources that can be used as OERs in classes.

Tijerina's Student Success Minutes segment concludes with an ungraded self-assessment on the terms and concepts presented in the segment and then a graded quiz on the same terms and concepts. Faculty participants were invited to practice with the ungraded self-assessment as much as desired before taking the graded quiz on the same information. Every faculty participant in phase one scored 100% on the graded quiz. Ungraded self-assessments [32] will be the topic of a future Student Success Minutes segment.

D. Student Success Minutes 4: The Quick Write

Saundra McGuire recommends a reflection activity as part of a class to engage students and enhance self-esteem [10]. It is hard to imagine that something so simple to implement can be such a powerful tool for student success. Stephen Bartlett, Associate Director of the ODE, created a short video on a type of reflection assignment called "The Quick Write" (Fig. 7).

As Bartlett explains, McGuire uses the Quick Write as a confidence booster. She asks students to remember a thing they learned that was hard and recall how they learned it [10]. Bartlett also recommends using the Quick Write as a reflection assignment to help cement information students have learned in a class period and to "check in" on students regarding to their progress in the class.

Asking students to take just one to three minutes to write about an aspect of the material that was just presented is a great way to support learning and engagement, and it also allows the professor to see whether students are paying attention or "getting the material" in an online class.

Figure 7. Student Success Minutes 4: a short video on the power of reflection.

E. Student Success Minutes 5: Weekly Modules

Universal Design for Learning Theory states that consistency is a key component for supporting increased success as it lightens cognitive load, freeing up more time and mental energy to assist the student in learning the course content [33].

It is important to be consistent in scheduling expectations for students in online courses. Students are used to organizing their college schedules by weeks in in-person classes, and it makes sense to use that structure in online courses as well. It also makes sense to create folders, organized by weeks, with everything a student needs in that folder to complete that week of class.

When faculty instead create modules of random lengths (module 1 is three weeks, module 2 is four days, module 3 is seven weeks, etc.), students who already struggle with time management can suffer severely. When faculty create overly long modules (one 16-week course with only four, four-week-long modules), students who wait until the last minute find out four weeks into the course that they have fallen too far behind to succeed.

Student Success Minutes 5: Weekly Modules (Fig. 8) provides the rationale for organizing the online course in a weekly fashion and examples of why it is the easiest way to support student success in an online course.

Weekly organization of online classes supports student success by providing consistency, reducing cognitive load, and helping students to organize their time [34]. This segment, created by Tamara Powell, ended with a quiz over the material presented in the short video. All faculty participants in phase one scored 100% on the quiz.

Figure 8. Student Success Minutes 5: a short video on weekly modules.

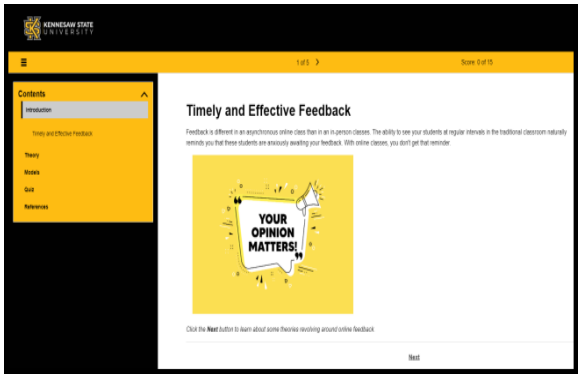


Figure 9. Student Success Minutes 6: a website with interactive exercises and a quiz that provide information about the importance of timely and effective feedback.

F. Student Success Minutes 6: Timely and Effective Feedback

A great deal of research on student success supports not only feedback, but timely and effective feedback [12], [27] [35]. For our last Student Success Minutes segment in the pilot, Sam Lee created a website that included an interactive presentation on the importance of timely and effective feedback.

As Darby points out, "It's easy for online students to feel isolated and unsupported" [12]. Feedback, even small notes about low or no stakes assignments, can motivate students to invest more time in the course. Such feedback can also alert students that they are not doing enough to succeed in the class—or are on the wrong track—long before they fail a high stakes assignment. In this way, timely and effective feedback promotes student success.

As the reader may remember, this project was inspired partly as a way to provide student success strategies to faculty who were already strapped for time. And, as we know very well, suggesting faculty take time to provide more feedback is not a timesaver. However, in the age of technology, faculty can often use the learning management system to "work smarter, not harder." Specifically, many learning management systems have automated feedback tools to allow faculty to set up bots to, for example, send out a congratulatory email to students who did well on a test or send study tips to students who did not do so well on a test.

Solutions that support student success and reduce faculty workload are not always possible, but in this case, the student success strategy was able to support both positive outcomes

This module included the interactive presentation, mentioned above, along with a practice quiz that allowed participants to check their understanding of the material. After the practice quiz, participants in phase one took a graded quiz with the same questions. The quiz was worth 20 points, total, and the average grade was 75%. This information suggests that the presentation on timely and effective feedback needs adjustment to increase participant retention of the information.

VII. OER Student Success Resource

In the literature review, it was noted that faculty often did not have access to professional development [22] or professional development was limited to full time faculty [24]. Additionally, in the survey of KSU faculty, time was the number one reason preventing further training/education and/or teaching in the online environment. To address these concerns, phase two of the project was created. As referenced previously, phase two of the project included offering the modules described above in a standalone format in SoftChalk and hosted on the internet. This training is called "Student Success Workshop" [36]. These modules were available online to anyone, anywhere, and users were asked to complete a survey at the end. The survey measured, among other things, intent to transfer. The SoftChalk ScoreCenter showed that 51 people accessed the open, online training. However, only five persons completed the cumulative assessment and received a certificate by December 2021. Clearly, at this point, we need to examine why the training is not engaging users and moving them to completion of the certificate requirements.

A. Survey Results

Of the 51 people who accessed the training, only eight completed the survey linked to phase two of the training. All of this information was reported anonymously. Of those eight, four indicated they were faculty at Kennesaw State University, and two indicated that they were not. Two did not respond to that question. Six respondents indicated that they remembered all six topics presented to them in the training, and two did not respond. This result indicated that moving the training into a standalone format seemed to make it easier for participants to recall the topics.

The segment that respondents indicated they found most helpful was the "Timely and Effective Feedback" segment created by Sam Lee. All six participants who were still responding to the survey found that segment helpful. Four found the "GroupMe" segment helpful—also by Lee.

Participants were queried regarding their intent to transfer the information presented in training using this survey question: "The following are the titles of the six strategies. Select any you plan to incorporate into your own course(s). In addition, please share your ideas regarding implementing these strategies in your course(s)." Of the four participants who responded to this question, at least one person found each strategy helpful and intended to incorporate it in a course. While "Timely and Effective Feedback" and "GroupMe" were deemed the strategies participants would use the most, "OER and Creative Commons" had only one user indicate that that information would be used in future courses.

Participants were also asked to share their ideas regarding implementing these strategies in their courses. With regard to "GroupMe," participants shared that they would start to make their own GroupMe in line with the workshop suggestion to try to deter student cheating with the social media tool. One participant wrote, "I have been largely ignoring GroupMe until this semester and one class. In the future, I will try using the create my own GroupMe strategy." Responses to the "Timely and Effective Feedback" segment indicated that participants engaged with the material. One participant wrote, "I like the rubric method, and it does reduce grading time. Maybe I should create lower stakes activities to provide more beneficial feedback." Another shared, "I will specifically separate out a 'for next time' in my feedback." And a third response was
very enthusiastic: "I'm re-working office hours to gain more attendance—love the Doodle survey idea!"

As might logically follow, the "OER and Creative Commons" had very few responses. One respondent wrote, "I was not aware of the various types, and I would like to become more familiar to prevent misusing OER material." This response might indicate that the segment was too packed with information for users to easily digest. Alternately, inclusion of OER may be seen as a larger, more systematic change than the incremental adjustments needed to incorporate the other techniques presented.

"Scaffolding," "The Quick Write," and "Weekly Modules" were also moderately popular. The written responses to the survey that addressed scaffolding and weekly modules indicated that these were familiar strategies that were already widely used in online classes. However, the "Quick Write" or reflection segment seemed to also spur participants to consider this small change in a course. One participant responded, "I have heard of this method, but I have not tried it yet. I plan to use it in my upper-level courses." Another respondent shared, "Small reading sections could really benefit from quick writes."

An additional survey question asked, "If you are not going to use any of these strategies, please share your reasons." One answer referenced content, that the OER material wasn't relevant to that instructor's courses because the instructor taught ancient languages and the material was already out of copyright. Another shared that the strategies were already familiar. A third shared that some were already in use in that instructor's course, concluding that "If this were new to me, I'd be plugging in everything you taught and be excited about it!"

When asked if we should continue the workshop in this format, three participants answered, "Yes," and two answered, "Other." No one answered, "No." The two persons who marked, "Other" provided explanations. One wrote, "I think they should be part of a series covering each strategy more in depth and providing assistance in creating them." A second shared, "It would depend on that person's level of familiarity with the pedagogy," which we interpreted as the person thought we were asking if the workshop should be recommended to another person.

VIII. Conclusion and Future Work

The project was borne of several motivations: a foundation for securing student success funding, a desire to increase student success, and a desire to provide easily accessible and time-efficient faculty development. The execution of the project was facilitated by the fact that faculty surveyed showed that they were intrinsically motivated to engage in professional development to improve their courses. The survey from KSU faculty also showed that faculty needed guidance regarding what students want in online courses. In addition, faculty did not always recognize what key elements of an online course support student success. The team's work resulted in an openly available faculty development resource that attempts to support faculty development needs while respecting faculty workloads. The team used information gleaned from both primary and secondary research to deliver a product that served a wide range of needs.

In phase one of the project, the summaries of each Student Success Minutes segment showed that faculty participants did engage with the materials—although they were least successful with the assessment included in segment 6. Four faculty members completed the survey regarding intent to transfer. (The survey is anonymous.) The faculty members remembered all of the Student Success Minutes and liked segments 2, 5, and 6 (GroupMe, weekly modules, and feedback) the best. All faculty members indicated that they intended to implement at least one of the strategies in the course they were building. When asked if these three segments should be included in future trainings, three of the respondents answered "yes." The fourth shared that it depended. Even respondents with previous training made comments such as "This was good--well put-together. Thanks! It added a few small changes that I think will have big effects to my class, so it was worth the time."

In phase two, the segments were isolated and offered as a freely available, online, asynchronous faculty training on student success, and participants were surveyed in those trainings as well. While we would have liked for the 51 persons who accessed the training to have completed the training and the survey, user feedback to date has been constructive and positive. For instance, when asked, "Is there anything else you would like to share regarding the 'Strategies for Student Success' workshop?" one participant responded, "I think the short sessions are fantastic. It would be nice to have these minute sessions posted on the faculty information

website," in reference to the training segments. The team appreciated that feedback and promptly did so. Another respondent wrote, "I really liked multiple styles of feedback to really reach different learning styles." With the limited feedback that we were able to obtain, we did see that those who participated in the survey intended to transfer the information to their teaching practice. While that is heartening, it is also clear that there is a need for extended research into application of these student success initiatives across time.

In the future, we will make the survey more prominent and include an explanation of how we will use the information in hopes of enticing more people to participate. Our goal is to have a set of strategies all faculty can easily incorporate into their courses to support student success and to gather data showing a reduction in DFWI rates. Preliminary results indicate that faculty who engage with the Student Success Minutes find them helpful and will be implementing them in their courses. However, we will continue to work to collect more data to create a stronger argument with regard to the effectiveness of this project across time. We will also incorporate user suggestions into our revision strategies for future offerings. At the end of the next several semesters, we will collect DFWI information for the entire college to see if the needle moves in a positive direction with regard to student success. To reiterate our earlier statements, many student success efforts are ongoing at KSU, and a moving needle would not indicate that this project alone made a difference. To determine whether our work is making a difference directly, we will need to solicit volunteers to allow us to survey students in classes where these specific strategies are being applied. With that information, we will have a clearer picture whether to continue in this direction with follow-up success strategies or pursue another path.

REFERENCES

[1] T. Powell, J. Newell, S. Bartlett, S. Lee, B. Milam, L. Snider, "Student Success Innovations vs. Faculty Workload Concerns: How to Find a Balance for Success," eLmL 2021: The Thirteenth International Conference on Mobile, Hybrid, and On-line Learning. Conference Proceedings, Nice, France, pp. 15-20, July 18-22, 2021. Available from: https://www.thinkmind.org/download_full.php?instance=eLmL+2021 2021.12.08

[2] Pew Trusts Foundation, "Two decades of change in federal and state higher education funding: Recent trends across levels of government," October 15, 2019. Available from: https://www.pewtrusts.org/en/research-and-analysis/issue-briefs/2019/10/two-decades-of-change-in-federal-and-state-higher-education-funding 2021.12.08

[3] S. Baum and M. Johnson, "Financing public higher education: the evolution of state funding," Research Report. The Urban Institute. pp. 1-24, 2015. Available from: https://www.urban.org/research/publication/financing-public-higher-education-evolution-state-funding/view/full_report 2021.12.08.

[4] M. Rothlisberger. Executive Director for Academic & Fiscal Operations at Kennesaw State University. Personal Interview, 2021.05.02.

[5] P. Freire. Pedagogy of the Oppressed. 30th anniversary edition. Translated by Myra Bergman Ramos. Contiuum. New York, 2005. Available from: https://envs.ucsc.edu/internships/internship-readings/freire-pedagogy-of-the-oppressed.pdf 2021.12.08

[6] G. D. Kuh, "High-Impact Educational Practices: What They Are, Who Has Access to Them, and Why They Matter." Qtd in "High Impact Educational Practices: A Brief Overview." Association of American Colleges and Universities. Available from: https://www.aacu.org/node/4084 2021.12.08

[7] J. Kinzie, A. C. McCormick, R. M. Gonyea, B. Dugan, and S. Silberstein. "Assessing Quality and Equity in High-Impact Practices." Center for Postsecondary Research." Indiana School of Education. Available from: https://nsse.indiana.edu/research/special-projects/hip-quality/index.html 2021.12.08

[8] J. Kinzie and G. Kuh. "Reframing student success in college: Advancing know-what and know-how," *Change*, vol. 49, no. 3, pp. 19–27, May 2017, doi:10.1080/00091383.2017.1321429.

[9] D. Stewart and Z. Nicolazzo, "High impact of [whiteness] on trans* students in postsecondary education," Equity & Excellence in Education," vol. 51, no.2, pp. 132-145, 2018, doi:10.1080/10665684.2018.1496046.

[10] S. McGuire. Teach Students How to Learn: Strategies You Can Incorporate Into Any Course to Improve Student Metacognition, Study Skills, and Motivation. Sterling, VA: Stylus, 2015.

[11] M. Winkelmes, "The National Teaching and Learning Forum." National Teaching and Learning Forum, vol. 24, no. 2, pp. 1-4, Feb. 2015, doi:10.1002/ntlf.20019.

[12] F. Darby. Small Teaching Online. San Francisco, CA: John Wiley and Sons, 2019.

[13] A. Kamenetz, DIY U: Edupunks, Edupreneurs, and the Coming Transformation of Higher Education. Hartford, VT: Chelsea Green Publishing, 2010.

[14] M. S. Hoyert and C. D. O'Dell, "Developing faculty communities of practice to expand the use of effective pedagogical tecniques," Journal of the Scholarship of Teaching and Learning, vol. 19, no. 1, pp. 80–85, Feb. 2019.

[15] R. Arum and J. Roksa, Academically Adrift: Limited Learning on College Campuses. Chicago, IL: University of Chicago Press, 2011.

[16] E. Toufaily, T. Zalan, and D. Lee. "What do learners value in online education? An emerging market perspective," E-Journal of Business Education and Scholarship of Teaching, vol. 12, no. 2, pp. 24-39, 2018.

[17] A. J. Magda and C. B. Aslanian, "Online College Students 2018: Comprehensive Data on Demands and Preferences," Louisville, KY: The Learning House Inc. Available from: https://distance-educator.com/online-college-students-2018-comprehensive-data-on-demands-and-preferences/ 2021.12.08

[18] P. Ralston-Berg, "What Makes a Quality Online Course? A Student Perspective," Presented at Distance Teaching and Learning 2009, Madison, Wisconsin. August 7, 2009. Available from: https://www.slideshare.net/plr15/what-makes-a-quality-online-course-the-student-perspective-18294402021.12.08

[19] P. S. Muljana and T. Luo, "Factors contributing to student retention in online learning and recommended strategies for improvement: A systematic literature review," Journal of Information Technology Education: Research, vol. 18, pp. 19-57, 2019, doi:10.28945/4182.

[20] K. E. Brinkley, "Learning to Teach Online: An Investigation of the Impacts of Faculty Development Training on Teaching Effectiveness and Attitudes toward Online Instruction," PhD diss., University of Tennessee, 2016. Available from: https://trace.tennessee.edu/utk_graddiss/4127 2021.12.08

[21] C. J. Daly, "Faculty learning communities: Addressing the professional development needs of faculty and the learning needs of students," Currents in Teaching & Learning vol. 4, no. 1, pp. 3-16, 2011.

[22] A. Sun and X. Chen, "Online education and its effective practice: A research review," Journal of Information Technology Education vol. 15, pp. 157–190, 2016, doi:10.28945/3502.

[23] C. A. VanLeeuwen, G. Veletsianos, O. Belikov, and N. Johnson, "Institutional Perspectives on Faculty Development for Digital Education in Canada," Canadian Journal of Learning & Technology, vol. 46, no. 2, pp. 1–19, Dec. 2020, doi:10.21432/cjlt27944.

[24] T. L. Brady, Implementing an Adjunct Training and Development Model to Increase Student Success A Mixed Method Study, PhD Diss., William Howard Taft University, April 2020.

[25] M. S. Andrade, "Teaching online: A theory-based approach to student success," Journal of Education and Training Studies, vol 3, no. 5, pp. 1-9, September 2015, doi:10.11114/jets.v3i5.904.

[26] M. Perrow, "Designing professional learning to support student success: Lessons from the Faculty Writing Fellows Seminar," College Teaching, vol. 66, no. 4, pp. 190–198, Oct. 2018.

[27] J. Neuhaus. Geeky Pedagogy. Morgantown, WV; West Virginia University Press, 2019.

[28] K. Hye-Sook and S. Yu. "Structural relationship among environment, motivation, engagement and transfer of training of teachers in distance education." KEDI Journal of Educational Policy, vol. 17, no. 2, pp. 221–245, Dec. 2020, doi:10.22804/kjep.2020.17.2.004.

[29] GroupMe, Homepage, 2021. Available from: https://groupme.com/ 2021.12.08

[30] K. Kristoff, "What's behind the soaring cost of college textbooks," Moneywatch, CBSNews, January 26, 2018. Available from: https://www.cbsnews.com/news/whats-behind-the-soaring-cost-of-college-textbooks/ 2021.12.0.08

[31] N. B. Colvard, C. E. Watson, and H. Park, "The impact of open educational resources on various student success metrics," International Journal of Teaching and Learning in Higher Education, vol. 30, no. 2, pp. 262-276, 2018. Available from: https://files.eric.ed.gov/fulltext/EJ1184998.pdf 2021.06.10

[32] J. Eustace and P. Pathak, "Retrieval practice, enhancing learning in electrical Science," Proceedings of the 11th International Conference on Computer Supported Education (CSEDU 2019), vol. 1, pp. 262-270, May 2019, doi:10.5220/0007674102620270. Available from: https://www.researchgate.net/publication/332858972_Retrieval_Practice_Enhancing_Learning_in_Electrical_Science 2021.12.08

[33] S. Johnson, "Design, consistency, access," Online Course Development Resources. Vanderbilt University. Available from: https://www.vanderbilt.edu/cdr/module1/design-consistency-and-access/ 2021.12.08

[34] S. Negash and T. Powell. "What if We Put Best Practices into Practice?: A Report on Course Design Beyond Quality Matters," Proceedings, Learner Conference (2013) University of the Aegean, Rhodes, Greece. Available from: https://radow.kennesaw.edu/ode/docs/BestPractices.Powell.Negash.2013.doc 2021.06.10.

[35] S. Laato, E. Lipponen, H. Salmento, H. Vilppu, and M. Murtonen. "Minimizing the Number of Dropouts in University Pedagogy Online Courses." CSEDU 2019 11th International Conference on Computer Supported Education. Heraklion, Crete, Greece. May 2-4, 2019. Proceedings, Volume 1. Edited H. Lane, Susan Zvacek, and James Uhomoibhi. pp. 587-596

[36] Student Success Workshop, Homepage, 2021. https://alg.manifoldapp.org/projects/student-success-workshop 2021.12.08

An Investigation of Japanese Twitter Users Who Disclosed Their Personal Profile Items in Their Tweets Honestly

Yasuhiko Watanabe, Hiromu Nishimura, Yuuya Chikuki, Kunihiro Nakajima, and Yoshihiro Okada
Ryukoku University
Seta, Otsu, Shiga, Japan
Email: watanabe@rins.ryukoku.ac.jp, t160405@mail.ryukoku.ac.jp, t160389@mail.ryukoku.ac.jp,
nakajima.k216@gmail.com, okada@rins.ryukoku.ac.jp

***Abstract*—These days, many people use a Social Networking Service (SNS). Most SNS users are careful in protecting the privacy of personal information: name, age, gender, address, telephone number, birthday, etc. However, some SNS users disclose their personal information that can threaten their privacy and security even if they use non-real name accounts. In this study, we investigated tweets disclosing submitters' personal profile items which many of us think are not true. We collected 565 tweets where submitters used non-real name accounts and made promises to disclose their personal profile items, surveyed the details of their personal profile items disclosed by themselves, especially their ages, genders, heights, and foot sizes, and analyzed them statistically by applying the Shapiro-Wilk test of normality and the Welch's test. The results of these tests showed that most of the submitters disclosed their ages, genders, heights, and foot sizes honestly.**

Keywords–personal information; Twitter; SNS; privacy risk; Shapiro-Wilk test of normality; Welch's test.

I. INTRODUCTION

These days, many people use a Social Networking Service (SNS) to communicate with each other and try to enlarge their circle of friends. SNS users are generally concerned about potential privacy risks. To be specific, they are afraid that unwanted audiences will obtain information about them or their families, such as where they live, work, and play. As a result, SNS users are generally careful in disclosing their personal information. They disclose their personal information only when they think the benefits of doing so are greater than the potential privacy risks. However, some SNS users, especially young users, disclose personal information on their profiles, for example, real full name, gender, hometown and full date of birth, which can potentially be used to identify details of their real life, such as their social security numbers. In order to discuss this phenomenon, many researchers investigated how much and which type of information is disclosed in SNSs, especially, on Facebook. Researchers might think that personal information disclosed on Facebook is reliable, or it is possible to check whether personal information disclosed on Facebook is true. This is because

- Facebook users are required to register and disclose their real names when they first start using Facebook.
- Facebook users would be criticized by their friends if they disclose their information dishonestly.

On the other hand, a small number of researchers investigated how much and which type of information is disclosed by non-real name account users, such as Twitter users. Researchers might think that personal information disclosed by non-real name account users is unreliable. This is because

- nobody criticizes non-real name account users when they disclose their personal information dishonestly.
- true personal information can threaten their privacy and security even if they use non-real name accounts.

As a result, many of us think that it is natural for non-real name account users not to disclose their personal information honestly. Figure 1 shows tweets submitted by non-real name account user, *Rina*. In these tweets, *Rina* disclosed her personal profile items: her age, gender, birthday, zodiac sign, height, and foot size. Many of us think that these personal profile items were not true. However, we cannot check whether *Rina*

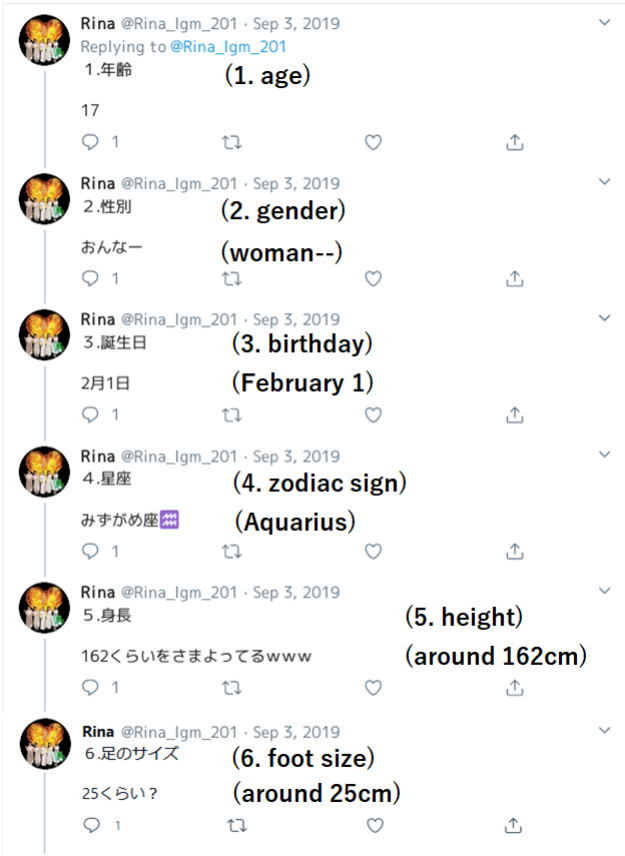


Figure 1. A non-real name account user, *Rina*, disclosed her personal profile items in her tweets.

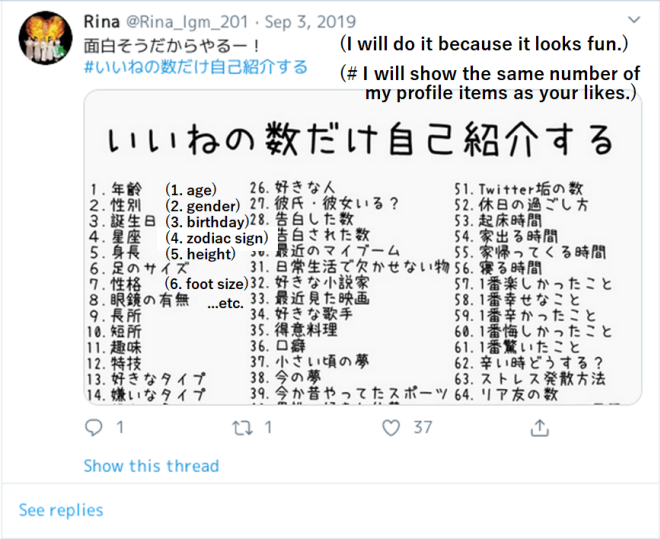


Figure 2. A tweet promising to disclose the same number of submitters' personal profile items as likes to it.

disclosed her personal profile items honestly because it is difficult to do it. To solve this problem, we collected tweets where non-real name account users made promises to disclose their personal profile items, analyzed them statistically, and showed that it is likely that most of the non-real name account users, especially young users, disclosed their personal information honestly [1]. In this paper, we survey and analyze these tweets by taking into account foot size that has not been paid attention to before and discuss whether their submitters disclosed their ages, genders, heights, and foot sizes honestly.

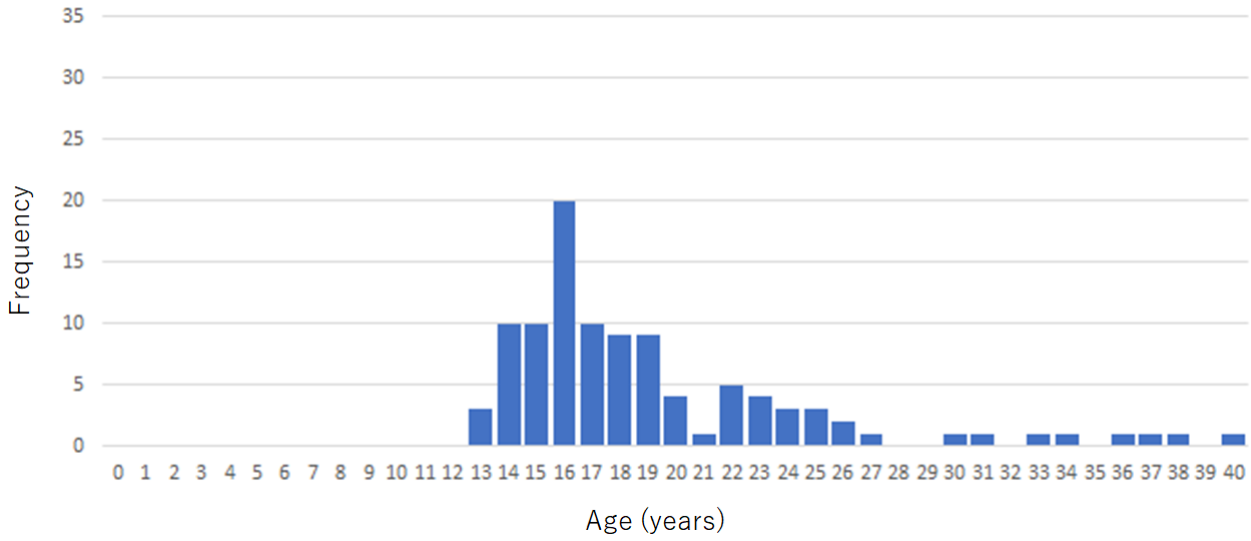
The rest of this paper is organized as follows: In Section II, we survey the related works. In Section III, we show how to collect tweets disclosing submitters' personal profile items. In Section IV, we survey the details of submitters' personal profile items, analyze them statistically, and show that it is likely that most of the submitters disclosed their personal profile items honestly. Finally, in Section V, we present our conclusions.

II. Related work

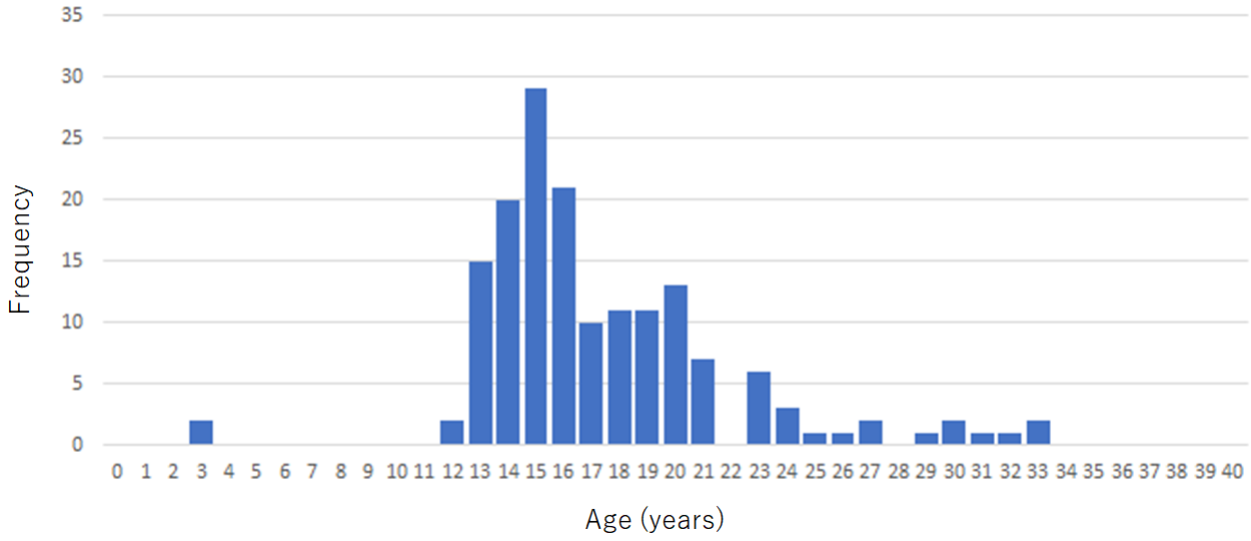
Personally identifiable information is defined as information which can be used to distinguish or trace an individual's identity such as social security number, biometric records, etc. alone, or when combined with other information that is linkable to a specific individual, such as date and place of birth, mother's maiden name, etc. [2] [3]. Unsafe disclosure of personal information on SNSs can lead to several concerns such as cyberbullies, addiction, risky behavior, and contact with dangerous communities [4]. Internet users are generally concerned about unwanted audiences obtaining personal information. Fox et al. reported that 86% of Internet users are concerned that unwanted audiences will obtain information about them or their families [5]. Also, Acquisti and Gross reported that students expressed high levels of concern for general privacy issues on Facebook, such as a stranger finding out where they live and the location and schedule of their classes, and a stranger learning their sexual orientation, name of their current partner, and their political affiliations [6]. However, Internet users, especially young users, tend to disclose personal information on their profiles, for example, real full name, gender, hometown and full date of birth, which can potentially be used to identify details of their real life, such as their social security numbers. As a result, many researchers discussed the reasons why young users willingly disclose personal information on their SNS profiles. Barnes argues that Internet users, especially teenagers, are not aware of the nature of the Internet and SNSs [7]. On the other hand, Hinduja and Patchin analyzed randomly sampled MySpace profile pages and reported that the majority of adolescents are responsibly using the web site [8]. Krasnova et al. reported that SNS users are primarily motivated to disclose personal information because SNSs are fun and convenient to develop their social networks [9]. Van der Heijden also reported that users' sense of enjoyment is a strong intention to use information systems [10]. For example, Liu et al. showed that the sense of enjoyment positively influences the intention of Chinese university students to use SNSs [11]. In contrast, Pavlou et al. reported that users' perception on privacy risk negatively influences their intention to use online services [12]. Viseu et al. reported that many online users believe the benefits of disclosing personal information in order to use an Internet site are greater than the potential privacy risks [13]. Joinson et al. reported that trust and perceived privacy had a strong effect on individuals' willingness to disclose personal information to a website [14]. Also, Tufekci found that concern about unwanted audiences had an impact on whether or not students revealed their real names and religious affiliation on MySpace and Facebook [15]. The authors also think that most students are seriously concerned about their privacy and security. However, they often underestimate the risk of their online messages and submit them. For example, Watanabe et al. reported that many students submit tweets concerning school events and these tweets may give a chance to other people, including unwanted audiences, to distinguish which schools students go to [16]. Hirai reported that many users experienced trouble in SNSs because they did not mind that strangers observed their communication with their friends [17]. On the other hand, Acquisti and Gross explain this phenomenon as a disconnection between the users' desire to protect their privacy and their actual behavior [6]. Dwyer concluded in her research that privacy is often not expected or undefined in SNSs [18]. Also, Livingstone points out that teenagers' concept of privacy does not match the privacy settings of most SNSs [19]. Hui et al. reported that online companies can induce users to disclose their personal information by offering benefits [20].

III. A collection of tweets disclosing submitters' personal profile items

It is difficult to collect tweets disclosing submitters' personal profile items, such as tweets in Figure 1, directly. To solve this problem, we focused on tweets where submitters promised their followers to disclose the same number of their own personal profile items as likes to their tweets. Figure 2 shows a tweet submitted by *Rina* on September 3, 2019. In this tweet, *Rina* promised her followers to disclose the same number of her personal profile items as likes to her tweet. Actually, *Rina* submitted 35 replies disclosing her personal profile items to her tweet shown in Figure 2 from September 3 to 9, 2019. The six tweets shown in Figure 1 were the first six



(a) The number of submitters who disclosed that they were men by age.



(b) The number of submitters who disclosed that they were women by age.

Figure 3. The number of submitters who disclosed their genders clearly by age.

replies submitted by *Rina* to her tweets shown in Figure 2. As of November 20, 2019, we confirmed that 37 likes were given to her tweet shown in Figure 2. Furthermore, we found many tweets promising to disclose the same number of their own personal profile items as likes to their tweets. As a result, it is easy to collect tweets disclosing submitters' personal profile items when we collect tweets promising to disclose submitters' personal profile items. The reasons why many Twitter users submitted tweets promising to disclose submitters' personal profile items might be

- they thought they looked fun,
- they wanted to draw attention, and
- they wanted to know how much attention was paid to their tweets.

In order to collect tweets promising to disclose submitters' personal profile items, we focused on images attached to these tweets. This is because many submitters attached the same image to their tweets and many personal profile items were listed in the image. As shown in Figure 2, *Rina* attached an image to her tweet and showed the list of personal profile items that she promised her followers to disclose in the image. Many twitter users attached the same image to their tweets promising to disclose their personal profile items. As a result, we used these shared images as key to collect tweets promising to disclose submitters' personal profile items. To be specific, we collected these tweets by using Twigaten [21]. Twigaten helps us to collect tweets to which the same image is attached. By

using Twigaten, we collected 565 Japanese tweets promising to disclose submitters' personal profile items on November 20, 2019. The obtained tweets were submitted from October 3, 2018 to November 20, 2019.

IV. An analysis of tweets disclosing submitters' personal profile items

It is difficult to determine whether an individual submitter disclosed his/her personal profile items honestly. For example, it is difficult to determine whether *Rina*, who submitted tweets in Figure 1 and Figure 2, was a woman. In this study, we discuss whether submitters disclosed their personal profile items honestly when they made promises to disclose them. In order to discuss this problem, we analyze submitters' genders, ages, heights, and foot sizes statistically.

A. Submitters' genders

As mentioned in Section III, we obtained the 565 tweets promising to disclose user's personal profile items. We surveyed these 565 tweets and their replies and, according to submitters' genders disclosed in the replies, classified them into

- 282 tweets (women)
- 156 tweets (men)
- 27 tweets (unclear)
- 100 tweets (no replies)

B. Submitters' ages

We also surveyed the 565 tweets and their replies and, according to whether submitters' ages were disclosed in their replies clearly, classified them into

- 276 tweets (clearly)
- 60 tweets (unclearly)
- 229 tweets (no replies)

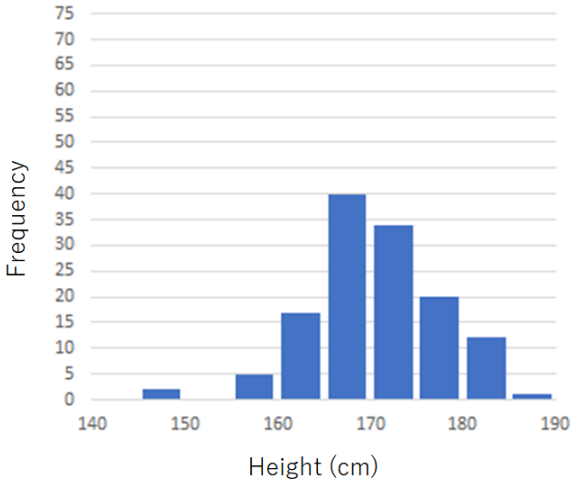
When submitter's age was disclosed such as "early 20s" and "over thirty", we determined that submitter's age was disclosed unclearly. Among the 276 tweets where submitters' ages were disclosed clearly, we found 102 and 161 tweets where submitters' genders were also disclosed clearly, men and women, respectively. Figure 3 shows the number of submitters, who disclosed their genders clearly, men and women, by age. As shown in Figure 3, the most popular age of men and women were 16 and 15 years old, respectively.

C. Submitters' heights

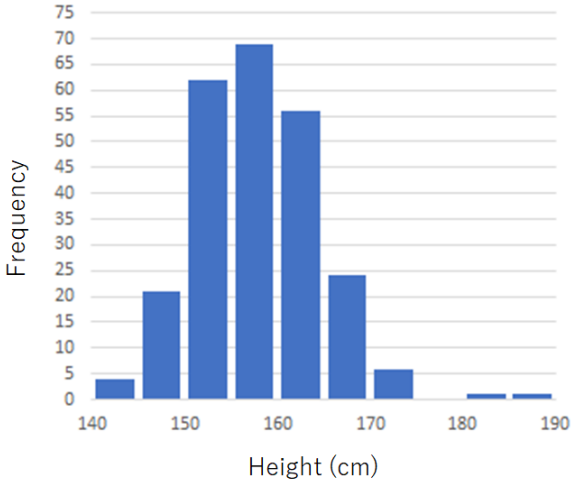
We also surveyed the 565 tweets and their replies and, according to whether submitters' heights were disclosed in their replies clearly, classified them into

- 401 tweets (clearly),
- 8 tweets (unclearly), and
- 156 tweets (no replies).

Among the 401 tweets where submitters' heights were disclosed clearly, we found 131 and 244 tweets where submitters' genders were disclosed clearly, men and women, respectively. Figure 4 shows the histogram of heights of submitters who disclosed their genders, men or women, clearly.



(a) the histogram of submitters' heights (disclosed genders: men).



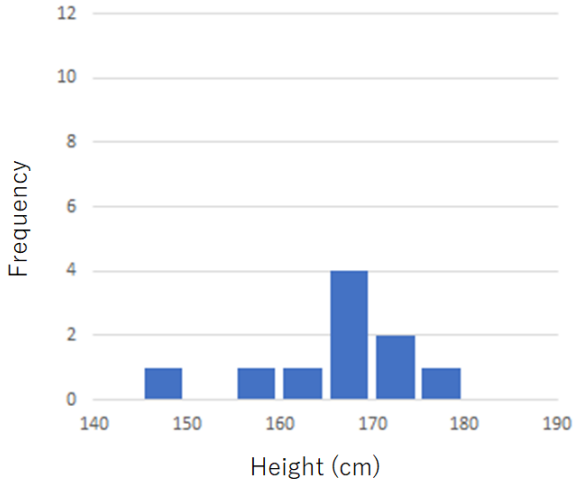
(b) the histogram of submitters' heights (disclosed genders: women).

Figure 4. The histogram of heights of submitters who disclosed their genders, men or women, clearly. (bin width = 5cm)

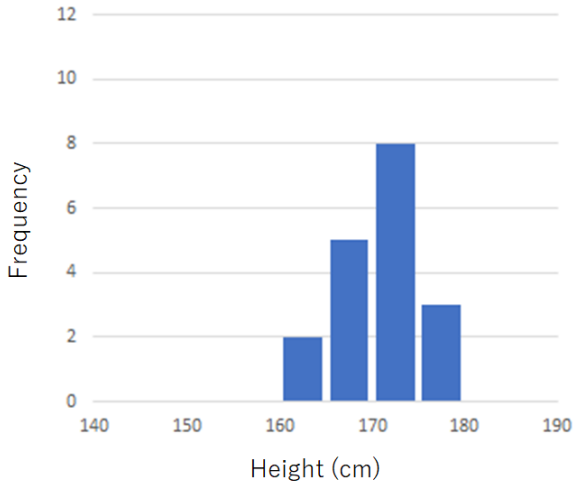
It is difficult to determine whether an individual submitter disclosed his/her personal profile items honestly. In this study, we statistically examine whether submitters disclosed their personal profile items honestly when they made promises to disclose their personal profile items and disclosed them in the same way as *Rina* did.

It is well known that our heights follow a normal (Gaussian) distribution [22]. As a result, if most of submitters disclose their ages, genders, and heights honestly, their heights would follow a normal distribution. Also, the average of their heights would be equal to the national average height in Japan. To solve this problem, in this paper, we conduct the statistical analysis on

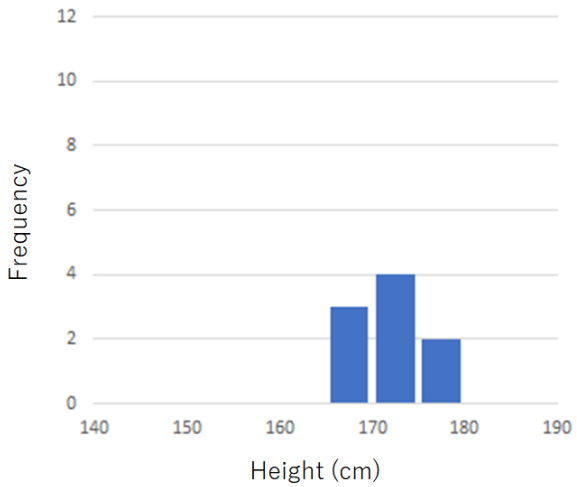
- 37 submitters who disclosed their genders (men), ages (15-17 years old), and heights clearly, and
- 60 submitters who disclosed their genders (women), ages (14-16 years old), and heights clearly.



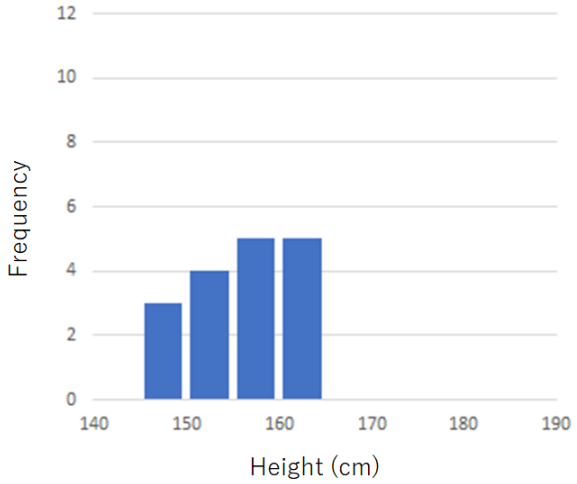
(a) men aged 15



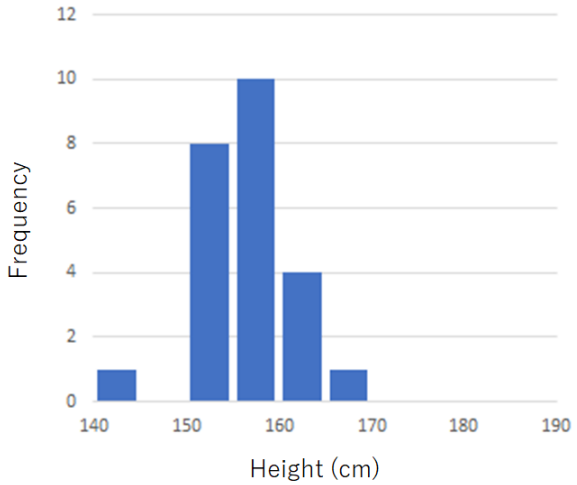
(b) men aged 16



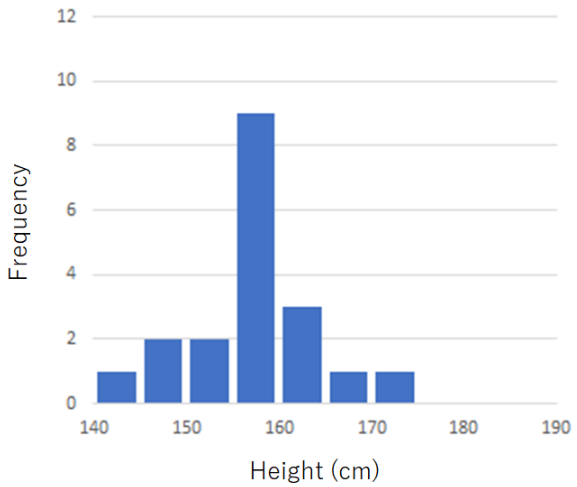
(c) men aged 17



(d) women aged 14



(e) women aged 15



(f) women aged 16

Figure 5. The histograms of heights of submitters who disclosed that they were men aged 15-17 and women aged 14-16 (bin width = 5cm).

TABLE I. THE RESULTS OF THE SHAPIRO-WILK TEST OF NORMALITY ON HEIGHTS

gender	age	sample size	W value	p-value
men	15	10	0.885	0.147
men	16	18	0.929	0.190
men	17	9	0.977	0.946
women	14	17	0.933	0.244
women	15	24	0.971	0.697
women	16	19	0.961	0.587

TABLE II. THE AVERAGE AND STANDARD DEVIATION OF SUBMITTERS' HEIGHTS

gender	age	sample size	average	standard deviation
men	15	10	165.5	7.58
men	16	18	169.2	4.08
men	17	9	171.3	3.84
women	14	17	155.0	6.11
women	15	24	155.7	4.99
women	16	19	156.9	6.54

TABLE III. THE NATIONAL AVERAGE AND STANDARD DEVIATION OF HEIGHTS IN JAPAN

gender	age	sample size	average	standard deviation
men	15	1411	168.37	5.75
men	16	1428	169.59	5.70
men	17	1427	170.46	5.82
women	14	1386	156.36	5.24
women	15	1413	156.76	5.36
women	16	1419	157.16	5.17

As shown in Figure 3, men aged 15-17 and women aged 14-16 were the most popular segments in the submitters' ages.

First, we discuss whether submitters' heights followed a normal distribution. Figure 5 shows the histograms of their heights. In order to discuss whether submitters' heights followed a normal distribution, we conducted the Shapiro-Wilk test of normality. The null hypothesis in this study was that submitters' heights followed a normal distribution. Table I shows the results of the Shapiro-Wilk test of normality on heights. As shown in Table I, the p-value in each case was greater than 0.05. As a result, the null hypothesis in each case was not rejected. In other words, submitters' heights, in each case of men aged 15-17 and women aged 14-16, followed a normal distribution.

Next, we discuss whether the average of submitters' heights was equal to the national average height in Japan. Table II shows the average of submitters' heights. Table III shows the national average height in Japan [23]. In order to discuss whether the average of their heights was equal to the national average height in Japan, we conducted the Welch's test. The null hypothesis in this study was that the average of submitters' heights was equal to the national average height in Japan. Table IV shows the results of the Welch's test. As shown in Table IV, the p-value in each case was greater than 0.05. As a result, the null hypothesis in each case was not rejected. In other words, in each case of men aged 15-17 and women aged 14-16, the average of submitters' heights was equal to the national average height in Japan.

TABLE IV. THE RESULTS OF WELCH'S TEST ON HEIGHTS

gender	age	Degrees of freedom	test statistic T	p-value
men	15	9.07	1.195	0.262
men	16	17.84	0.380	0.708
men	17	8.23	-0.675	0.518
women	14	16.29	0.914	0.374
women	15	23.91	1.060	0.300
women	16	18.30	0.179	0.860

The results of the Shapiro-Wilk test of normality and the Welch's test rarely happened when many submitters disclosed their ages, genders, and heights dishonestly. As a result, it is assumed that most of the submitters disclosed their ages, genders, and heights honestly. Furthermore, age, gender, and height were important personal information. It is likely that they disclosed not only their ages, genders, and heights but also other personal profile items honestly.

D. Submitters' foot sizes

We also surveyed the 565 tweets and their replies and, according to whether submitters' foot sizes were disclosed in their replies clearly, classified them into

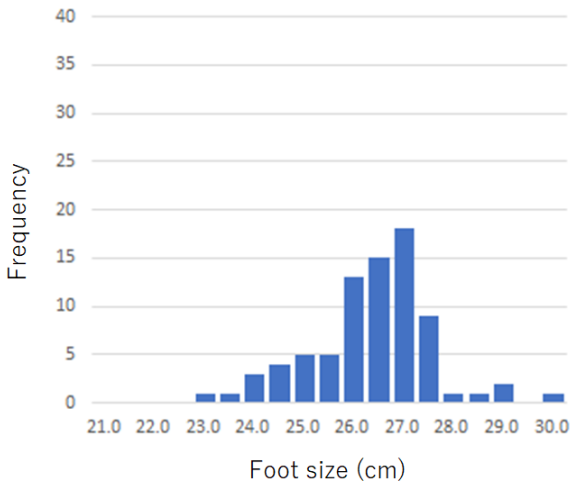
- 257 tweets (clearly),
- 20 tweets (unclearly), and
- 288 tweets (no replies).

Among the 257 tweets where submitters' foot sizes were disclosed clearly, we found 79 and 159 tweets where submitters' genders were disclosed clearly, men and women, respectively. Figure 6 shows the histogram of foot sizes of submitters who disclosed their genders, men or women, clearly. In this section, we statistically examine whether submitters disclosed their foot sizes honestly.

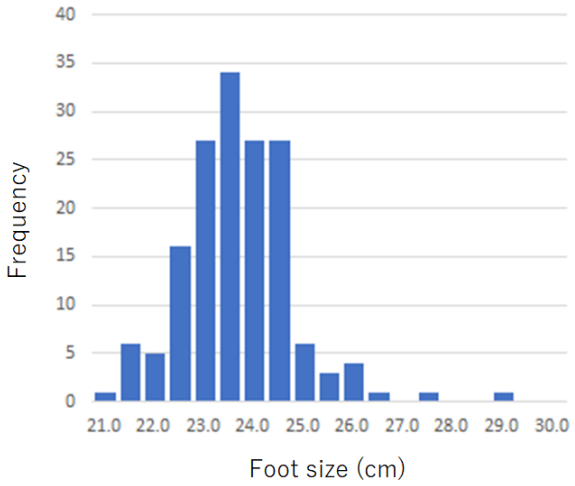
One thing to note is that the term *foot size* can be interpreted in two ways: foot length and shoe size. Katase et al. reported that most Japanese people do not know their own foot lengths, only know their shoe sizes, because they rarely have the opportunity to measure their foot lengths [24]. As a result, when Japanese submitters disclosed their foot sizes in their tweets, their values are mainly those of their shoe sizes. However, our shoe sizes are close to our foot lengths, and our foot lengths are generally considered to follow a normal distribution, like our heights. As a result, if most of submitters disclose their ages, genders, and foot sizes honestly, their foot sizes would follow a normal distribution. Also, the average of their foot sizes would be equal to the average of foot sizes surveyed in Japan. To solve this problem, in this paper, we conduct the statistical analysis on

- 32 submitters who disclosed their genders (men), ages (15-17 years old), and foot sizes clearly, and
- 57 submitters who disclosed their genders (women), ages (14-16 years old), and foot sizes clearly.

As shown in Figure 3, men aged 15-17 and women aged 14-16 were the most popular segments in the submitters' ages.



(a) the histogram of submitters' foot sizes (disclosed genders: men).



(b) the histogram of submitters' foot sizes (disclosed genders: women).

Figure 6. The histogram of foot sizes of submitters who disclosed their genders, men or women, clearly. (bin width = 1cm)

First, we discuss whether submitters' foot sizes followed a normal distribution. Figure 7 shows the histograms of their foot sizes. In order to discuss whether submitters' foot sizes followed a normal distribution, we conducted the Shapiro-Wilk test of normality. The null hypothesis in this study was that submitters' foot sizes followed a normal distribution. Table V shows the results of the Shapiro-Wilk test of normality on foot sizes. As shown in Table V, the p-values in the cases of men aged 15 and women aged 14 were less than 0.05, on the other hand, those in cases of men aged 16-17 and women aged 15-16 were greater than 0.05. As a result, the null hypothesis in four out of the six examined cases was not rejected. In other words, submitters' foot sizes followed a normal distribution in more than half of the examined cases.

Next, we discuss whether the average of submitters' foot sizes was equal to the average of foot sizes surveyed in Japan. Table VI shows the average of submitters' foot sizes. Table VII shows the average of shoe sizes in Japan, surveyed by Japan

TABLE V. THE RESULTS OF THE SHAPIRO-WILK TEST OF NORMALITY ON FOOT SIZES

gender	age	sample size	W value	p-value
men	15	8	0.779	0.023
men	16	16	0.917	0.153
men	17	8	0.826	0.063
women	14	16	0.857	0.017
women	15	23	0.930	0.108
women	16	18	0.908	0.080

TABLE VI. THE AVERAGE AND STANDARD DEVIATION OF SUBMITTERS' FOOT SIZES

gender	age	sample size	average	standard deviation
men	15	8	25.9	1.45
men	16	16	26.4	0.74
men	17	8	26.1	1.35
women	14	16	23.7	1.06
women	15	23	23.4	0.92
women	16	18	23.9	1.08

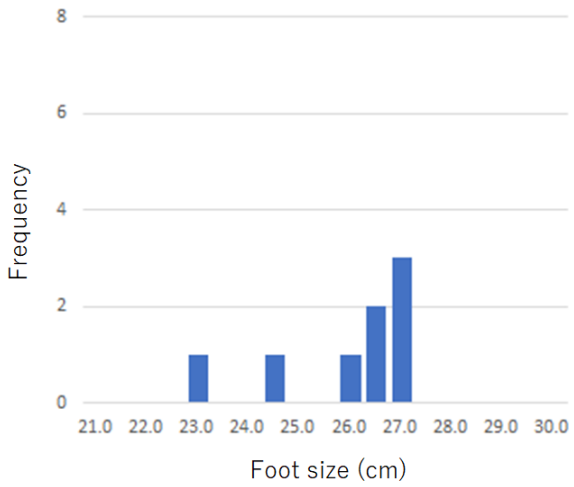
TABLE VII. THE AVERAGE AND STANDARD DEVIATION OF SHOE SIZES IN JAPAN (SURVEYED BY JLIA IN OCTOBER 2013)

gender	age	sample size	average	standard deviation
men	15	76	26.6	1.10
men	16	127	26.6	0.96
men	17	151	26.6	1.02
women	14	109	23.8	0.61
women	15	76	23.7	0.76
women	16	88	23.8	0.67

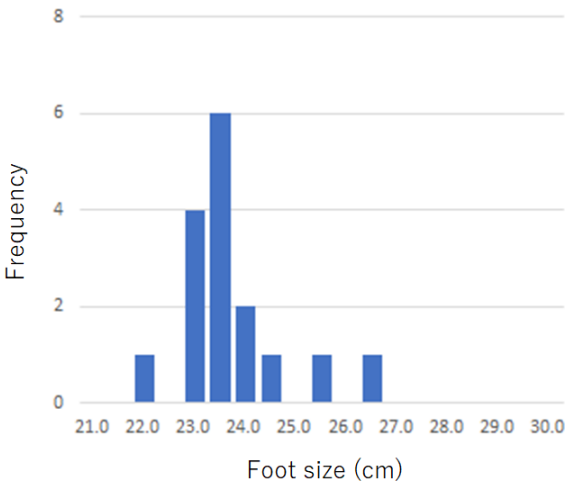
TABLE VIII. THE RESULTS OF WELCH'S TEST ON FOOT SIZES

gender	age	Degrees of freedom	test statistic T	p-value
men	15	7.87	1.326	0.222
men	16	21.92	0.982	0.337
men	17	7.43	1.032	0.334
women	14	16.49	0.369	0.717
women	15	31.63	1.424	0.164
women	16	19.76	-0.378	0.709

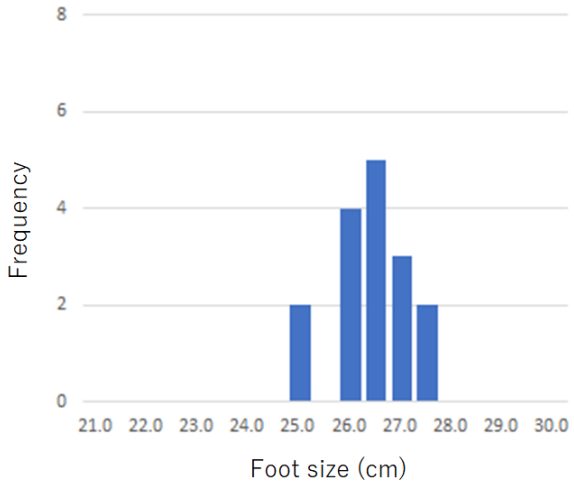
Leather and Leather Goods Industries Association (JLIA) in October 2013 [25]. As mentioned, when Japanese submitters disclosed their foot sizes in their tweets, their values are mainly those of their shoe sizes. As a result, we examine whether the average of submitters' foot sizes was equal to the average of shoe sizes in Japan surveyed by JLIA [25]. In order to discuss whether the average of their foot sizes was equal to the average of shoe sizes in Japan, we conducted the Welch's test. The null hypothesis in this study was that the average of submitters' foot sizes was equal to the average of shoe sizes in Japan. Table VIII shows the results of the Welch's test. As shown in Table VIII, the p-value in each case was greater than 0.05. As a result, the null hypothesis in each case was not rejected. In other words, in each case of men aged 15-17 and women aged 14-16, the average of submitters' foot sizes was equal to the



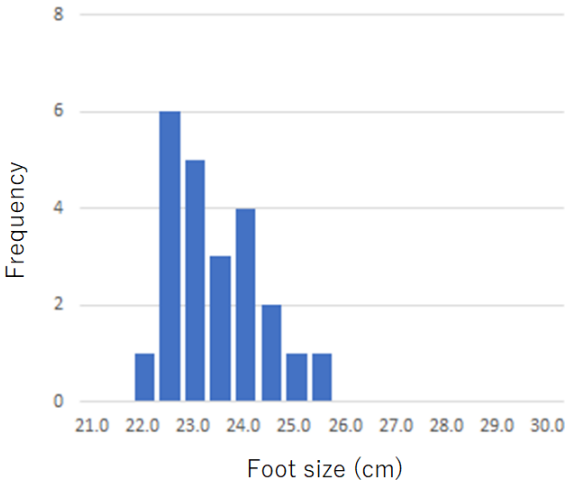
(a) men aged 15



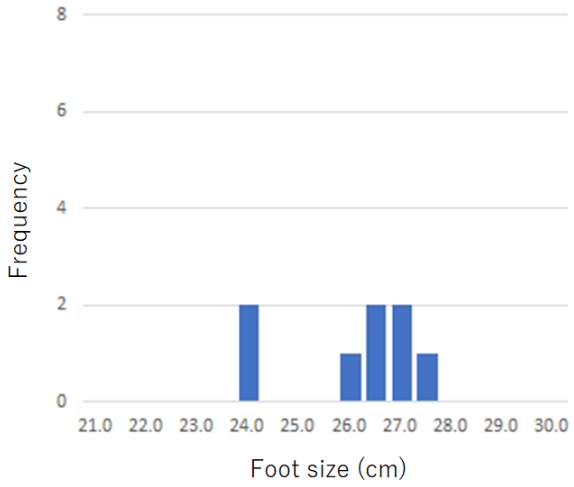
(d) women aged 14



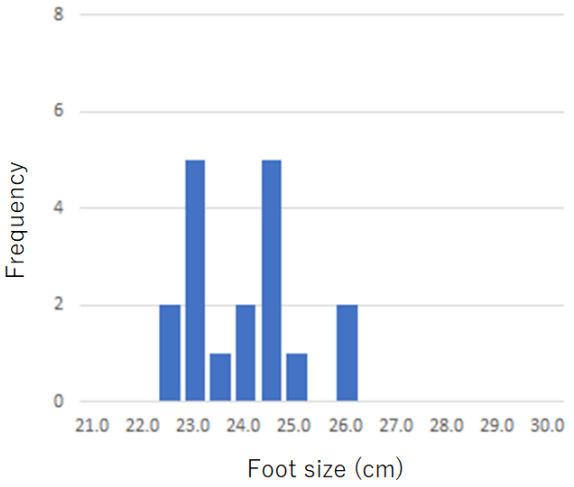
(b) men aged 16



(e) women aged 15



(c) men aged 17



(f) women aged 16

Figure 7. The histograms of foot sizes of submitters who disclosed that they were men aged 15-17 and women aged 14-16 (bin width = 1cm).

average of shoe sizes in Japan surveyed by JLIA.

The results of the Shapiro-Wilk test of normality and the Welch's test rarely happened when many submitters disclosed their ages, genders, and foot sizes dishonestly. As a result, it is assumed that most of the submitters disclosed their ages, genders, and foot sizes honestly.

V. Conclusion

In this paper, we investigated tweets disclosing submitters' personal profile items and analyzed submitters' ages, genders, heights, and foot sizes statistically. The results of the statistical analysis showed that it is likely that most of the submitters disclosed their personal profile items honestly. These personal profile items can threaten their privacy and security even if they use non-real name accounts. We are investigating whether submitters were concerned about their privacy and security risks caused by submitting tweets disclosing their personal profile items honestly [26]. Furthermore, we intend to conduct the same statistical analysis on tweets in languages other than Japanese.

References

[1] Y. Watanabe, H. Nishimura, Y. Chikuki, K. Nakajima, and Y. Okada, "An Investigation of Twitter Users Who Disclosed Their Personal Profile Items in Their Tweets Honestly," in Proceedings of the Sixth International Conference on Human and Social Analytics (HUSO 2020), October 2020, pp. 20–25. [Online]. Available: http://www.thinkmind.org/index.php?view=article&articleid=huso_2020_1_40_80035 [accessed: 2021-11-30]

[2] C. Johnson III, Safeguarding against and responding to the breach of personally identifiable information, Office of Management and Budget Memorandum, 2007. [Online]. Available: https://obamawhitehouse.archives.gov/sites/default/files/omb/assets/omb/memoranda/fy2007/m07-16.pdf [accessed: 2021-11-30]

[3] B. Krishnamurthy and C. E. Wills, "On the leakage of personally identifiable information via online social networks," Computer Communication Review, vol. 40, no. 1, 2010, pp. 112–117. [Online]. Available: https://doi.org/10.1145/1672308.1672328 [accessed: 2021-11-30]

[4] S. Valenzuela, N. Park, and K. F. Kee, "Is There Social Capital in a Social Network Site?: Facebook Use and College Students' Life Satisfaction, Trust, and Participation," Journal of Computer-Mediated Communication, vol. 14, no. 4, 07 2009, pp. 875–901. [Online]. Available: https://doi.org/10.1111/j.1083-6101.2009.01474.x [accessed: 2021-11-30]

[5] S. Fox et al., Trust and Privacy Online: Why Americans Want to Rewrite the Rules, The Pew Internet & American Life Project, 2000. [Online]. Available: http://www.pewinternet.org/2000/08/20/trust-and-privacy-online/ [accessed: 2021-11-30]

[6] A. Acquisti and R. Gross, Imagined Communities: Awareness, Information Sharing, and Privacy on the Facebook. Berlin, Heidelberg: Springer Berlin Heidelberg, 2006, pp. 36–58. [Online]. Available: https://doi.org/10.1007/11957454_3 [accessed: 2021-11-30]

[7] S. B. Barnes, "A privacy paradox: Social networking in the United States." First Monday, vol. 11, no. 9, 2006. [Online]. Available: http://firstmonday.org/article/view/1394/1312 [accessed: 2021-11-30]

[8] S. Hinduja and J. W. Patchin, "Personal information of adolescents on the Internet: A quantitative content analysis of MySpace," Journal of Adolescence, vol. 31, no. 1, 2008, pp. 125–146. [Online]. Available: https://doi.org/10.1016/j.adolescence.2007.05.004 [accessed: 2021-11-30]

[9] H. Krasnova, S. Spiekermann, K. Koroleva, and T. Hildebrand, "Online Social Networks: Why We Disclose," Journal of Information Technology, vol. 25, no. 2, 2010, pp. 109–125. [Online]. Available: https://doi.org/10.1057/jit.2010.6 [accessed: 2021-11-30]

[10] H. van der Heijden, "User acceptance of hedonic information system," Management Information Systems Quarterly, vol. 28, 12 2004, pp. 695–704. [Online]. Available: https://www.jstor.org/stable/25148660 [accessed: 2021-11-30]

[11] L. Liu, L. Zhang, P. Ye, and Q. Liu, "Influencing factors of university students' use of social network sites: An empirical analysis in china," International Journal of Emerging Technologies in Learning (iJET), vol. 13, 2018, pp. 71–86. [Online]. Available: https://online-journals.org/index.php/i-jet/article/view/8380/4839 [accessed: 2021-11-30]

[12] P. Pavlou, H. Liang, and Y. Xue, "Understanding and Mitigating Uncertainty in Online Exchange Relationships: A Principal-Agent Perspective," Management Information Systems Quarterly, vol. 31, 03 2007, pp. 105–136. [Online]. Available: https://www.jstor.org/stable/25148783 [accessed: 2021-11-30]

[13] A. Viseu, A. Clement, and J. Aspinall, "Situating privacy online: Complex perception and everyday practices," Information, Communication & Society, 2004, pp. 92–114. [Online]. Available: https://doi.org/10.1080/1369118042000208924 [accessed: 2021-11-30]

[14] A. N. Joinson, U.-D. Reips, T. Buchanan, and C. B. P. Schofield, "Privacy, trust, and self-disclosure online." Human-Computer Interaction, vol. 25, no. 1, 2010, pp. 1–24. [Online]. Available: https://www.tandfonline.com/doi/abs/10.1080/07370020903586662 [accessed: 2021-11-30]

[15] Z. Tufekci, "Can You See Me Now? Audience and Disclosure Regulation in Online Social Network Sites," Bulletin of Science, Technology & Society, vol. 28, no. 1, 2008, pp. 20–36. [Online]. Available: https://doi.org/10.1177/0270467607311484 [accessed: 2021-11-30]

[16] Y. Watanabe, H. Onishi, R. Nishimura, and Y. Okada, "Detection of School Foundation Day Tweets That Can Be Used to Distinguish Senders' Schools," in Proceedings of the Eleventh International Conference on Evolving Internet (INTERNET 2019), June 2019, pp. 34–39. [Online]. Available: https://www.thinkmind.org/index.php?view=article&articleid=internet_2019_2_30_40026 [accessed: 2021-11-30]

[17] T. Hirai, "Why does "Enjyo" happen on the Web? : An Examination based on Japanese Web Culture," Journal of Information and Communication Research, vol. 29, no. 4, mar 2012, pp. 61–71. [Online]. Available: http://doi.org/10.11430/jsicr.29.4_61 [accessed: 2021-11-30]

[18] C. Dwyer, "Digital Relationships in the "MySpace" Generation: Results From a Qualitative Study," in Proceedings of the 40th Annual Hawaii International Conference on System Sciences (HICSS '07), 2007, p. 19. [Online]. Available: https://ieeexplore.ieee.org/document/4076409 [accessed: 2021-11-30]

[19] S. Livingstone, "Taking risky opportunities in youthful content creation: teenagers' use of social networking sites for intimacy, privacy and self-expression." New Media & Society, vol. 10, no. 3, 2008, pp. 393–411. [Online]. Available: https://doi.org/10.1177/1461444808089415 [accessed: 2021-11-30]

[20] K.-L. Hui, B. C. Y. Tan, and C.-Y. Goh, "Online Information Disclosure: Motivators and Measurements," ACM Trans. Internet Technol., vol. 6, no. 4, Nov. 2006, p. 415441. [Online]. Available: https://doi.org/10.1145/1183463.1183467 [accessed: 2021-11-30]

[21] twigaten.204504byse.info. TwiGaTen. [Online]. Available: https://twigaten.204504byse.info/ [accessed: 2021-11-30]

[22] National Centre for Research Methods (NCRM). Using Statistical Regression Methods in Education Research. [Online]. Available: http://www.restore.ac.uk/srme/www/fac/soc/wie/research-new/srme/index.html [accessed: 2021-11-30]

[23] The Ministry of Education, Culture, Sports, Science and Technology (MEXT). the survey on physical strength and sporting ability (2018). [Online]. Available: https://www.e-stat.go.jp/stat-search/files?page=1&layout=datalist&toukei=00402102&tstat=000001088875&cycle=0&tclass1=000001133904&stat_infid=000031872003 [accessed: 2021-11-30]

[24] M. Katase, Y. Hirabayashi, S. Saitou, S. Watanabe, K. Kuribayashi, and K. Shionoya, "On the necessity of measuring children's shoe size (first report):preventing podiatry problems by including shoe size in physical measurements taken at school." Proceedings of the Annual

Meeting of Japan Ergonomics Society, vol. 43spl, no. 0, 2007, pp. 432–433. [Online]. Available: https://doi.org/10.14874/jergo.43spl.0.432.0 [accessed: 2021-11-30]

[25] Japan Leather and Leather Goods Industries Association (JLIA), "The report of the foot size measurement survey project (between 4 and 18 years old)," https://www.jlia.or.jp/library/member/foot/foot_sizes2013.pdf [accessed: 2021-11-30].

[26] Y. Watanabe, L. Mashimo, T. Nakano, H. Nishimura, and Y. Okada, "An Investigation of When Japanese Twitter Users Deleted Their Tweets Disclosing Their Personal Information," in Proceedings of the Seventh International Conference on Human and Social Analytics (HUSO 2021), July 2021, pp. 9–14. [Online]. Available: https://www.thinkmind.org/index.php?view=article&articleid=huso_2021_1_20_80016 [accessed: 2021-11-30]