International Journal of Interactive Multimedia and Artificial Intelligence

March 2021, Vol. VI, Number 5
ISSN: 1989-1660

unir LA UNIVERSIDAD EN INTERNET

"As leaders, it is incumbent on all of us to make sure we are building a world in which every individual has an opportunity to thrive. Understanding what AI can do and how it fits into your strategy is the beginning, not the end, of that process."
Andrew Ng

Special Issue on Artificial Intelligence, Paving the Way to the Future

INTERNATIONAL JOURNAL OF INTERACTIVE MULTIMEDIA AND ARTIFICIAL INTELLIGENCE
ISSN: 1989-1660 –VOL. 6, NUMBER 5

Editor's Note

Artificial Intelligence (AI) has become nowadays one of the main relevant technologies that is driven us to a new revolution, a change in society, just as well as other human inventions, such as navigation, steam machines, or electricity did in our past. There are several ways in which AI might be developed, and the European Union has chosen a path, a way to transit through this revolution, in which Artificial Intelligence will be a tool at the service of Humanity. That was precisely the motto of the 2020 European Conference on Artificial Intelligence ("Paving the way towards Human-Centric AI"), of which these special issue is a selection of the best papers selected by the organizers of some of the Workshops in ECAI 2020.

These workshops constitute a diverse list of different subjects that are relevant to AI at present, some that envision the future, and finally also multidisciplinary topics in an increasingly transversal discipline. This selection aims at proposing, discussing and finding ways to confront the many challenges that lie ahead and for which solutions need to be found. Designing the correct strategy is crucial to be able embrace a future in which AI ensures empowering people, making true the conference motto.

The selected papers belong to the following Workshops: Hybrid Intelligence for Natural Language Processing Tasks (HI4NLP); Applied Deep Generative Networks (ADGN); Declarative Problem Solving (DPSW); Advancing towards the Sustainable Development Goals: AI for a fair, just and equitable world (AI4EQ); Evaluating Progress in AI (epAI); Singular Problems for Health Care (SP4HC); Intelligent Information Processing and Natural Language Generation (IntelLanG); and Data Fusion for Artificial Intelligence (DAFUSAI).

With the workshop "Advancing Towards the Sustainable Development Goals (SDGs): Artificial Intelligence for a Fair, Just and Equitable World (AI4Eq)" we aimed to illustrate the R&D path that would confer a decisive role to AI in achieving the UN Agenda 2030. Eradicating poverty is a central objective of the SDGs so the emphasis is on AI benefits for low and middle income countries, and the growing pockets of underdevelopment in high income countries.

There is a growing interest in the role that AI can play in achieving SDGs on the part of international organisations, such as UN Global Pulse [1], UNHCR [2], the UNICEF Global Innovation Centre [3], the World Wide Web Foundation [4], the International Telecommunications Union [5], and even the World Economic Forum [6]. In order for AI to catalyse the necessary transformation promoted by the 2030 Agenda, a research agenda that is practice-oriented and that goes beyond cataloging AI risks and potentialities is required, in part as a counter-weight to the heavily-plugged corporate sector view on AI ethics, which is often little more than "ethicswash" for a program in which the effect of AI/S development and deployment will most likely be to increase inequality [7] [8]. The three papers selected from the submissions to AI4Eq for this special issue describe research on SDG-oriented AI applications, as well as AI tools conceived to support the development of AI respectful of, and even actively committed to, fundamental human rights, focusing particularly on protecting and empowering the most vulnerable and marginalized.

The paper "Achieving fair inference using error-prone outcomes" focusses on a field that is attracting increasing interest: the assessment of fairness criteria in supervised learning. The authors demonstrate that existing methods to assess and calibrate fairness criteria do not extend to the true target variables of interest, when error-prone proxy targets are used. They propose a framework that combines fair machine-learning methods, such as those found in the fairness literature, and measurement models found in the statistical literature; and illustrate their approach in a healthcare decision problem showing how a latent variable model to account for measurement error removes the unfairness detected previously.

The paper "Attesting Digital Discrimination Using Norms" also addresses the problem of digital discrimination arising from bias in machine-learning algorithms. In this case, the authors point to the need to provide non-expert users of machine-learning algorithms with simple tools to determine if a machine-learning system is potentially discriminatory, and to make explicit under which assumptions the systems are discrimination free. The authors suggest using "norms" as an abstraction to represent different situations that may lead to digital discrimination. In particular, they formalise non-discrimination norms in the context of machine-learning systems and propose a digital-discrimination attesting algorithm to check whether the systems violate these norms, illustrating its performance in three case studies where, in particular, gender and racial biases are identified.

"No App is an Island: Collective Action and Sustainable Development Goal-Sensitive Design" deals with the challenges of engineering ever more complex socio-technical systems to address "wicked" societal problems, with respect to satisfying qualitative human values and to assessing their impact on global challenges. The authors present a set of sets of design principles and an associated meta-platform, which focusses the design of socio-technical systems on the potential interaction of human and artificial intelligence with respect to three aspects: firstly, decision support regarding the codification of deep social knowledge; secondly, visualisation of community contribution to successful collective action; and thirdly, systemic improvement with respect to the SDGs through impact assessment and measurement. This SDG-sensitive design methodology is illustrated through the design of two collective action apps, one for encouraging plastic re-use and reducing plastic waste, and the other for addressing redistribution of surplus food.

The Workshop on Hybrid Intelligence for Natural Language Processing Tasks (HI4NLP) has provided a forum to discuss exciting research on hybrid technologies for NLP. In particular, our interest lies in those methodologies and architectures which combine and integrate symbolic information into statistical methods, including neural networks, thus allowing for building more transparent and interpretable models.

The paper "Assessing Lexical-Semantic Regularities in Portuguese Word Embeddings" introduces TALES, a new dataset with lexical-semantic word analogies. The authors use this resource to perform a detailed analysis of various word embeddings for Portuguese, including static representations such as GloVe [9] and word2vec [10], and current contextualized models (e.g., BERT [11]). Interestingly, this paper also discusses how distributional models can be used to enlarge lexical-semantic knowledge bases, which can be beneficial to various natural language processing tasks.

On "The Semantics of History. Interdisciplinary Categories and Methods for Digital Historical Research", Travé et al. present a conceptual framework for interdisciplinary research in History, focusing on data modeling and labeling methods. In particular, they propose identifying minimum units of information (units of topography, units of stratigraphy, and actors) and their relations, for which methodological aspects are described. A detailed case study on landscape archaeology shows the usefulness of the proposed framework, which takes advantage of knowledge obtained from several sources.

DOI: 10.9781/ijimai.2021.02.014

ECAI 2020 also hosted the first edition of the Declarative Problem Solving Workshop (DPSW), gathering researchers from different AI disciplines with a common interest in solving computational problems via their explicit representation in some declarative language. This covers, for instance, the solution of combinatorial problems, optimization, numerical constraints, planning, scheduling, temporal constraints, etc, or combinations of these categories provided that their specification is made in terms of some declarative formal language. The workshop spanned two days, August 29th and 30th, and included the presentations corresponding to thirteen accepted regular papers. The average audience was around 25 participants, reaching a maximum of 40 during the invited talk by Vladimir Lifschitz, that closed the event. Although the papers covered different disciplines like constraints, planning, natural language or pattern mining, perhaps the most frequent topic was the problem solving paradigm of Answer Set Programming (ASP). Two contributions obtained the workshop best paper award in a tie, and were extended into full journal papers included in this volume. The first best paper, "Smoke Test Planning using Answer Set Programming" by Tobias Philipp, Valentin Roland and Lukas Schweizer, presented a declarative method for optimizing the automated generation of smoke tests for hardware devices, that is, quick tests of main functionalities that may spot a large error in the early stages of hardware design. The second best paper, "An Application of Declarative Languages in Distributed Architectures: ASP and DALI microservices", by Stefania Costantini, Giovanni de Gasperis and Lorenzo de Lauretis, introduced an innovative combination of the microservices architecture with a modular variant of ASP, showing potential applications of declarative problem solving to Multi-Agent Systems, Internet of Things (IoT) or Cloud Computing.

The workshop on Intelligent Information Processing and Natural Language Generation (IntelLang) aimed to identify challenges and explore current results that arise from the interaction of Intelligent Information Processing techniques and research in Natural Language Generation (NLG), both at the level of models and applications. The use of intelligent data and information processing techniques can help in many relevant aspects of the NLG problem, for example in the contribution of formalisms for knowledge modeling and management, or in the development of models for the evaluation of the quality of the proposals, among many others. The workshop provided a forum for discussion of these new research directions and attracted a broad spectrum of contributions, emphasising either or both of the workshop's main themes - NLG and Information Processing. Our hope is that these contributions will serve to enhance the sharing of ideas among the two communities.

The paper "Improving Asynchronous Interview Interaction with Follow-up Question Generation" proposes a follow-up question generation model (followQG) capable of generating relevant and diverse follow-up questions based on the previously asked questions, and its answers. This model is integrated in a 3D virtual interviewing system, Maya, with capability of follow-up question generation, taking advantage of the implicit knowledge from deep pre-trained language models to generate rich and varied natural language follow-up questions. Empirical results suggest that followQG generates questions that humans rate as high quality, achieving 77% relevance, and a comparison with strong baselines of neural network and rule-based systems shows that it produces better quality questions.

The paper "Neural Scoring of Logical Inferences from Data using Feedback" proposes a neural network model that generates personalised lifestyle insights based on a model of their significance, and feedback from the user. These insights are derived from wearable sensors in smartwatches or sleep trackers, and their generation should adapt automatically to the preferences and goals of the user. Simulated analysis of the presented model shows its ability to assign high scores to a) insights with statistically significant behaviour patterns and b) topics related to simple or complex user preferences at any given time. The authors believe that the proposed neural networks model could be adapted for any application that needs user feedback to score logical inferences from data.

The first workshop on Evaluating Progress in Artificial Intelligence (EPAI) took place on September 4th and comprised 13 presentations and one invited talk from Professor Barry O'Sullivan, President of the European AI Association. There were over 30 attendants and a reasonably good number of (very active) attendees. EPAI 2020 served not only as a meeting point for people from different backgrounds and goals, but also to identify the most challenging/urgent needs for AI evaluation [12]. In this regard, it is very well-known that AI capabilities are growing at an unprecedented rate. Countless AI approaches and applications are being developed and can be expected over the long term. In hindsight, one would say that progress certainly has taken place just looking at the range of tasks that AI are able to solve autonomously today (according to the benchmarks, challenges, and competitions [13]) and were not solvable a few years ago, from machine translation to medical image analysis or self-driving vehicles [14]. Moreover, progress in AI is widely believed to have substantial social and economic benefits, and possibly to create unprecedented challenges. In order to properly prepare policy initiatives for the arrival of such technologies, accurate forecasts and timelines are necessary to enable timely action among policymakers and other stakeholders. However, there is still much uncertainty over how to assess and monitor the state, development, uptake, and impact of AI as a whole, including its future evolution, progress and benchmarking capabilities. While measuring the performance of state-of-the-art AI systems on narrow tasks is useful and fairly easy to do, where the assessment really becomes difficult, though, is in trying to map these narrow-task performances onto more general AI and how it can have an impact on society in terms of benefits, risks, interactions, values, ethics, oversight into these systems, etc.

EPAI papers covered different formalisations, methodologies and testbenches for the evaluation of AI systems with the final goal of measuring the field's rates of development, progress, and impact. Two contributions obtained, respectively, the workshop best paper award and runner-up award, and have been extended into full journal papers included in this volume. The best paper, "Artificial Canaries: Early Warning Signs for Anticipatory and Democratic Governance of AI" by Carla Zoe Cremer and Jess Whittlestone, propose a method for identifying early warning signs of transformative progress in AI, and discuss how these can support the anticipatory and democratic governance of AI. Their method combines expert elicitation and collaborative causal graphs to identify key mile-stones and identify the relationships between them.

The runner-up award paper, "Efficient and Robust Model Benchmarks with Item Response Theory and Adaptive Testing", by Hao Song and Peter Flach, investigate adaptive approaches to achieve better efficiency in model benchmarking. In this regard, they propose and analyse methods that allow machine learning practitioners to pick only a few representative datasets to quantify the quality of a technique, from which to extrapolate the performance on other datasets. To this end, they adapt existing approaches from psychometrics, Item Response Theory and Adaptive Testing specifically, by implementing certain modifications following the requirements of machine learning experiments, and present experimental results to validate the approach.

The workshop "Singular Problems for Healthcare (SP4HC)" was devoted to advances in Artificial Intelligence applied to Healthcare and Well-Being, with an active interest in frontier-of-knowledge Machine Learning subjects, sometimes named as singular Machine Learning problems. In particular, imbalanced classification, ordinal classification or multi-label classification, which are pervasive in important practical problems in healthcare, have consequently generated a tremendous interest. From the scientific point of view, the workshop intended to serve as a basis for proposing and discussing advances in the artificial intelligence arena, with a range of applications. Some contributions have dealt with common challenges in healthcare applications, as imbalanced classes and feature selection using simple interpretable classifiers like logistic models and decision trees. Others used feature extraction from images and artificial neural networks approaches. Natural language processing, reinforcement learning and model-based reinforcement techniques, recommender systems or echo state networks, which are an alternative to standard recurrent neural networks, have also been examined. Finally, several temporal modelling approaches to manage the concept drift phenomenon have been applied for identification and classification tasks. From the medical perspective, the papers of this workshop had coped with different medical topics like melanoma skin cancer detection, wellness application for providing personalized health activities, Type 1 diabetes blood glucose control for insulin dose decisions, and antimicrobial multidrug resistance in Intensive Care Units (ICUs) for their characterization and prediction.

The paper selected as best of this workshop, "Antimicrobial Resistance Prediction in Intensive Care Unit for Pseudomonas Aeruginosa using Temporal Data-Driven Models" proposes new paradigms to address the problem of the increasing bacterial resistance to antibiotics, a particularly serious problem in hospital's ICUs because of the vulnerability of these patients. Knowing in advance whether a concrete bacterium is resistant or susceptible to an antibiotic is a crux step for clinicians to determine an effective antibiotic treatment. This article focuses on cultures of the Pseudomonas Aeruginosa bacterium because is one of the most frequent, dangerous and difficult to treat in the ICU. Several temporal data-driven models are proposed and analysed to predict the resistance or susceptibility to a specific antibiotic family previously to obtain the result of the antimicrobial susceptibility test and only using historical data registered in the electronic health system. The approach provides reasonably accurate results for some antimicrobial families, and could be used by clinicians as an early-warning system to support the election of the antibiotic therapy. This early prediction can save valuable time to start the adequate treatment for an ICU patient.

The Workshop on Data Fusion for Artificial Intelligence (DAFUSAI) was dedicated to discuss this crucial problem from both theoretical and applied point of views. Classiffication, image processing, decision-making, big data or deep learning require collecting data and fusing them in appropriate ways in order to solve specific problems. For this reason, a huge effort is devoted to the developments and analysis of data fusion methods [15]. Aggregation functions are one of the most widely used methods in this sense. They are defined as monotone functions with appropriate boundary conditions and include, among others, most of the means or functions such as the product, the minimum or the maximum. However, in recent years it has been shown that the concept of aggregation function can be too restrictive, as it does not cover some examples which can provide good results in particular applications, as it is the case of the mode. Furthermore, some data fusion functions more general than aggregations,- the so called pre-aggregation functions [16]-, have been proposed to deal with problems ranging from classiffication [17] to the computational brain [18], with very promising results.

From the papers presented at the DAFUSAI Workshop, we have selected the paper by Sicui Zhang et al., entitled "Towards Multi-perspective Conformance Checking with Fuzzy Sets". The paper points out the problem faced by the organizations concerning the necessity to employ data-driven techniques to audit their business processes and ensure they comply with laws and internal/external regulations, which can be inefficient and subject to frauds or abuses. An increasingly popular approach to automatically assess the compliance of the executions of organization processes is represented by alignment-based conformance checking. These techniques present several advantages, e.g., by comparing real process executions with the models, they can show possible discrepancies. However, there are also some drawbacks, e.g., as they perform a crisp evaluation of process compliance, a behavior process is classified as either compliant or d-viant (even if such deviation is not severe). In this paper, the authors discuss about these problems, proposing a novel conformance checking approach aimed at representing actors' tolerance with respect to process deviations, taking it into account when assessing the severity of the deviations. Additionally, as a proof of concept, the authors performed a set of synthetic experiments to assess the approach. The obtained results clearly show the advantages of considering a more flexible evaluation of process deviations, and the impact on the quality and the interpretation of the obtained diagnostics.

With this special issue, we have attempted to describe progress in various areas of Artificial Intelligence by outlining some of the most interesting papers presented in several selected workshops of the European Conference in AI in 2020. These papers reveal a mature discipline, stating novel aspects, and opening questions and problems that require equally novel approaches. We hope that its reading can inspire new directions and solutions that can lead us to both theoretical and practical developments, helping us to advance to a future in which humans undoubtedly will be at the center of the technology.

Acknowledgements

We would like to thank the other co-organizers of the workshops involved, as well as the reviewers, for their invaluable contribution to the selection and quality of the articles included in this special issue.

Amparo Alonso-Betanzos[1], Pedro Cabalar[2],
Gracaliz P. Dimuro[10,11], Marcos Garcia[3], José Hernández-Orallo[4],
Raquel Hervás[5], Ángeles Manjarés[6], Fernando Martínez-Plumed[7,4],
Inmaculada Mora-Jiménez[8], Miquel Sànchez-Marré[9]

[1] Research Center on Information and Communication Technologies (CITIC). Universidad da Coruña (Spain)

[2] Universidade da Coruña (Spain)

[3] Research Center in Intelligent Technologies (CiTIUS). Universidade de Santiago de Compostela, Galicia (Spain)

[4] Valencian Research Institute for Artificial Intelligence (VRAIN). Universitat Politècnica de València (Spain)

[5] Facultad de Informática, Instituto de Tecnología del Conocimiento

Universidad Complutense de Madrid (Spain)

[6] Dept. Inteligencia Artificial ETSI, UNED (Spain)

[7] European Commission Joint Research Centre, Seville (Spain)

[8] Department of Signal Theory and Communications, Telematics and Computing Systems. Universidad Rey Juan Carlos, Madrid (Spain)

[9] Intelligent Data Science and Artificial Intelligence Research Centre (IDEAI-UPC) Dept. of Computer Science, Universitat Politècnica de Catalunya, Barcelona (Spain)

[10] Centro de Ciências Computacionais (C3) Universidade Federal do Rio Grande (Brazil)

[11] Departamento de Estadística, Informática y Matemáticas Universidad Pública de Navarra, Pamplona (Spain)

References

[1] UN Global Pulse, "Big data for development: Challenges and opportunities," 2012. [Online].

[2] UN, "Human Rights Based Approach to development," 2003. [Online].

[3] UNICEF Global Innovation Centre, "Generation AI," 2003. [Online].

[4] World Wide Web Foundation, "Artificial Intelligence: The Road Ahead in Low and Middle-Income Countries," World Wide Web Foundation, 2017. [Online].

[5] International Telecommunications Union, "AI for good global summit 2018," International Telecommunication Union, Geneva, 2018. [Online].

[6] World Economic Forum, "Harnessing artificial intelligence for the earth," 2018. [Online].

[7] R. Ochigame, "The invention of "Ethical AI"," 2019. [Online].

[8] Y. Benkler, "Don't let industry write the rules for AI," *Nature*, vol. 569, no. 7754, pp. 161–162, 2019.

[9] J. Pennington, R. Socher, C. D. Manning, "Glove: Global vectors for word representation," in *Empirical Methods in Natural Language Processing (EMNLP)*, 2014, pp. 1532–1543.

[10] T. Mikolov, K. Chen, G. Corrado, J. Dean, "Efficient estimation of word representations in vector space," in *Workshop Proceedings of the International Conference on Learning Representations (ICLR) 2013*, 2013. arXiv preprint arXiv:1301.3781.

[11] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, Minneapolis, Minnesota, June 2019, pp. 4171–4186, Association for Computational Linguistics.

[12] F. Martínez-Plumed, J. Hernández-Orallo, E. Gómez, "Tracking AI: The capability is (not) near," *Frontiers in Artificial Intelligence and Applications*, vol. 320: ECAI 2020, pp. 2915–2916, 2020.

[13] J. Hernández-Orallo, M. Baroni, J. Bieger, N. Chmait, D. L. Dowe, K. Hofmann, F. Martínez-Plumed, C. Strannegård, K. R. Thórisson, "A new AI evaluation cosmos: Ready to play the game?," *AI Magazine*, vol. 38, no. 3, pp. 66–69, 2017.

[14] F. Martínez-Plumed, E. Gómez, J. Hernández-Orallo, "Futures of artificial intelligence through technology readiness levels," *Telematics and Informatics*, vol. 58, p. 101525, 2021.

[15] G. Pereira Dimuro, J. Fernández, B. in Bedregal, R. Mesiar, J. Antonio Sanz, G. Lucca, H. Bustince, "The state-of-art of the generalizations of the Choquet integral: From aggregation and pre-aggregation to ordered directionally monotone functions," *Information Fusion*, vol. 57, pp. 27 – 43, 2020, doi: https://doi.org/10.1016/j.inffus.2019.10.005.

[16] G. Lucca, J. Sanz, G. Pereira Dimuro, B. Bedregal, R. Mesiar, A. Kolesárová, H. Bustince Sola, "Pre-aggregation functions: construction and an application," *IEEE Transactions on Fuzzy Systems*, vol. 24, pp. 260–272, April 2016, doi: 10.1109/TFUZZ.2015.2453020.

[17] M. Elkano, M. Galar, J. A. Sanz, P. F. Schiavo, S. Pereira, G. caliz Pereira Dimuro, E. N. Borges, H. Bustince, "Consensus via penalty functions for decision making in ensembles in fuzzy rule-based classification systems," *Applied Soft Computing*, vol. 67, pp. 728 –740, 2018, doi: https://doi.org/10.1016/j.asoc.2017.05.050.

[18] L. Ko, Y. Lu, H. Bustince, Y. Chang, Y. Chang, J. Ferandez, Y. Wang, J. A. Sanz, G. Pereira Dimuro, C. Lin, "Multimodal fuzzy fusion for enhancing the motor-imagery-based brain computer interface," *IEEE Computational Intelligence Magazine*, vol. 14, no. 1, pp. 96–106, 2019, doi: 10.1109/MCI.2018.2881647.

TABLE OF CONTENTS

OPEN ACCESS JOURNAL

ISSN: 1989-1660

The International Journal of Interactive Multimedia and Artificial Intelligence is covered in Clarivate Analytics services and products. Specifically, this publication is indexed and abstracted in: *Science Citation Index Expanded, Journal Citation Reports/ Science Edition, Current Contents®/Engineering Computing and Technology.*

Achieving Fair Inference Using Error-Prone Outcomes

Laura Boeschoten**, Erik-Jan van Kesteren**, Ayoub Bagheri, Daniel L. Oberski *

Utrecht University, Department of Methodology & Statistics, Utrecht (The Netherlands)
**Shared first authorship. These authors contributed equally to the current work

Received 16 October 2020 | Accepted 20 December 2020 | Published 22 February 2021

Abstract

Recently, an increasing amount of research has focused on methods to assess and account for fairness criteria when predicting ground truth targets in supervised learning. However, recent literature has shown that prediction unfairness can potentially arise due to measurement error when target labels are error prone. In this study we demonstrate that existing methods to assess and calibrate fairness criteria do not extend to the true target variable of interest, when an error-prone proxy target is used. As a solution to this problem, we suggest a framework that combines two existing fields of research: fair ML methods, such as those found in the counterfactual fairness literature and measurement models found in the statistical literature. Firstly, we discuss these approaches and how they can be combined to form our framework. We also show that, in a healthcare decision problem, a latent variable model to account for measurement error removes the unfairness detected previously.

Keywords

Algorithmic Bias, Fair Machine Learning, Latent Variable Model, Measurement Error, Measurement Invariance.

DOI: 10.9781/ijimai.2021.02.007

I. Introduction

Supervised learning is used to guide human decisions across a wide range of different fields. In sensitive areas such as healthcare or criminal justice, a key issue is that these decisions are equitable and fair. To this end, an active area of research investigates how fairness criteria can be incorporated into supervised learning [1]–[6]. This literature has focused on supervised learning for a single objective, assumed to be the target variable of interest.

However, focusing on fair inference for a single objective is not sufficient in many real-world applications. The motivating example for this paper is presented in [7]: a commercial health prediction algorithm, widely used by health insurance companies and affecting millions of patients, exhibits significant racial bias –at a given risk score, black patients are considerably sicker than white patients, as evidenced by signs of uncontrolled illnesses. The bias arises because the algorithm predicts healthcare costs rather than illness, but unequal access to care means that less money is spent caring for black patients than for white patients. Thus, substantial racial biases arise, despite healthcare cost appearing to be an effective proxy for health by some measures of predictive accuracy, and despite these predictions complying with conventional standards of fair inference on outcomes [8]. The situation presented in [7] is but one example of a more general common framework of using a proxy to measure outcomes which cannot be directly measured – another example would be predicting true criminal recidivism using only observed recidivism, which is an error-prone proxy [9]. In this paper, we suggest using an approach from the field of social science: to make use of *multiple* observable proxies to build a measurement model representing the unobserved (latent) variable of interest. We propose to integrate such an approach when developing prediction models. This issue cannot be ignored because fairness is generally conceptualised on a level more abstract than the proxy label [10]; for example, it is reasonable to require that fairness in a healthcare need prediction system should extend to a person's true health status. However, it is challenging to measure a patient's true health status, as such measures are typically impossible to observe directly. In social science, a common approach is to make use of multiple observable indicators to build a measurement model representing the unobserved (latent) variable of interest. We propose to integrate such an approach when developing prediction models.

This paper addresses the problem of prediction unfairness arising from measurement error. By considering the supervised learning problem at the level of a latent variable of interest, we reformulate the problem as one of adequate *measurement modelling*. In effect, instead of requiring perfect measurement to achieve fairness, we propose that researchers developing a prediction model to be used for decision-making collect several independent, possibly error-prone, measures of the variable of interest (e.g. health).These measures act like error-prone labels made by independent annotators, each containing some information about the true health status (similar to, e.g., [11],[12]). We then suggest to combine measurement models from the statistical literature with techniques from the literature on fair ML to assess and ameliorate the problem of unfair predictions in the face of measurement error.

Our contributions are as follows:

- We illustrate that existing methods to examine unfairness in error-prone outcomes are insufficient;
- We suggest a framework, based on the existing measurement modelling literature, to investigate and ameliorate such issues;
- We perform an exemplary analysis to demonstrate the sug-gested approach. In an existing healthcare application, this demonstrates that replacing one proxy with another does not lead to parity, while our approach does.

* Corresponding author.

E-mail addresses: l.boeschoten@uu.nl (Laura Boeschoten), e.j.vankesteren@uu.nl (Erik-Jan van Kesteren), a.bagheri@uu.nl (Ayoub Bagheri), d.l.oberski@uu.nl (Daniel L. Oberski).

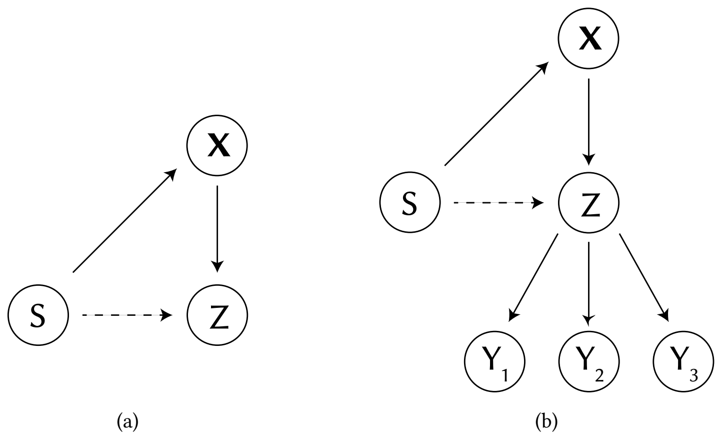


Fig. 1. Graphical representation of causal relations between the sensitive feature (S), the predictors (**X**), and the error-prone outcome (Z) in the naive case (A), in the measurement error framework (B), and in the measurement error framework with differential item functioning on the Y_1 proxy (C). The dotted arrow indicates the discriminatory causal pathway (as in [8]) which is blocked when performing fair inference, evaluating $E[Z \mid \mathbf{X}, S]$ to compute a risk score $\hat{Z}$.

In Section II, we provide a summary of basic concepts in fairness. In Section III prior approaches with respect to fair inference are discussed. In Section IV, the failure of these approaches is discussed when making use of proxies, and the proposed framework is introduced based on existing measurement models. In Section V the proposed framework is then applied to the exemplary data set provided by [7].

II. Problem Definition

We consider probabilistic classification and regression prob-lems with a set of features **X** and true outcome Z. Among the features, there is a sensitive feature $S \in \mathbf{X}$ (e.g. race, gender), with respect to which discriminatory predictions are to be avoided. Furthermore, although the prediction problem is with respect to the true outcome Z – e.g. "health" or "crime" – this outcome is not directly observed; instead, we have observed a set of error-prone proxy variables **Y**. For example, in practice a proxy for "health", $Y \in \mathbf{Y}$, might be the costs of healthcare or the number of chronic conditions experienced by the patient, whereas, instead of "crime", the number of arrests might be measured. Following [8], we represent the goal of the regression or classification problem as a query on the (generative) joint distribution $p(Z, \mathbf{X})$, potentially after conditioning on a set of "fixed" covariates C, i.e. the (discriminative) conditional joint $p(Z, \mathbf{X} \backslash \mathbf{C} \mid \mathbf{C})$. Typically, this query will be the point prediction $\hat{Z} := E(Z \mid \mathbf{X})$.

Following standard social-scientific measurement theory [13], the fact that **Y** is a measurement proxy for Z is reflected by a *causal model*, in the sense of [14], [15], in which $Z \longrightarrow \mathbf{Y}$, i.e., the true outcome is a common cause of all available proxy variables. Because Z is an unobserved latent variable, our causal model will be identifiable only through additional assumptions of conditional independence; we discuss these assumptions later. The key point to note here is that, generally,

$$E(Z \mid \mathbf{X}) \neq E(Y \in \mathbf{Y} \mid \mathbf{X}) \tag{1}$$

i. e. predictions using error-prone proxies as labels, $\hat{Y}$, will, of course, differ from the $\hat{Z}$ that would have been obtained had the true labels been available.

III. Related Work

A large and growing literature on fairness of predictions for the error-free outcome Z exists, with divergent and sometimes mutually exclusive definitions of the notion of algorithmic fairness. An excellent overview of this literature can be found in [6], which identified 20 separate definitions. Broadly, a distinction can be made between statistical metrics, distance-based measures, and causal reasoning [6].

Statistical metrics define fairness as the presence or absence of a (conditional) independence in the joint distribution $p(Z, \hat{Z}, S)$. For example, take a classification problem in which the decision is taken as $d := I(Z > \tau)$, where I is the indicator function and τ is some threshold on the predicted score. *Statistical parity* ("group fairness") is then defined as

$$p(d = 1 \mid S = s) = p(d = 1 \mid S = s') \tag{2}$$

for all $s \neq s'$, i.e., the decision should not depend on the sensitive attribute, whereas *predictive parity* is defined as

$$p(Z = 1 \mid d = 1, S = s) = p\,(Z = 1 \mid d = 1, S = s'\,) \tag{3}$$

for all $s \neq s'$—i.e. the positive predictive value should not depend on the sensitive attribute. Further definitions include conditional statistical parity [2], overall accuracy equality [1], and well calibration [4].

Distance-based measures of fairness account for the non-sensitive predictors $\mathbf{X}\backslash S$, in addition to the observed and predicted outcomes and sensitive attribute. The well-known "fairness through awareness" framework [3] generalises several of the preceding notions, such as statistical parity, by defining fairness as "similar decisions for similar people". Consider a population of potential applicants P, and consider any randomised output from the prediction algorithm, $M(x \in P)$. Fairness is achieved whenever the distance among the decisions M made for two people is at least as small as the distance between these people, i.e. when

$$D\,(M(x), \mathrm{M}(y)) \leq d(x, y) \tag{4}$$

for any $x, y \in P$. Here, D and d are arbitrary metrics on the distance between outputs and people, respectively. Careful choice of these metrics can yield some of the above definitions as special cases. Since the fairness condition can be trivially achieved, for example by always outputting a constant regardless of the input, the prediction model should be trained by minimising a loss function under the above constraint.

Finally, in recent years, results from the causal modelling literature have been leveraged to define and achieve "counterfactual" fairness [5], [8]. In these definitions one first considers a causal model involving Y, $X\backslash S$, and S such as Panel A of Fig. 1. This causal model then induces a counterfactual distribution $p_{do(s)}(\hat{Z}|X)$, i.e. the distribution we would observe if S were set to the value s [14]. [5] then defined counterfactual fairness as

$$p_{do(s)}(\hat{Z}|X) = p_{do(s')}(\hat{Z}|X) \tag{5}$$

Note that this definition looks superficially similar to the definition of statistical parity (group fairness), but is distinct because it refers to an individual. This definition has as a disadvantage that *any* causal effect of the sensitive attribute on the prediction is deemed illegitimate. Based on the same framework, [8] suggested a more general definition: some causal pathways originating in S are denoted discriminatory, while others are not. Fairness is then achieved by performing inference on a distribution $p^*(Z, \mathbf{X})$, in which the "fair world" distribution $p^*(Z, \mathbf{X})$ is close in a Kullback-Leibler sense to the original $p(Z, \mathbf{X})$, but all discriminatory pathways have been blocked (up to a tolerance) using standard causal inference techniques. Note that, if all causal pathways originating in S are deemed discriminatory and the tolerance set to zero, the counterfactual fairness criterion by [5] will be satisfied.

IV. Proposed Framework

A. Fair Inference in Error-prone Outcomes

The existing methods from Section III do not consider the target Z to be error-prone. However, in practice, the target feature $Y \in \mathbf{Y}$ in the data set is not a perfect representation of the true underlying outcome Z. There can be several sources for this imperfect representation. For example, the true underlying outcome of interest may not be directly measurable at all (i.e., $Z \neq Y$ for any possible Y). In this case, the outcome of interest will only partially explain any feature used as its proxy. For example, in using healthcare costs Y as a proxy for health Z, the observed value will in part be determined by other factors besides Z, such as the location of residence of the patient. Then, even if the outcome of interest were "true healthcare costs" – thus in principle measurable – the observed feature will in practice still not be an infallible proxy, because health records are never perfect observations and always contain some form of noise [16]. Together, such sources of noise in the observation process are termed "measurement error", and any outcome Z containing measurement error can be considered *latent* [17] and modelled as such.

Crucially, the presence of measurement error may result in unfair inferences for the error-prone outcome, even after applying the procedures presented in Section III to account for unfairness. This is shown in a compelling example by [7], who concluded that commercial algorithms used by insurance companies for patient referral contain a fundamental racial bias. In the algorithm under consideration, healthcare costs $Y \in \mathbf{Y}$ are used as a proxy for health Z. [7] illustrated that although there is no bias in healthcare costs, there is strong racial bias in other proxies of health such as whether patients have chronic conditions. Specifically, in order to be referred to a primary care physician, the true underlying health status Z of black patients was worse than that of white patients.

[7] concluded that fair inference requires selecting a better proxy for health as the outcome variable Z. Indeed, their analyses were possible precisely due to the availability of different proxies of health, such as the number of chronic conditions. However, we note that solving racial bias in a new proxy does not guarantee the absence of racial bias in other proxies indicating other aspects of health. Instead, here we suggest incorporating several proxies, or *indicators* $\mathbf{Y}$ in a measurement model for the unobserved, error-prone outcome Z [18]. In the next section, we introduce the existing literature on measurement models and its approach to fair inference.

B. Fair inference in Measurement Models

When outcomes are thought to be error-prone, an existing literature suggests the use of measurement models [16], [19]. At their core, measurement models describe the causal relationship between observed scores $\mathbf{Y}$ and unobserved "true scores" Z as $Z \longrightarrow \mathbf{Y}$. A measurement model adequately represents the empirical conditions of measurement if conditional independence can be assumed [20]. More specifically, measurement models assume that Y_1 and Y_2 are conditionally independent given Z, i.e.,

$$p\,(Y_1, Y_2 \mid Z) = p\,(Y_1 \mid Z)\, p\,(Y_2 \mid Z) \qquad (6)$$

A plethora of variations of measurement models assuming conditional independence have been developed, such as latent class models [21], item response models [22], mixture models [23], factor models [24], structural equation models [25], and generalised latent variable models [26].

Measurement models are suggested here as a convenient way to account for a latent variable's relationship to sensitive features. The measurement error of a proxy variable (e.g. Y_1) is then assumed to differ over different groups of S. To account for group differences in proxy variables, a large body of literature is available where this issue is known under different labels. Generally, these approaches are applied within the structural equation modelling (SEM) framework [27], as SEM explicitly separates the measurement model ($Z \longrightarrow \mathbf{Y}$) from the structural model ($\mathbf{X} \longrightarrow Z$). Approaches for investigating how features S influence Z are investigating item bias [28], Differential Item Functioning (DIF) [29] and measurement invariance [30]. For an extensive overview of the different approaches and their benefits and drawbacks, we refer to [30]–[33].

C. Proposed Method for Fair Inference on Latent Variables

We propose our framework for fair inference on outcomes which are measured only through error-prone proxies in a step-by-step manner. To clarify the framework and make it more comparable to earlier work, we use the running example of health risk score prediction from [7]. Their healthcare data set contains several clinical features $\mathbf{X}$ at time point $t-1$ (e.g., age, gender, care utilisation, biomarker values and comorbidities) which are used to predict healthcare cost Z at time t. In addition, the patient's race is the sensitive feature S, coded as $S = b$ for black patients and $S = w$ for white patients. The relations between these features are shown in panel A of Fig. 1.

Based on X, the expectation of a persons' healthcare cost is used as a risk score $\hat{Z} := E[Z \mid \mathbf{X}, S]$. The risk score is used to make a decision D to refer a patient to their primary care physician to consider program enrolment. More specifically $d = 1$ if Z is above the 55th percentile. In this setting, attributes $\mathbf{X}$ can be legitimately controlled. However, conditional on $\mathbf{X}$ both groups in S should have equal probability of being referred:

$$p\,(d = 1 \mid \mathbf{X} = x, S = b) = p\,(d = 1 \mid \mathbf{X} = x, S = w) \qquad (7)$$

As mentioned in Section A and shown by [7], this procedure leads to bias in other proxies of Z, such as a patient's number of chronic conditions.

Our proposed framework is a SEM implementation of the second and third panels of Fig. 1. The general structure of the model is that of a Multiple Indicator, Multiple Causes (MIMIC) model.

In SEM, a latent variable (a hypothetical construct that is not directly observed) can be related to observable variables, such as indicators and causes of the latent variable, through sets of regression equations [34] and where parameters are typically estimated by means of maximum-likelihood [35]. A MIMIC model is a particular structure of a SEM model where a latent variable is simultaneously related to both observed indicator and cause variables [36]. In our model, the outcome variable Z (e.g., health) has multiple proxy indicators (e.g., chronic conditions, healthcare costs, hypertension), and the $\mathbf{X}$ features predict Z directly (thus the proxies only indirectly). A graphical representation of the MIMIC SEM model is shown in Fig. 2. This implementation imposes additional assumptions on the general causal

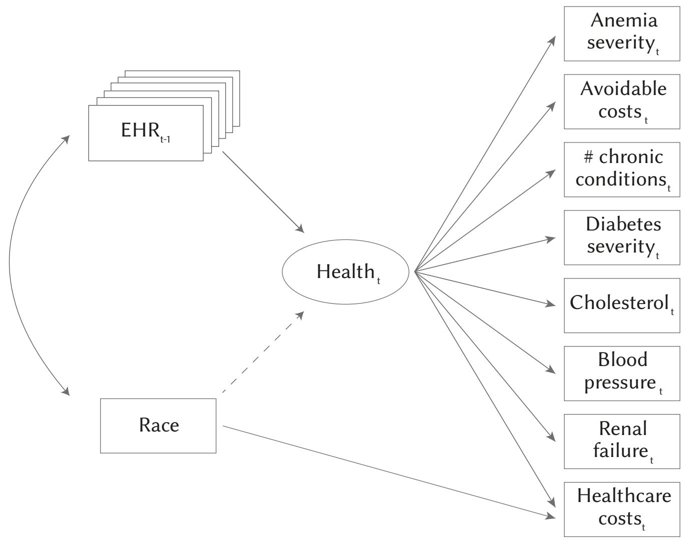


Fig. 2. Structural equation model for the proposed framework on the healthcare data set. For clarity, residual variances of the endogenous variables are not drawn in the diagram. EHR stands for Electronic Health Record. For more information on the variables used in the model, see [7].

graphs, most notably linear relationships between the variables and multivariate Gaussian residuals.

We implement our proposed correction procedure on the outcome variable Z in an existing fair inference approach [8] by means of the following steps:

1. The data-set is split in half to obtain a training set and a test set.
2. Regression parameters ($\mathbf{X}, S \longrightarrow Z$) are estimated on the training set using the MIMIC model.
3. The path from race to health is blocked by setting $S = b$ for all rows in the test set.
4. Predictions are generated for the adjusted test set by using the parameter estimates obtained in step 2.

To summarise, during estimation of the regression parameters ($\mathbf{X} \longrightarrow Z$), health is conditioned on race, but during prediction the path from race to health is blocked by setting $S = b$. Following the notation of [8], this yields a "fair world" distribution $p^*(Z, \mathbf{X})$. The expectation $\hat{Z} = E[Z \mid \mathbf{X}, S]$ is then computed from this distribution, meaning for two participants who differ only on S but not on $\mathbf{X}$, the risk score $\hat{Z}$ will be exactly the same. Because in SEM the latent outcome Z is modelled as a linear combination of the different proxies, the risk score is a reflection of the underlying health rather than only health cost.

V. Experiments

In this section, we evaluate the proposed framework on an application of the procedures discussed in this paper. We first prepare the data set as provided by [7] to create a basic risk score based on healthcare cost similar to the commercial risk score reported in their paper. Then, we illustrate our argument from Section A: we perform fair inference on the proxy measure for health (healthcare cost) to show that this does not solve the issue of unfairness in other proxy measures. This is a reproduction of the results shown by [7]. Next, we use the SEM framework from Section C to show how including a formal measurement model for Z – as in panel B of Fig. 1 – can largely solve the issue of unfairness in the proxies. Last, we show how existing differential item functioning (DIF) methods in the SEM framework – panel C of Fig. 1 – can aid in interpreting the extent to which proxy measures contain unfairness. Fully reproducible R code for this section is available as supplementary material to this paper at the following DOI: 10.5281/zenodo.3708150.

A. Data Preparation and Feature Selection

Log-transformations are applied to highly skewed variables at time-point t, such as costs, to meet the assumption of normally distributed residuals in regression procedures. As an additional normalisation step, the predictors at time-point $t-1$ are re-scaled to homogenise their levels of variance. The data set is then split into a training and a test set. In this section, estimation is always done on the training set and inference is done on the test set.

To simplify our proposed framework for the purpose of this application, we select a subset of features at time-point $t-1$ for prediction of the target of interest at time point t, health. We want our procedure to be comparable to the commercial algorithm which produces the risk scores described in [7]. If the features we select are the same features used by the commercial algorithm, then our procedure would yield very similar results upon generating a risk score. Unfortunately, the predicted risk scores used by [7] cannot be replicated exactly using the provided data set.

To select the subset of predictor features for further use in our procedure, we performed a LASSO regression [37] where all available features at time-point $t-1$ are used as predictor variables, and the provided algorithmic risk score at time-point t is used as a target. Following the guidelines by [38], we used cross-validation to select the optimal λ penalty value. This yields a set of non-zero predictors which predict the algorithmic risk score well.

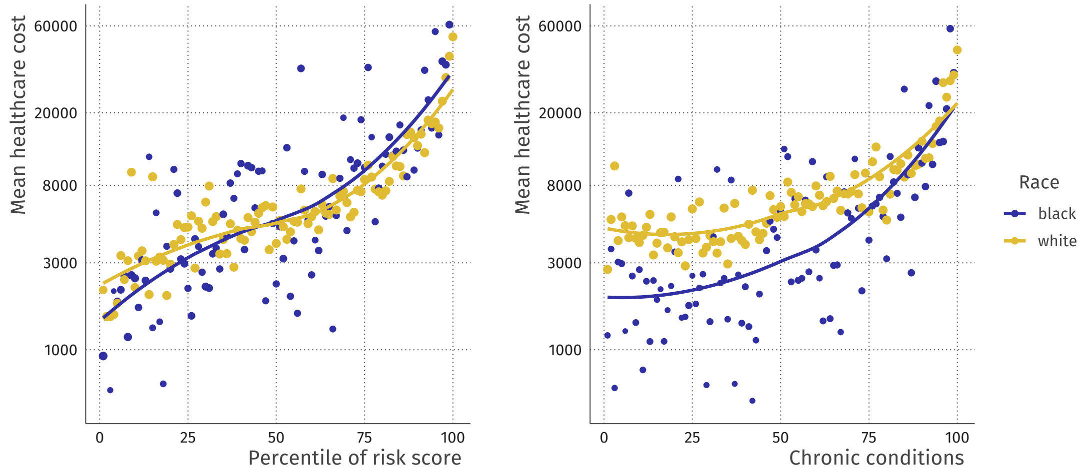
Superman's rank correlation between the commercial and the replicated risk score is high $\rho = .82$, indicating that the commercial and replicated risk scores perform similarly in the rank-based cutoff applied in [7]. The predictors selected in this model are used as predictors X in the structural equation models of the following sections.

B. Fair Inference on Cost as a Proxy of Health

Pane A of Fig. 1 illustrates conditional statistical parity as defined by [6]. To perform standard statistical parity correction, the outcome Z is conditioned on sensitive feature S when estimating the coefficients of the prediction model ($X \longrightarrow Z$), and during prediction all subjects are assumed to have the same level of S, e.g., $S = b$, such that

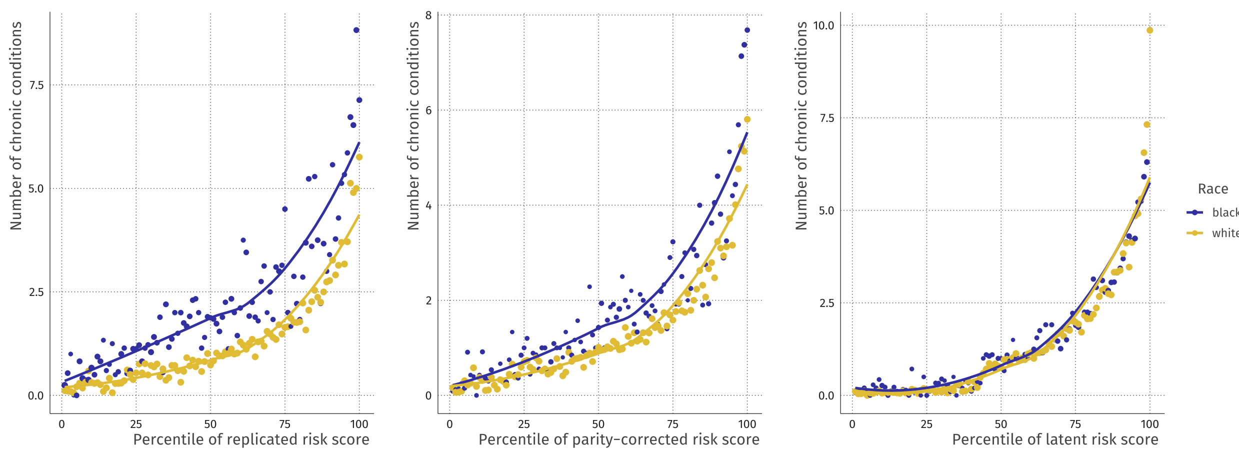
$$p\,(Z = z \mid \mathbf{X} = x, S = b) = p\,(Z = z \mid \mathbf{X} = x, S = w) \tag{8}$$

However, in the current situation we do not measure Z directly, but only a proxy $Y \in \mathbf{Y}$. Standard parity correction for this proxy does not necessarily mean the parity is achieved for other proxies [7]. The



(a) (b)

Fig. 3. Although the risk score displays statistical parity on healthcare costs (no differences between the lines in panel A), these costs conditional on health (as measured by chronic illness) depends on race (panel B). This causes statistical disparity for the risk score on the level of health (Fig. 4, panel B). Figure replicated from [7].



(a) (b) (c)

Fig. 4. Effect of including a measurement model in constructing risk scores. The first panel shows the uncorrected risk score based on healthcare cost, the middle panel shows the same risk score but corrected for the sensitive feature, and the third panel shows the corrected risk score based on the latent health outcome using a measurement model.

reason for this is explained in Fig. 3. Pane A illustrates that statistical parity is present when plotting the risk score against healthcare costs, meaning that for a given risk score, the healthcare costs for both races are approximately equal. However, Pane B illustrates that when the number of chronic conditions are plotted against healthcare costs, there are differences between the two race groups, meaning that for a given amount of chronic conditions, white patients cost more than black patients.

As a result, standard statistical parity correction on healthcare cost does not remove the disparity in chronic conditions. This becomes visible when comparing Pane B of Fig. 3 with Pane A of Fig. 4. In addition, from Pane B of Fig. 4 it can be seen that the results improve compared to not including race at all (Pane A of Fig. 4), yet race differences remain for the chronic conditions proxy. As a consequence, individuals belonging to $S=b$ will still have a lower health status when being selected for intervention.

C. Fair Inference on Latent Health

A cause for the fact that conditional statistical parity is not met when following Pane A of Fig. 1 can be that $\hat{Z}$ is a (bad) proxy. Instead of using one bad proxy, it is better to use multiple (bad) proxies as indicators of an unobserved latent variable measuring 'true health'. How such a model can be specified is illustrated in Pane B of Fig. 1. Such a model can be applied in practice by following the steps in the framework described in Section C. Similarly to [6], the sensitive feature is excluded during prediction.

Fig. 4 shows the effect of including a measurement model in constructing risk scores. The figure illustrates that using a measurement model with multiple imperfect measurements of health as indicators for 'true health' substantially improves conditional statistical parity, when compared to either the uncorrected risk score on a proxy, or a parity-corrected risk score on the proxy. Additionally, Table I shows a numerical summary which corroborates this finding. Here, we created a prediction model for the number of chronic conditions using both risk score and race. The parameter for race then indicates whether a race difference exists for health, conditional on the risk score. This conditional dependence becomes close to 0 when using the latent risk score (95% CI = [´0.113, 0.012]). Thus, by using this measurement model, the problem that individuals belonging to $S=b$ had a lower health status when being selected for intervention is minimised.

TABLE I. Estimated Conditional Parity on the Number of Chronic Conditions for Different Risk Scores. β Parameters Are Linear Regression Parameters, Indicating the Deviation of White Patients From Black Patients in the Number of Chronic Conditions, Conditional on Risk Score. For Example, a Value of -0.963 Means that White Patients Have on Average a 0.963 Fewer Chronic Conditions for the Same Risk Score

Risk score	β	2.5%	97.5%
Replicated	-0.963	-1.063	-0.864
Parity-corrected	-0.577	-0.677	-0.478
Latent	-0.051	-0.113	0.012

D. Investigating Unfairness in Proxies

When using a measurement model with multiple imperfect measurements of health as indicators of 'true health', differences in measurement error over the different groups of the sensitive feature can still be present. Panel C of Fig. 1 illustrates how differences over the sensitive feature groups in the error prone indicator variables can be incorporated directly when estimating 'true health'. For example, differences in measurement error of healthcare cost can be present for the different groups of race.

Including a DIF parameter δ on the healthcare cost variable yields a model which fits significantly better on the test set than the model without the DIF parameter ($\chi^2(1) = 50,\ p < 0.001$). The value of the DIF parameter on cost is estimated as $\delta = 0.198$ (95% CI = [0.172, 0.225]). This means that for the same level of health, the log-healthcare costs of the white race class in this data set is estimated to be 0.198 higher. This means that the cost of healthcare for white patients is $(e^{0.198} - 1) \cdot 100\% = 21.9\%$ higher than that for black patients, given an equal level of health as measured by the measurement model (95% CI = [18.7, 25.2]).

Applying the same procedure to the other indicators leads to estimates of DIF for those indicators. The results are shown in Table II. This table shows that some proxies have stronger DIF than others, meaning some proxies are more unfair than other proxies. Notable, the avoidable healthcare cost and the renal failure items have low levels of DIF for race, whereas the healthcare cost and the number of active chronic conditions have strong DIF.

TABLE II. Estimated Differential Item Functioning Parameters for Each Indicator (Proxy) of Health. δ Parameters Should Be Interpreted as the Mean Deviation of the Black Patients Compared to the White Patients Given Health.

Indicator	δ	2.5%	97.5%
No. active chronic conditions	0.453	0.364	0.541
Mean blood pressure	-0.262	-0.320	-0.204
Diabetes severity (HbA1c)	-0.343	-0.391	-0.296
Anemia severity (hematocrit)	0.250	0.231	0.268
Renal failure (creatinine)	-0.019	-0.025	-0.014
Cholesterol (mean LDL)	-0.235	-0.317	-0.153
Healthcare cost (log)	0.198	0.172	0.225
Avoidable healthcare cost (log)	-0.052	-0.096	-0.008

VI. Conclusion

In this paper, we have argued that when measurement error is at play, performing fair inference on a proxy measure of the outcome is insufficient to achieve a fair inference on the true outcome. This manifests itself, as shown in [7], as unfairness in other proxy measures of the outcome of interest. Alternatively, in this study we proposed to make use of existing measurement models containing multiple error-prone proxies for the outcome of interest. In addition, fair inference can be accounted for in each of these proxies simultaneously if needed by allowing for measurement error in proxies to differ over groups defined by differing values of a sensitive feature. We provided a framework to perform these estimations and applied this framework to the exemplary data set provided by [7]. Here, it was concluded that fair inference was accounted for when multiple proxies were used in a measurement model instead of a single proxy. Additionally accounting for differences in measurement error over race groups was not needed to further improve fairness in predicted risk scores, although substantive group differences were found for some proxies.

References

[1] R. Berk, H. Heidari, S. Jabbari, M. Kearns, A. Roth, "Fairness in criminal justice risk assessments: The state of the art," *Sociological Methods & Research*, p. 0049124118782533, 2018.

[2] S. Corbett-Davies, E. Pierson, A. Feller, S. Goel, A. Huq, "Algorithmic decision making and the cost of fairness," in *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2017, pp. 797–806.

[3] C. Dwork, M. Hardt, T. Pitassi, O. Reingold, R. Zemel, "Fairness through awareness," in *Proceedings of the 3rd innovations in theoretical computer science conference*, 2012, pp. 214–226.

[4] J. Kleinberg, S. Mullainathan, M. Raghavan, "Inherent trade-offs in the fair determination of risk scores," *arXiv preprint arXiv:1609.05807*, 2016.

[5] M. J. Kusner, J. Loftus, C. Russell, R. Silva, "Counterfactual Fairness," in *Advances in Neural Information Processing Systems 30*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, R. Garnett Eds., Curran Associates, Inc., 2017, pp. 4066–4076.

[6] S. Verma, J. Rubin, "Fairness definitions explained," in *Proceedings of the International Workshop on Software Fairness*, FairWare '18, Gothenburg, Sweden, May 2018, pp. 1–7, As-sociation for Computing Machinery.

[7] Z. Obermeyer, B. Powers, C. Vogeli, S. Mullainathan, "Dissecting racial bias in an algorithm used to manage the health of populations," *Science*, vol. 366, pp. 447–453, Oct. 2019, doi: 10.1126/science.aax2342.

[8] R. Nabi, I. Shpitser, "Fair Inference on Outcomes," in T*hirty-Second AAAI Conference on Artificial Intelligence*, Apr. 2018.

[9] A. Chouldechova, "Fair prediction with disparate impact: A study of bias in recidivism prediction instruments," *Big data*, vol. 5, no. 2, pp. 153–163, 2017.

[10] A. Z. Jacobs, H. Wallach, "Measurement and fairness," *arXiv preprint arXiv:1912.05511*, 2019.

[11] A. P. Dawid, A. M. Skene, "Maximum likelihood estimation of observer error-rates using the em algorithm," *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, vol. 28, no. 1, pp. 20–28, 1979.

[12] V. C. Raykar, S. Yu, L. H. Zhao, G. H. Valadez, C. Florin, L. Bogoni, L. Moy, "Learning from crowds.," *Journal of Machine Learning Research*, vol. 11, no. 4, 2010.

[13] D. Borsboom, "When does measurement invariance matter?," *Medical care*, vol. 44, no. 11, pp. S176–S181, 2006.

[14] J. Pearl, *Causality models, reasoning, and inference.* Cambridge: Cambridge University Press, 2013. OCLC: 956314447.

[15] P. Spirtes, C. N. Glymour, R. Scheines, D. Heckerman, *Causation, prediction, and search.* MIT press, 2000.

[16] T. B. Brakenhoff, M. Mitroiu, R. H. Keogh, K. G. Moons, R. H. Groenwold, M. van Smeden, "Measurement error is often neglected in medical literature: a systematic review," *Journal of clinical epidemiology*, vol. 98, pp. 89–97, 2018.

[17] D. Borsboom, "Latent variable theory," *Measurement: Interdisciplinary Research and Perspectives*, vol. 6, no. 1-2, pp. 25–53, 2008, doi: 10.1080/15366360802035497.

[18] N. Kilbertus, M. R. Carulla, G. Parascandolo, M. Hardt, D. Janzing, B. Schölkopf, "Avoiding discrimination through causal reasoning," in *Advances in Neural Information Processing Systems*, 2017, pp. 656–666.

[19] W. A. Fuller, *Measurement error models*, vol. 305. John Wiley & Sons, 2009.
[20] H. M. Blalock, A. B. Blalock, "Methodology in social research," 1968.
[21] A. L. McCutcheon, *Latent class analysis*. No. 64, Sage, 1987.
[22] G. Rasch, *Probabilistic models for some intelligence and attainment tests*. ERIC, 1993.
[23] G. J. McLachlan, K. E. Basford, *Mixture models: Inference and applications to clustering*, vol. 38. M. Dekker New York, 1988.
[24] F. M. Lord, *Applications of item response theory to practical testing problems*. Routledge, 2012.
[25] K. A. Bollen, *Structural equations with latent variables*. Wiley series in probability and mathematical statistics Applied probability and statistics, New York, NY Chichester Brisbane Toronto Singapore: Wiley, 1989. OCLC: 18834634.
[26] A. Skrondal, S. Rabe-Hesketh, *Generalized latent variable modeling: Multilevel, longitudinal, and structural equation models*. CRC Press, 2004.
[27] K. G. Jöreskog, "Testing structural equation models," *Sage focus editions*, vol. 154, pp. 294–294, 1993.
[28] G. J. Mellenbergh, "Item bias and item response theory," *International Journal of Educational Research*, vol. 13, pp. 127–143, Jan. 1989, doi: 10.1016/0883-0355(89)90002-5.
[29] P. W. Holland, H. Wainer, *Differential Item Functioning*. New York: Routledge, 1993.
[30] N. Schmitt, G. Kuljanin, "Measurement invariance: Review of practice and implications," *Human resource management review*, vol. 18, no. 4, pp. 210–222, 2008.
[31] P. Flore, "Stereotype threat and differential item functioning: A critical assessment," 2018.
[32] J.-B. E. Steenkamp, H. Baumgartner, "Assessing measure-ment invariance in cross-national consumer research," *Journal of consumer research*, vol. 25, no. 1, pp. 78–90, 1998.
[33] R. J. Vandenberg, C. E. Lance, "A review and synthesis of the measurement invariance literature: Suggestions, practices, and recommendations for organizational research," *Organizational research methods*, vol. 3, no. 1, pp. 4–70, 2000.
[34] B. M. Byrne, *Structural equation modeling with Mplus: Basic concepts, applications, and programming*. routledge, 2013.
[35] Y. Rosseel, "Lavaan: An r package for structural equation modeling and more. version 0.5–12 (beta)," *Journal of statistical software*, vol. 48, no. 2, pp. 1–36, 2012.
[36] K. G. Jöreskog, A. S. Goldberger, "Estimation of a model with multiple indicators and multiple causes of a single latent variable," *Journal of the American Statistical Association*, vol. 70, no. 351a, pp. 631–639, 1975.
[37] R. Tibshirani, M. Saunders, S. Rosset, J. Zhu, K. Knight, "Sparsity and smoothness via the fused lasso," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 67, no. 1, pp. 91–108, 2005.
[38] T. Hastie, R. Tibshirani, J. Friedman, *The elements of statistical learning: data mining, inference, and prediction*. Springer Science & Business Media, 2009.

Laura Boeschoten

Laura Boeschoten is a postdoctoral researcher at the Methodology and Statistics department of Utrecht University. She is particularly interested in measurement error, missing data and latent variable models. Her PhD was on a joint project between Tilburg University and Statistics Netherlands where she developed a new methodology that estimates and corrects for measurement error in combined survey-register data-sets (Tilburg University, The Netherlands, 2019) and she currently continues this line of research by further investigating how this methodology can be implemented in the production if various official statistics. Other lines of research focus on developing an infrastructure that enables the use of so-called 'Data Download Packages' for scientific research and on investigating the validity and reliability of measurements obtained from these 'Data Download Packages'.

Erik-Jan van Kesteren

Erik-Jan van Kesteren is an assistant professor of human data science at Utrecht University, the Netherlands. His educational background is mainly in social science and statistics. In between studies he worked in a data management team at a large company in the Netherlands. His PhD research was on extending latent variable models for modern data problems, with a focus on regularization, optimization, and software (Utrecht University, The Netherlands, 2021). Currently, Erik-Jan works with a small social data science team at ODISSEI (odissei-data.nl), helping social scientists with their computational and data problems.

Ayoub Bagheri

Ayoub Bagheri is an assistant professor in applied data science at the Methodology and Statistics department of Utrecht University. The focus of his academic career has been to develop intelligent systems for improving health, education, and social sciences by mining big data, especially text data. His current research in terests include machine learning, text mining, and natural language processing. As part of the Human Data Science group, he works on several projects in the domain of applied data science for health and social sciences. He is also part of the organization team of the special interest group Text Mining of the focus area applied data science in Utrecht University.

Daniel Oberski

Daniel Oberski holds a joint appointment as associate professor of data science methodology at Utrecht University, department of Methodology Statistics, and at the University Medical Center Utrecht (UMCU), department of Biostatistics. His work focuses on latent variable modeling and data science applications in the social, behavioral, and biomedical sciences. He leads the social data science team at the national research infrastructure for the social sciences in the Netherlands, ODISSEI. He is also lead data scientist of UMCU's "digital health" program, which works to implement data science in clinical care at the hospital.

International Journal of Interactive Multimedia and Artificial Intelligence, Vol. 6, Nº5

Attesting Digital Discrimination Using Norms

Natalia Criado, Xavier Ferrer, Jose M. Such *

Department of Informatics, King's College London (United Kingdom)

Received 19 October 2020 | Accepted 20 December 2020 | Published 22 February 2021

Abstract

More and more decisions are delegated to Machine Learning (ML) and automatic decision systems recently. Despite initial misconceptions considering these systems unbiased and fair, recent cases such as racist algorithms being used to inform parole decisions in the US, low-income neighborhood's targeted with high-interest loans and low credit scores, and women being undervalued by online marketing, fueled public distrust in machine learning. This poses a significant challenge to the adoption of ML by companies or public sector organisations, despite ML having the potential to lead to significant reductions in cost and more efficient decisions, and is motivating research in the area of algorithmic fairness and fair ML. Much of that research is aimed at providing detailed statistics, metrics and algorithms which are difficult to interpret and use by someone without technical skills. This paper tries to bridge the gap between lay users and fairness metrics by using simpler notions and concepts to represent and reason about digital discrimination. In particular, we use norms as an abstraction to communicate situations that may lead to algorithms committing discrimination. In particular, we formalise non-discrimination norms in the context of ML systems and propose an algorithm to attest whether ML systems violate these norms.

Keywords

Discrimination, Attestation, Norms, Machine Learning.

DOI: 10.9781/ijimai.2021.02.008

I. Introduction

Digital discrimination is a form of discrimination in which algorithms, often based on AI techniques (such as machine learning), make automatic decisions that result in users treated unethically, unfairly, or just differently based on their personal data [1] such as gender, ethnicity, or religion, among others. It is an important and relevant problem to tackle as more and more tasks are being delegated to automated decision-making systems which embed these AI techniques on the background, and that can be found in mobile phones, computer systems, and even searching and recommendation engines [2]. For instance, a relevant example among many is that some firms in the UK use automated decision-making systems to screen or hire candidates on[1].

Frequently the users of such machine learning (ML) systems are not technical experts and cannot assess by themselves if these algorithms are discriminatory. For example, many public organizations would like to reduce operational costs and delegate some decisions to algorithms, but at the same time need some guarantees about the ML systems not breaking anti-discrimination laws. Our approach has been precisely designed to allow non-technical users to determine if ML systems are potentially discriminatory and to make explicit under which assumptions the systems are discrimination free.

This paper is organised as follows: Section II introduces background knowledge on discrimination legislation; Section III introduces our formalization of non-discrimination norms in the context of ML systems; Section IV contains our attesting algorithm; Section V illustrates the performance of our algorithm in three case studies; Section VI contains related work; and Section VII contains a discussion of the paper contribution.

II. Background

Legislation about discrimination in general, not necessarily just about digital discrimination, is varied and extensive. National and international governments and organisations have legislation that specifically prohibits discrimination; e.g., the European Convention for the Protection of Human Rights. Most of this legislation defines a non-exhaustive list of criteria or protected attributes, e.g., race, gender, sexual orientation, based on which discrimination is prohibited [1]. This means that, from a legal perspective, discrimination are usually the actions, procedures, etc., that disadvantage citizens based on their membership of particular groups defined by those protected attributes.

Legislation about discrimination typically distinguishes between two main types of discrimination [3]:

1. Direct discrimination (disparate treatment). This type of discrimination considers the situation in which an individual is treated differently because of their protected attributes. This means that different social groups defined by their protected attributes are treated differently, with one or more groups being disadvantaged by these differences. One example of (explicit) direct discrimination would be a policy that does not allow to hire candidates who are foreign. We differentiate two types of direct discrimination:
   a) Explicit, as in the previous example, in which members of a particular social group, i.e., foreign people, are explicitly

[1] http://www.bbc.co.uk/news/business-36129046

* Corresponding author.

E-mail addresses: natalia.criado@kcl.ac.uk (N. Criado), xavier.ferrer_aran@kcl.ac.uk (Xavier Ferrer), jose.such@kcl.ac.uk (Jose M. Such).

disadvantaged by a decision, i.e., foreign candidates will be treated differently and not considered for hiring.

b) Implicit, in which the discriminated group is not explicitly mentioned or considered. Coming back to the previous example, the same company could replace the hiring policy with a new policy of not hiring candidates with qualifications from international institutions. The new policy would not explicitly consider the relevant social group (foreign people), yet it may accomplish the same exact objective, because foreign people are more likely to have qualifications from another country.

2. Indirect discrimination (disparate impact). This type of discrimination, also named disparate impact, considers situations in which a neutral treatment has a disproportionately negative effect on the members of a particular group defined by their protected attributes. This is considered discrimination, even if: i) there is no intention to discriminate against that particular group, and ii) there is not any unconscious prejudice motivating the act. For example, a company having the policy to not approve any part-time work requests, without having any reason to reject such requests, may have a disproportionate effect on female employees, when compared to their male counterparts. In this case, the company may not have an intention to discriminate against female employees, but the policy may disproportionately disadvantage them.

III. Digital Discrimination Normative Model

The term digital discrimination refers to those direct or indirect discriminatory acts that are based on the automatic decisions made by an ML system. In this section, we formalise the notion of digital discrimination norms accounting for the different types of discrimination introduced in the previous section: explicit, implicit, and indirect discrimination.

An ML system can be defined by a set of input features $\mathcal{I} = \{I_1, ... I_m\}$, where each feature I_i takes values from a discrete domain D_{I_i}; and an output feature O, which also takes values from a discrete domain D_O.[2] Note that, in this paper, we are interested in ML systems where the input may contain, directly or indirectly, personal information about individuals in order to attest discrimination. For this reason, the set of *protected* features is also defined; i.e., $\mathcal{P} = \{P_1, .., P_n\}$, where each protected feature $P_i \in \mathcal{P}$ takes values from a discrete domain D_{P_i}. It may be that protected features are part of the input directly used by an ML system, but it is not necessary, e.g., as we will see later, protected features could be strongly *associated* with the inputs even if not directly used as inputs.

The decisions of an ML system can be represented as a dataset *DS* formed by tuples $(p_1, ... p_n, i_1, ... i_m, o)$, where each tuple represents a previous decision made by the ML system about a particular individual with protected attributes $p_1, ... p_n$, input attributes $i_1, ... i_m$, and algorithm outcome[3] o. In particular, each $p_i \in D_{P_i}$, $i_i \in D_{I_i}$ and $o \in D_O$.

In the following, we provide a formalization of non-discrimination norms for ML systems and define how domain knowledge can be represented using norm exceptions. These normative notions are illustrated with an example.

A. Digital Discrimination Norms

As aforementioned, in the legislation around the world, we find the following types of discrimination: direct (also known as disparate treatment), which further classifies into explicit and implicit; and indirect (disparate impact) [5]. Next, we contextualise these notions in the context of digital discrimination and we formally represent them as computational norms using deontic logic[4]. These deontic norms express anti discrimination rules of behaviour for ML systems using concepts and terminology easily understood by non-technical users.

1. Direct Discrimination

Direct Discrimination is the unequal behavior toward someone because of a protected characteristic. We consider the two types of direct discrimination identified in previous literature, as discussed in Section II: explicit and implicit discrimination.

Explicit Discrimination. In terms of ML systems, this type of discrimination is equivalent to having some of the protected attributes considered in the systems' input. Norms preventing explicit discrimination can be formalised as prohibitions to include protected attributes in the input of the system as follows:

$$\forall P_i \in \mathcal{P} : \mathbf{F}(P_i \in \mathcal{I})$$

The set of all explicit discrimination norms is denoted by N_E and has a size of $|\mathcal{P}|$.

Implicit Discrimination. This type of discrimination can be formalised as a situation where the values of a set of input attributes of an ML system correlate with the value of one or more of the protected attributes. Therefore, norms preventing implicit discrimination can be formalised as follows:

$$\forall P_i \in \mathcal{P} : \mathbf{F}(P_i \text{ is a function of } \mathcal{I})$$

Note that P_i is a function of $\mathcal{I}$ is defined in terms of a process to detect associations, correlations or dependencies between attributes (Section VI provides more details about techniques and metrics that can be used for this). Also note that the set of all implicit discrimination norms is denoted by N_I and has a size of $|\mathcal{P}|$.

Remark 1. If an explicit discrimination norm for a protected feature P_i is violated, then the implicit discrimination norm for P_i is also violated. The inverse inference, however, does not hold.

2. Indirect Discrimination

Indirect Discrimination (disparate impact) refers to decisions that adversely affect one group of people of a protected characteristic more than another. This equals to state that for a particular protected attribute value $p \in D_{P_i}$, the probability of a given outcome $o \in D_o$ is x times lower than that of the values of the same protected attribute P with the highest probability. Formally, we can define the norm prohibiting indirect discrimination as:

$$\forall P_i \in \mathcal{P}, \forall p \in P_i, \forall o \in D_o : \mathbf{F}\,(P_i \downarrow^p_o)$$

where $P_i \downarrow^p_o$ denotes:

$$Pr(O = o | P_i = p) < x \times \max_{\forall p' \in D_{P_i} \setminus \{p\}} Pr(O = o | P_i = p')$$

with $Pr(O = o | P_i = p)$ standing for the probability that the outcome o is given to an individual with protected attribute p. Therefore, the norm states that it is forbidden that for a given group, characterised by having p as the value for the protected feature P_i, the probability of an outcome o is x times lower than the probability of the same outcome o for all the alternative groups, which are characterised by having the other values for P_i (i.e., $D_{P_i} \setminus \{p\}$).

Note that different methods can be used to estimate this probability.

[2] For simplicity we assume domains are discrete, but this is without loss of generality, as any continuous domain can be discretized.

[3] Note that it is possible to consider discrimination in an algorithm by considering the ground-truth labels as well. See Appendix B for more details about this particular type of discrimination, which in some cases is known as disparate mistreatment [4].

[4] For simplicity, we don't consider compound discrimination in the main part of this paper. For a definition of compound discrimination norms see Appendix A.

In Section VI, we provide a review of the different techniques that may be used. Also note that the value $x \in [0, 1]$ is a constant representing the extent of the disproportion allowed in a particular domain[5].

The set of all disparate impact norms is denoted by N_D and has a size of $|\mathcal{P}| \times \overline{D_\mathcal{P}} \times |Do|$ where $\overline{D_\mathcal{P}}$ denotes the average number of values belonging to the domain of protected attributes. That is, there is one disparate impact norm per each group, characterised by having a particular value for a given protected feature, and each possible outcome.

3. Norm Violations

Based on the definitions above, the full set of anti-discrimination norms considered is represented as a collection denoted by $N = (N_E, N_I, N_D)$, where N_E, N_I, N_D are as defined above, representing norms against explicit, implicit and indirect discrimination.

Whenever any of the norms in N are violated, there may then be a case of discrimination. However, some of these violations could be considered inconsequential, as we describe next, or there may also be domain-dependent exceptions (as defined later on in Section B).

In this paper, we define inconsequential norm violations as those violations which can be considered trivial, since they have little effect on the decisions made by the ML system. Importantly, inconsequential violations are anyway worth considering, as they may be an indicator of bad practices (e.g., considering disability status of students in university admissions may be immoral even if that information is ultimately not influencing much the decision).

Remark 2. If an explicit discrimination norm for a protected feature P_i is violated and no indirect discrimination norm for P_i is violated, then the violation is inconsequential as the protected feature P_i is not affecting significantly the decision-making process. If an implicit discrimination norm for protected feature P_i is violated and no indirect discrimination norm for P_i is violated, then the violation is inconsequential as the protected feature P_i is not affecting significantly the decision-making process.

B. Norm Exceptions

The previous section formalises the general definition of anti-discrimination norms. In general, when these norms are violated there is a potential case of digital discrimination. However, there are domains in which the violation of these norms is justifiable, and hence not result in discrimination. To allow for such type of domain knowledge to be explicitly represented and accounted for, we use the notion of domain permission norms, which define exceptions to the general anti-discrimination norms.

1. Exceptions to Direct Discrimination Norms

Exception to Violate Explicit Norms. This refers to the cases where permission to use protected attributes in decision making may be justified. For example, legislation does not usually consider discriminatory to use religion as a criteria for hiring a religion teacher at a school. An explicit permission to use a protected attribute $P_i \in \mathcal{P}$ can be defined as follows:

$$\mathbf{P}(P_i \in I)$$

The set of all exceptions to explicit discrimination norms is denoted by E_E.

Exception to Violate Implicit Norms.This refers to the cases where permission to allow for correlations between a protected attribute and input attributes is justified. For example, for some particular jobs, (e.g., firefighters) the candidates may need to demonstrate physical strength, which is correlated with gender. In such cases, it may be lawful to consider the results of fitness tests in hiring decisions. This allowed correlation between a protected attribute $P_i \in \mathcal{P}$ and a subset of the input attributes $I \subset \mathcal{I}$ can be represented as a permission norm as follows:

$$\mathbf{P}(P_i \text{ is a function of } I)$$

The set of all exceptions to implicit discrimination norms is denoted by E_I.

Remark 3. An exception to an explicit discrimination norm about protected attribute P_i entails an exception for the implicit discrimination norm related to P_i and all input attributes. The inverse relationship does not hold.

2. Exceptions to Indirect Discrimination Norms

This refers to the cases where permission to treat different groups disparately may be explainable. For example, on average, women Uber drivers are paid less than men drivers [7], but that could be explained by factors such as driver experience, time and location of rides, etc. An exception to allow for a significant difference on an outcome $o \in D_o$ for a particular protected group $p \in D_{P_i}$ where $P_i \in \mathcal{P}$ can be formalised as follows:

$$\mathbf{P}(P_i \downarrow_o^p)$$

The set of all exceptions to indirect discrimination norms is denoted by E_D.

Remark 4. An exception to an explicit discrimination norm about protected attribute P_i does not entail an exception to any indirect discrimination norms for P_i. An exception to an implicit discrimination norm about protected attribute P_i does not entail an exception to any indirect discrimination norms for P_i.

There may be cases in which it is lawful to consider protected attributes in the decision-making process, either explicitly or implicitly, as long as that information is not used to disproportionately disadvantage the members of a certain group; e.g., positive discrimination practices allows the use of gender and race information to increase the number of employees from minority groups in a company or business, which are known to have been discriminated against in the past. In this case there is an exception to an explicit discrimination norm about gender and race, as long as that information is not used to adversely affect any group; e.g., gender information can be used by the ML system as long as all genders do not have disproportional probabilities to obtain the different outcomes.

Domain exceptions to discrimination norms are represented as a collection denoted by $E = (E_E, E_I, E_D)$.

C. Example: Credit Risk Assessment

To illustrate the different types of norms and exceptions let us consider an example of a decision making system that classifies individuals as high or low risk in a credit risk assessment scenario.

The attributes used to describe invididuals are:

$$\mathcal{I} = \{Age, Job, Salary\}$$

where and $Age \in \{[20, 30], [30, 40], ...\}$, $Job \in \{Unemployed, Unskilled, ...\}$, and $Salary \in \{[0, 20k], [20k, 30k], ...\}$. According to common discrimination law, protected attributes are defined as:

$$\mathcal{P} = \{Gender, Age\}$$

where $Gender \in \{Male, Female\}$. The output variable is:

$$O = Risk$$

where $Risk \in \{High, Low\}$.

[5] For example, the US *fourth-fifth rule* from the Equal Employment Opportunity Commission (1978) states a job selection rate for the protected group of less than 4/5 of the selection rate for the unprotected group [6].

In this example the following norms are generated considering protected attributes:

$$\mathbf{F}(Gender \in \mathcal{I}),\ \mathbf{F}(Age \in \mathcal{I})$$

$$\mathbf{F}(Gender \text{ is a funtion of } \mathcal{I}),\ \mathbf{F}(Age \text{ is a funtion of } \mathcal{I})$$

$$\mathbf{F}(Gender \downarrow^{Male}_{High}),\ \mathbf{F}(Gender \downarrow^{Male}_{Low}),$$

$$\mathbf{F}(Gender \downarrow^{Female}_{High}),\ \mathbf{F}(Gender \downarrow^{Female}_{Low}),$$

$$\mathbf{F}(Age \downarrow^{[20,30]}_{High}),\ \mathbf{F}(Age \downarrow^{[20,30]}_{Low}),\ ...$$

$$...,\ \mathbf{F}(Age \downarrow^{[70,80]}_{High}),\ \mathbf{F}(Age \in \downarrow^{[70,80]}_{Low})$$

In addition, in this example, there are also several exceptions to the norms as follows:

$$\mathbf{P}(Age \in \mathcal{I})$$

$$\mathbf{F}(Gender \text{ is a funtion of } \{Salary\}),$$

$$\mathbf{P}(Age \downarrow^{[20,30]}_{High}),\ \mathbf{P}(Age \downarrow^{[20,30]}_{Low}),\ ...$$

$$...,\ \mathbf{P}(Age \downarrow^{[70,80]}_{High}),\ \mathbf{P}(Age \in \downarrow^{[70,80]}_{Low})$$

In particular, it may be considered lawful to use age in the credit risk assessment, as it is common practice to use age to estimate health risks, insurance, unemployment rates, etc. By Remark 3, it is implicitly permitted that age is a function of input attributes. The pay gap phenomenon also explains a degree of correlation between salary and gender. In this case, however, the use of salary for credit risk assessment may be considered lawful (i.e., salary has not been used as a way to discriminate women, but as a way to determine the capability of individuals to pay a credit back). Finally, it is considered permitted to allow age to have a significant impact on credit risk assessment decisions and any age groups to be discriminated on this basis.

IV. Digital Discrimination Attesting Process

The digital discrimination attesting process, which is depicted with all its steps in Fig. 1, takes as input a decision dataset and the domain exceptions defined by the user, and it returns a discrimination report with information about any potential discrimination cases (i.e., the minimal list of norm violations) and the assumptions made in the attesting process (i.e., the list of exceptions provided by the user and the allowed disproportion ratio)[6].

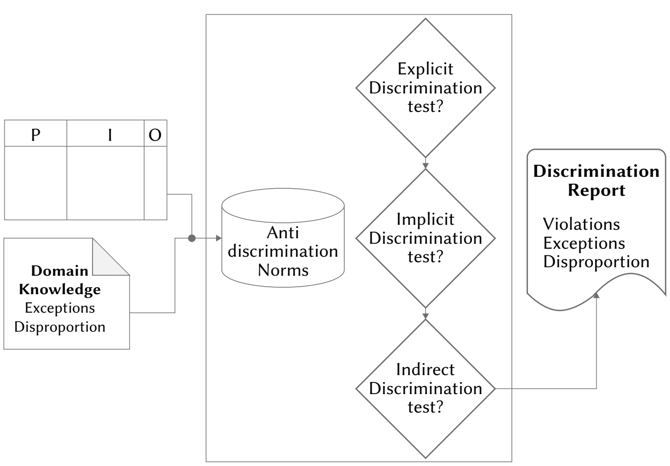


Fig. 1. Overview of the Attesting Process.

[6] Note the purpose of our paper is to allow non-technical users to attest whether ML systems discriminate. We do not focus on the mitigation of discrimination when found. For examples of the growing research field on mitigating discrimination see [8]-[11].

The attesting algorithm (see Algorithm 1) starts by generating the list of discrimination norms based on the input, protected and output attributes (line 7), and then, it checks compliance with the different types of norms.

Explicit direct discrimination. The algorithm starts by checking compliance with explicit direct discrimination norms (lines 8-15). In particular, for each protected attribute it checks if there is an permission norm allowing the ML system to use it as input (line 9). If not, it checks if the explicit discrimination norm is violated, which is the same as checking for set membership. For each explicit norm that is violated, a new inconsequential violation is added (line 11); later on the algorithm will confirm if this violation is actually inconsequential or not. Finally, the implicit norm related to that protected attribute is removed (line 12). Note that our goal is to produce the minimal set of violations and, by Remark 1, the explicit norm is more general.

Implicit direct discrimination. The algorithm checks for implicit direct discrimination in lines 16-24. For each implicit norm, the algorithm first searches for exceptions to explicit norms related to the same protected attribute (as stated in Remark3). If no exceptions are found, the algorithm uses the dataset *DS* as a representative sample to check if the norm is violated.[7] An implicit norm is violated when there is a subset of input attributes determining the value of a protected attribute. If the norm is violated, the algorithm checks for a permission norm allowing for that particular violation. In particular, the algorithm checks if there is an exception for that set of input attributes, or a subset of it, determining the protected attribute (lines 18-19). Again, if the norm is finally considered to be violated, a new inconsequential violation is created (line 21). As before, the algorithm will determine later on if that violation is actually inconsequential or not.

Indirect discrimination. The algorithm checks for indirect discrimination in lines 25-39. The algorithm starts by checking for each indirect norm whether there is an exception to it (line 26). If there is not, it checks if the indirect norm is violated (line 27). To determine if an indirect norm is violated the dataset *DS* is used as a representative sample to calculate probabilities associated to each outcome and protected group[8]. For each indirect norm that is violated, a new violation is created (line 28). As stated in Remark 2, if there are inconsequential violations of explicit norms related to that protected attribute, these are converted into consequential ones (lines 29-32). The violation of the indirect norm associated to a protected attribute demonstrates that decisions are having a disproportionate impact based on that protected attribute. Similarly, if there are inconsequential violations of implicit norms related to that protected attribute, these are converted into consequential ones (lines 33-36).

Discrimination Report. Finally, the algorithm outputs the list of inconsequential and consequential violations found. Note that the discrimination report will contain not only the information about norm violations (if any), but also the information about the exceptions considered in its analysis and the level of allowed disproportion specified by the user.

Complexity. The complexity of the algorithm to attest digital discrimination is determined by the size of the biggest norm set (or exception set). In this case, the complexity is given by $\mathbf{O}(|\mathcal{P}| \times \overline{D_{\mathcal{P}}} \times |Do|)$. This assumes that the norm violation checks are performed offline and can be retrieved in constant time. Section VI discusses different methods to check compliance of implicit and indirect norms (note checking compliance of explicit norms equates to checking set membership).

[7] Different statistical methods can be used to determine if there is a correlation between input attributes and protected attributes. Refer to Section VI for more details.

[8] Different statistical methods can be used to determine the probability of obtaining an outcome value for a particular protected group. Refer to Section VI for more details.

Algorithm 1. Digital Discrimiation Attesting

DiscriminationAttesting $(\mathcal{P}, \mathcal{I}, O, DS, E, x)$
  inputs: A set of protect attributes $\mathcal{P}$
    A set of input attributes $\mathcal{I}$
    An output attribute O
    A dataset DS
    A collection of exceptions (E_E, E_I, E_D)
    $x \in [0, 1]$ a constant representing the disproportion allowed
  output: A collection of violated norms (V_E, V_I, V_D)
    A collection of norms that have been violated inconsequentially (I_D, I_I)
  $V_E \leftarrow \varnothing$
  $V_I \leftarrow \varnothing$
  $V_D \leftarrow \varnothing$
  $I_E \leftarrow \varnothing$
  $I_I \leftarrow \varnothing$
  $(N_E, N_I, N_D) \leftarrow GenerateNorms(\mathcal{P}, \mathcal{I}, O)$
  `// Attesting Explicit Discrimination`
  foreach $\mathbf{F}(P_i \in \mathcal{I}) \in N_E$ **do**
    if $\nexists \mathbf{P}(P_i \in \mathcal{I}) \in E_E$ **then**
      if $P_i \in \mathcal{I}$ **then**
        $I_E \leftarrow I_E \cup \{\mathbf{F}(P_i \in \mathcal{I})\}$
        $N_I \leftarrow N_I \backslash \{\mathbf{F}(P_i \text{ is a function of } \mathcal{I})\}$
      end
    end
  end
  `// Attesting Implicit Discrimination`
  foreach $\mathbf{F}(P_i \textit{ is a function of } \mathcal{I}) \in N_I$ **do**
    if $\nexists \mathbf{P}(P_i \in \mathcal{I}) \in E_E$ **then**
      foreach $I \subseteq \mathcal{I} : I$ *is the minimal set such that* P_i *is a function of* I **do**
        if $\nexists \mathbf{P}(P_i \textit{ is a function of } I') : I \subseteq I'$ **then**
          $I_I \leftarrow I_I \cup \{\mathbf{F}(P_i \text{ is a function of } \mathcal{I})\}$
        end
      end
    end
  end
  `// Attesting Indirect Discrimination`
  foreach $\mathbf{F}(P_i \downarrow_o^p) \in N_D$ **do**
    if $\neg \exists \mathbf{P}(P_i \downarrow_o^p) \in E_D$ **then**
      if $\exists p' \in D_{P_i} : \frac{Pr(O=o|P_i=p)}{Pr(O=o|P_i=p')} < x$ **then**
        $V_D \leftarrow V_D \cup \{\mathbf{F}(P_i \downarrow_o^p)\}$
        if $\mathbf{F}(P_i \in \mathcal{I}) \in I_E$ **then**
          $I_E \leftarrow I_E \backslash \{\mathbf{F}(P_i \in \mathcal{I})\}$
          $V_E \leftarrow V_E \cup \{\mathbf{F}(P_i \in \mathcal{I})\}$
        end
        if $\mathbf{F}(P_i \textit{ is a function of } \mathcal{I}) \in I_I$ **then**
          $I_I \leftarrow I_I \backslash \{\mathbf{F}(P_i \text{ is a function of } \mathcal{I})\}$
          $V_I \leftarrow V_I \cup \{\mathbf{F}(P_i \text{ is a function of } \mathcal{I})\}$
        end
      end
    end
  end
  $V \leftarrow (V_E, V_I, V_D)$
  $I \leftarrow (I_E, I_I)$
  return *V,I*

V. Case Studies

In this section, we illustrate the performance of our digital discrimination attesting algorithm using three well-known datasets: the German dataset[9], the Adult dataset[10], and the COMPAS Recidivism dataset[11].

In our implementation[12], we have used the *sklearn* library for normalised mutual information [12] to detect violations of implicit discrimination norms. The normalised mutual information (NMI) is a measure of the mutual dependence between the two variables that quantifies the "amount of information" obtained about one random variable through observing the other random variable. The NMI returns 0 when there is no mutual information between the variables tested, and 1 when there exist a perfect correlation. In the implementation, the minimum coefficient for mutual information can be configured; we used a minimum threshold of 0.6 in the experiments below as indicative of a strong correlation between input and protected attributes. To detect indirect discrimination we have set to 0.8 the allowed disproportion ratio, inspired by the US *fourth-fifth* rule from the Equal Employment Opportunity Commission (1978), a threshold commonly used to detect disparate impact in domains like employee selection procedures[13]; and we have calculated the probabilities using the frequencies in the dataset as a representative sample. Also, due to the small size of the datasets used in the case studies, we have used the Chi-Squared Test [13] to determine those violations of indirect discrimination norms that are statistically significant (p-value < 0.05). To discretise numeric values, we have used quantile discretisation, which is a well-known method for discretising continuous variables in ML [14].

A. Adult Dataset

The Adult dataset uses 14 attributes to determine if a given person makes over 50K a year. The attributes include education, work class, age, sex, race, and occupation, among others. The dataset contains 48842 instances.

Let us assume that the gender, age, native country and race are protected and that the other attributes are the inputs of a ML system.

$$\mathcal{I} = \{workclass, education, education\_num, occupation, capital\_gain, capital\_loss, hours\_per\_week, fnlwgt\}$$

Note attribute *education_num* represents the number of education years, and *fnlwgt* represents the number of people the census believes the entry represents.

$$\mathcal{P} = \{age, gender, native\_country, relationship, marital\_status, race\}$$

$$O = income$$

where $income = \{<=50k, >50k\}$. In this case age is related to experience and seniority so it is considered lawful to use age to discriminate:

$$\mathbf{P}(age \downarrow_{<=50k}^{[0,16)}), \mathbf{P}(age \downarrow_{>50k}^{[0,16)}),$$
$$...$$
$$\mathbf{P}(age \downarrow_{<=50k}^{[75,99)}), \mathbf{P}(age \downarrow_{>50k}^{[75,99)})$$

After executing our algorithm several violations of indirect discrimination norms are detected. For example:

$$\mathbf{F}\ gender \downarrow_{>50k}^{female}$$

$$\mathbf{F}\ race \downarrow_{>50k}^{black}$$

[9] https://archive.ics.uci.edu/ml/datasets/statlog+(german+credit+data)

[10] https://archive.ics.uci.edu/ml/datasets/adult

[11] https://github.com/propublica/compas-analysis/

[12] Available on Github at https://github.com/xfold/NormativeApproachToDiscrimination

[13] http://www.uniformguidelines.com

$$\mathbf{F}\ native\_country \downarrow^{Nicaragua}_{>50k}$$

$$\mathbf{F}\ marital\_status \downarrow^{Married-civ-spouse}_{<=50k}$$

The first violation above indicates that females have a disproportionate lower probability of being classified as making more than 50k when compared with males. In particular, the dataset contains 21790 male instances out of which 6662 are classified as high income (i.e., the probability of income greater than 50k for male is 30%), whereas only 1179 female records out of 10771 are classified as high income (i.e., the probability of income greater than 50k for female is 11%). In this case 11% < 0.8 × 30% and it is considered disproportionate. The other violations above indicate that black people and nicaraguans have a disproportionate lower probability of being classified as making more than 50k when compared with other groups, in accordance with previous reports of discrimination in the dataset [15]. On the contrary, married people are significantly less likely of being classified as making less than 50k. Found violations are associated with particular values of gender, native country, relationship and marital-status attributes. This indicates that the decision making process may have a disparate impact on people belonging to particular protected groups.

B. German Credit Dataset

The German dataset contains information about people who ask for a credit. Each person is classified as good or bad credit risks. This is the inspiration for the small example contained in section C. In particular, the full dataset uses 20 attributes to represent each person, which include information like age, employment status, gender and personal status of the applicant; and the duration, amount and purpose of the credit. The dataset contains 1000 instances.

Let's us assume an ML system where age, personal status and sex, and being a foreign worker are considered protected attributes, and the rest of the features in the German dataset are considered inputs:

$$\mathcal{I} = \{job, housing, savings, .., amount, duration, purpose\}$$

$$\mathcal{P} = \{age, personal\_status\_and\_sex, foreign\_worker\}$$

$$O = risk$$

where $risk = \{high, low\}$. In this case, it is considered lawful to use age to discriminate credit risks as people are less likely to repay credits as they become older, hence, we consider age as an exception:

$$\mathbf{P}(age \downarrow^{[0,16)}_{good}), \mathbf{P}(age \downarrow^{[0,16)}_{bad}),$$

$$...$$

$$\mathbf{P}(age \downarrow^{[75,99)}_{good}), \mathbf{P}(age \downarrow^{[75,99)}_{bad})$$

After executing our algorithm, the following violation is detected:

$$\mathbf{F}(foreign\_worker \downarrow^{yes}_{good})$$

The violation means that foreign workers have a disproportionate low probability of being considered a good credit risk.

C. COMPAS Recidivism Dataset

The COMPAS (Correctional Offender Management Profiling for Alternative Sanctions) algorithm is a popular commercial algorithm used by judges and parole officers for scoring criminal defendant's likelihood of re-offending (recidivism), and is increasingly being used in pretrial and sentencing. The dataset has been widely used to study automatic decision systems related with recidivism [16], and it was found to be strongly biased against blacks [17]. The original dataset contains 28 columns which correspond to the variables used by the COMPAS algorithm to make its predictions, including data regarding sex, ethnicity, and marital status (among others), together with the final assessment made by the algorithm, the estimated recidivism score.

In the analysis below, we focus on *pretrial* instances and assessments about the *risk of recidivism* and *risk of violence*, following the analysis performed by ProPublica[14], considering sex and ethnicity as protected variables:

$$\mathcal{I} = \{marital\_status, legal\_status, \dots\}$$

$$\mathcal{P} = \{sex, ethnicity\}$$

$$O = recidivism\_score$$

where $recidivism\_score = \{low, medium, high\}$. After executing our algorithm, the following violation is detected:

$$\mathbf{F}(ethnicity \downarrow^{African-American}_{low})$$

The violation means that African-Americans have a disproportionate low probability of being considered with a low recidivism score when compared with other sub-populations, coinciding with the results reported in [17]. The reported bias becomes especially apparent when comparing African-American with Caucasian ethnicities, with African-Americans being consistently tagged by the COMPAS algorithm with *higher* and *medium* recidivism scores way more frequently than the Caucasian sub-population.

VI. Related Work

Recent research has addressed the problem of discrimination and bias in machine learning. These novel tools are most of the time aimed at technical users capable of interpreting different statistical results, programming, etc. Our algorithm is, on the contrary, aimed at non-technical users (albeit they may be domain experts). The notion of norm and exception is a suitable abstraction to represent the results these statistical analysis to non-technical users. For example, IBM's AI Fairness 360 Open Source Toolkit[15] and Google's What-if-tool[16], are probably two of the most comprehensive toolkits offering a great choice of bias metrics. However, its intended audience are technical users with previous knowledge of machine learning and statistics. Indeed, there are a large number of fairness metrics that may be appropriate for a given application [5], [18]. Also it is difficult for non-technical users to represent domain knowledge in a way that it can be taken into account by the metrics.

Closely related to our work is [19], where the authors proposed to infer classification rules from a given dataset and to detect those classification rules that can cause direct and indirect discrimination. They also allow for domain knowledge, expressed as rules, to be taken into account. Despite the similarities with this work, our proposal has two additional, potential benefits: it doesn't assume that meaningful rules can be inferred, note that it may be impossible to infer rules from complex decision-making algorithms; and it hides to the user the complexities of the analysis process using the notions of norms and exceptions.

Implicit Discrimination. Tramèr et al. [20] developed a methodology and toolkit combining different metrics for discovering associations, or proxies, between attributes. In particular, they studied different metrics that can be used to analyse the relationship between protected attributes and input attributes such as the Pearson correlation, which only works for scalar attributes linearly related; and Mutual Information, which can be applied to categorical attributes.

Indirect Discrimination. There have been many different metrics proposed to measure indirect discrimination both in the raw data used for training as well as the decisions made by the systems.

[14] https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm

[15] https://aif360.mybluemix.net

[16] https://pair-code.github.io/what-if-tool/

We refer the reader to [21] for an extensive survey in the topic, which discusses various statistical measures that could be used to measure discrimination and algorithmic fairness. In particular, the authors classify the metrics into: statistical tests, which are used to compute and calculate whether there is discrimination in a dataset; absolute measures, which are used to calculate the magnitude of the discrimination present in a dataset; conditional measures, which are used to assess the weights and importance that protected attributes have in the differences between groups, and how they relate to other characteristics; and structural measures, which are used to identify discriminated individuals in the dataset. Next, we also give some more detailed examples of work on indirect discrimination.

In [22], the authors proposed metrics to assess to what degree input attributes influence the outputs of an automated decision-making system. Although that paper is not intended to detect indirect discrimination per se, the measures the authors of the paper propose have the potential to increase the transparency and explainablity of decisions, which, in turn, may increase user trust and provide useful information for the detection of discrimination [23]. Other works have also attempted to propose metrics to capture discrimination in particular applications of Machine Learning. For instance, one example is the work that has attempted to detect discrimination in the applications of ML to Natural Language processing [24]-[26]. In these works, the approach followed is to explore the relationships between the words learned by the ML model to detect whether particular words or meanings are more associated to particular individuals based on their personal characteristics.

In addition to the work on detecting discrimination, there is also work focusing on making ML models fairer to start with. For instance, in [27], they test for fairness based on a similarity measure between individuals. For fairness to hold, the distance between the distributions of outputs for individuals should at most be the distance between the two individuals as estimated by means of the similarity metric. In [28], the authors first gather human judgments about the different protected features in the context of two real-world scenarios using Amazon Mechanical Turk. Using the set of *human-assessed* protected features, they compare the accuracy of different classifiers to test the trade-off between process fairness and output accuracy. In [29], they assume fairness can be attested by means of a directed causal graph, in which attributes are presented as nodes joined by edges which, by means of equations, represent the relations between attributes. Finally, the set of violations presented in our approach could also be extended with recent advances in explainable AI. One example is the post-hoc approach of Local Interpretable Model-Agnostic Explanations (LIME), which makes use of adversarial learning to generate counterfactual explanations [30].

VII. Conclusion

Digital discrimination is becoming a significant problem as more decisions are delegated to ML systems. Indeed, recent legislation and citizen initiatives are demanding more transparency about the way in which decisions are made using their data. In response to that, several metrics and tools have been proposed to analyse biases in ML systems. However, these tools often require expert ML or statistical knowledge that many users of ML systems do not necessarily possess.

In this paper, we proposed to use normative notions as an abstraction that may be more easily understood by non-technical users; simplifying the representation of the potential discrimination risks and the input of domain knowledge. Our digital discrimination attesting algorithm not only checks if ML systems are potentially discriminatory but also makes explicit under which assumptions these systems are discrimination free.

As future work, we plan to: i) investigate different metrics to be used in the attesting algorithm and to identify the most usable ones; ii) conduct user studies to further refine the way in which norms could be accessed and influenced by non-technical users to help them understand discrimination risks; and iii) explore interfaces to allow non-technical users to easily introduce exceptions and explanations to communicate the algorithm outputs to these users.

Appendix

A. Compound Discrimination

Compound discrimination is discrimination based on a combination of protected attributes. In that case of compound discrimination the previous discrimination norms are rewritten as follows:

- Direct.
  - Explicit. There is no need to change the definition of explicit discrimination norms to account for compound discrimination, since the prohibition to include a set of protected attributes in the input can be represented by a set of explicit norms referring to each individual protected attribute.
  - Implicit. There is no need to change the definition of implicit discrimination norms to account for compound discrimination, since the prohibition to have a set of protected attributes as a function of input attributes can be represented by a set of implicit norms referring to each individual protected attribute.
- Indirect (disparate impact). In this case the norms need to represent that for a particular combination of protected attribute values $p_1, ..., p_k$, where each $p_i \in P_i$; the probability of a given outcome $o \in D_o$ is x times lower than for values of the same protected attributes with the highest probability:

$$\forall \{P_1, ..., P_k\} \subseteq \mathcal{P}, (p_1, ..., p_k) \in D_1 \times ... \times D_{Pk}, o \in D_o :$$
$$\mathrm{F}(\{P_1, ..., P_k\} \downarrow_o^{(p_1,...,p_k)})$$

where $\{P_1, ..., P_k\} \downarrow_o^{(p_1,...,p_k)}$ denotes:

$$Pr(O = o \mid P_1 = p_1, ..., P_k = p_k) < x \times \max_{\forall \{(p'_1, ..., p'_k)\} \in D_{P_1} \times ... \times D_{P_k}} Pr(O = o \mid P_1 = p'_1, ..., P_k = p'_k)$$

B. Discrimination in Classification Process

In this paper we have focused on digital discrimination; i.e., discriminatory acts facilitated by the automatic decisions made by a ML system. However, it is possible to consider the discrimination in the algorithm itself. This is also know as disparate mistreatment [4]. In those cases it is necessary consider not only the outcome of the algorithm but also the ground-truth labels for the individuals, denoted by G. In those cases, it could be possible to formalise that for no particular value of a protected attribute the ML system can perform significantly worse than for the others groups. This equals to state that for a particular protected attribute value $p \in D_{P_i}$, the probability of the ML assigning the correct outcome ($O = g$) is x times lower than that of the values of the same protected attribute P with the highest probability. Formally, we can define the norm prohibiting disparate treatment as:

$$\forall P_i \in \mathcal{P}, \mathrm{p} \in D_{P_i}, g \in D_G : \mathbf{F}\, (P_i \uparrow_g^p)$$

where $P_i \uparrow_g^p$ represents:

$$Pr(O = g|P_i = p, G = g) < x \times \max_{\forall p' \in D_{P_i}} Pr\,(O = g \mid P_i = p', G = g)$$

$Pr(O = g|P_i = p, G = g)$ stands for probability that the algorithm outcome O is equal to the ground-truth label g for an individual with protected attribute $P_i = p$.

Acknowledgment

The research reported in this article was funded by the Engineering and Physical Sciences Research Council (EPSRC) under grant EP/R033188/1. This research is part of the cross-disciplinary project Discovering and Attesting Digital Discrimination (DADD) – visit our project website for further details: https://dadd-project.org.

References

[1] N. Criado, J. Such, "Digital discrimination," in *Algorithmic Regulation*, Oxford University Press, 2019.

[2] C. O'neil, *Weapons of math destruction: How big data increases inequality and threatens democracy*. Broadway Books, 2016.

[3] A. Altman, "Discrimination," 2011.

[4] M. B. Zafar, I. Valera, M. Gomez Rodriguez, K. P. Gummadi, "Fairness beyond disparate treatment & disparate impact: Learning classification without disparate mistreatment," in *Proceedings of the 26th international conference on world wide web*, 2017, pp. 1171–1180.

[5] S. Verma, J. Rubin, "Fairness definitions explained," in *2018 IEEE/ACM International Workshop on Software Fairness (FairWare)*, 2018, pp. 1–7, IEEE.

[6] S. Barocas, A. Selbst, "Big Data's Disparate Impact," *California law review*, vol. 104, no. 1, pp. 671–729, 2016, doi: http://dx.doi.org/10.15779/Z38BG31.

[7] C. Cook, R. Diamond, J. Hall, J. A. List, P. Oyer, "The gender earnings gap in the gig economy: Evidence from over a million rideshare drivers," *National Bureau of Economic Research*, 2018.

[8] S. Hajian, J. Domingo-Ferrer, "A methodology for direct and indirect discrimination prevention in data mining," *IEEE transactions on knowledge and data engineering*, vol. 25, no. 7, pp. 1445–1459, 2012.

[9] F. Calmon, D. Wei, B. Vinzamuri, K. N. Ramamurthy, K. R. Varshney, "Optimized preprocessing for discrimination prevention," in *Advances in Neural Information Processing Systems*, 2017, pp. 3992–4001.

[10] F. Kamiran, T. Calders, "Data preprocessing techniques for classification without discrimination," *Knowledge and Information Systems*, vol. 33, no. 1, pp. 1–33, 2012.

[11] M. Feldman, S. A. Friedler, J. Moeller, C. Scheidegger, S. Venkatasubramanian, "Certifying and removing disparate impact," in *proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining*, 2015, pp. 259–268.

[12] T. M. Cover, J. A. Thomas, *Elements of information theory*. John Wiley & Sons, 2012.

[13] W. G. Cochran, "The $\chi 2$ test of goodness of fit," *The Annals of Mathematical Statistics*, pp. 315–345, 1952.

[14] F. Freese, *Elementary statistical methods for foresters*. No. 317, US Department of Agriculture, 1967.

[15] R. K. Bellamy, K. Dey, M. Hind, S. C. Hoffman, S. Houde, K. Kannan, P. Lohia, J. Martino, S. Mehta, A. Mojsilović, et al., "Ai fairness 360: An extensible toolkit for detecting and mitigating algorithmic bias," *IBM Journal of Research and Development*, vol. 63, no. 4/5, pp. 4–1, 2019.

[16] A. Chouldechova, "Fair prediction with disparate impact: A study of bias in recidivism prediction instruments," *Big data*, vol. 5, no. 2, pp. 153–163, 2017.

[17] J. Angwin, J. Larson, S. Mattu, L. Kirchner, "Machine bias risk assessments in criminal sentencing," *ProPublica*, May, vol. 23, 2016.

[18] X. Ferrer, T. van Nuenen, J. M. Such, M. Coté, N. Criado, "Bias and Discrimination in AI: a cross-disciplinary perspective," *IEEE Technology and Society Magazine* (forthcoming), 2020.

[19] D. Pedreschi, S. Ruggieri, F. Turini, "Integrating induction and deduction for finding evidence of discrimination," in *Proceedings of the 12th International Conference on Artificial Intelligence and Law*, 2009, p. 157–166.

[20] F. Tramer, V. Atlidakis, R. Geambasu, D. Hsu, J. P. Hubaux, M. Humbert, A. Juels, H. Lin, "Fairtest: Discovering unwarranted associations in data-driven applications," in 2017 *IEEE European Symposium on Security and Privacy (EuroS&P)*, 2017, pp. 401–416, IEEE.

[21] I. Zliobaite, "A survey on measuring indirect discrimination in machine learning," *arXiv preprint arXiv:1511.00148*, 2015.

[22] A. Datta, S. Sen, Y. Zick, "Algorithmic transparency via quantitative input influence: Theory and experiments with learning systems," in *2016 IEEE symposium on security and privacy (SP)*, 2016, pp. 598–617, IEEE.

[23] T. van Nuenen, X. Ferrer, J. M. Such, M. Cote, "Transparency for whom? assessing discriminatory artificial intelligence," *Computer*, vol. 53, no. 11, pp. 36–44, 2020.

[24] A. Caliskan, J. J. Bryson, A. Narayanan, "Semantics derived automatically from language corpora contain human-like biases," *Science*, vol. 356, no. 6334, pp. 183–186, 2017, doi: 10.1126/science.aal4230.

[25] N. Garg, L. Schiebinger, D. Jurafsky, J. Zou, "Word embeddings quantify 100 years of gender and ethnic stereotypes," *PNAS 2018*, vol. 115, no. 16, pp. E3635–E3644, 2018.

[26] X. Ferrer, T. van Nuenen, J. M. Such, N. Criado, "Discovering and categorising language biases in reddit," in *International AAAI Conference on Web and Social Media (ICWSM 2021)(forthcoming)*, 2020.

[27] C. Dwork, M. Hardt, T. Pitassi, O. Reingold, R. Zemel, "Fairness through awareness," in *ITCS 2012*, 2012, pp. 214–226, ACM.

[28] N. Grgić-Hlača, M. Zafar, K. Gummadi, A. Weller, "Beyond Distributive Fairness in Algorithmic Decision Making," *AAAI*, pp. 51–60, 2018, doi: 10.2174/138161282166150416100516.

[29] N. Kilbertus, M. Carulla, G. Parascandolo, M. Hardt, D. Janzing, B. Schölkopf, "Avoiding discrimination through causal reasoning," in *NIPS'17*, 2017, pp. 656–666.

[30] S. T. Mueller, R. R. Hoffman, W. Clancey, A. Emrey, G. Klein Macrocognition, "DARPA XAI Literature Review p. Explanation in Human-AI Systems: A Literature Meta-Review Synopsis of Key Ideas and Publications and Bibliography for Explainable AI Prepared by Task Area 2 Institute for Human and Machine Cognition," no. February 2019, 2019.

Natalia Criado

Dr Natalia Criado is Senior Lecturer in Computer Science in the Department of Informatics at King's College London and a co-Director of the UKRI Centre for Doctoral Training in Safe and Trusted AI (www.safeandtrustedai.org). Her research interests are computational norms and normative multi-agent systems, as well as and the application of multi-agent systems, data science and artificial intelligence to enhance cyber security and privacy. Contact her by email at natalia.criado@kcl.ac.uk.

Xavier Ferrer

Dr Xavier Ferrer Aran is a Research Associate in Digital Discrimination at the Department of Informatics, King's College London. Xavier obtained his PhD in Informatics in 2017 from the Artificial In-telligence Institute of the Spanish National Research Council (IIIA-CSIC) and the Universitat Autonoma de Barcelona (UAB). Dr Xavier Ferrer Aran is a Research Associate in Digital Discrimination at the Department of Informatics, King's College London. Xavier obtained his PhD in Informatics in 2017 from the Artificial Intelligence Institute of the Spanish National Research Council (IIIA-CSIC) and the Universitat Autonoma de Barcelona (UAB).

Jose M. Such

Dr Jose M. Such is Reader (Associate Professor) in the Department of Informatics at King's College London, and Director of the KCL Cybersecurity Centre. His research interests are at the inter-section of Artificial Intelligence, Human-Computer Interaction and Cyber Security, with a strong focus on human-centred AI security, ethics, and privacy. He has been PI for a number of large projects funded by EPSRC, including the Discovering and Attesting Digital Discrimination (DADD) project, and the Secure AI Assistants (SAIS) project. Contact him on Twitter at @josemsuch, and by email at jose.such@kcl.ac.uk.

No App is an Island: Collective Action and Sustainable Development Goal-Sensitive Design

Steph Pitt, Marlína van Meelis Lacey, Ed Scaife, Jeremy Pitt*

Imperial College London, London (UK)

Received 30 November 2020 | Accepted 20 December 2020 | Published 15 February 2021

Abstract

The transformation to the Digital Society presents a challenge to engineer ever more complex socio-technical systems in order to address wicked societal problems. Therefore, it is essential that these systems should be engineered with respect not just to conventional functional and non-functional requirements, but also with respect to satisfying qualitative human values, and assessing their impact on global challenges, such as those expressed by the UN sustainable development goals (SDGs). In this paper, we present a *set of sets* of design principles and an associated *meta-platform*, which focus design of socio-technical systems on the potential interaction of human and artificial intelligence with respect to three aspects: firstly, decision-support with respect to the codification of deep social knowledge; secondly, visualisation of community contribution to successful collective action; and thirdly, systemic improvement with respect to the SDGs through impact assessment and measurement. This methodology, of *SDG-Sensitive Design*, is illustrated through the design of two collective action apps, one for encouraging plastic re-use and reducing plastic waste, and the other for addressing redistribution of surplus food. However, as with the inter-connectedness of the SDGs, we conclude by arguing that the inter-connectedness of the Digital Society implies that system development cannot be undertaken in isolation from other systems.

Keywords

Socio-Technical Systems, Value-Sensitive Design, Collective Action, Plug-in Architecture, Sustainable Development Goals.

DOI: 10.9781/ijimai.2021.02.005

I. Introduction

Transitioning to the Digital Society, as envisaged through the Digital Transformation, involves the increasing use of digi tal tools and technologies in the reconfiguration of business processes, organisational structures, and commercial transactions, and the recasting of social processes, physical infrastructure and relational interactions. This is having a profound impact on the nature of ownership (e.g. possession of goods being replaced by subscription to services), the sense of belonging to communities, and citizen's access to infrastructure for education, water, energy, medical treatment, and transportation; systems of manufacture, delivery and disposal; systems of justice, governance and political engagement; and monitoring and control over communal resources and the local environment.

In this context, there is a challenge to engineer ever more complex socio-technical and cyber-physical systems to support and enhance this full spectrum of human activities, to address wicked societal problems. Therefore, as outlined in Section II, it is essential that these systems should be engineered with respect not just to conventional functional and non-functional requirements, but also with respect to satisfying qualitative human values, and assessing their impact on global challenges, such as those expressed by the UN sustainable development goals (SDGs), a set of inter-connected goals intended to achieve a fairer and more inclusive future world.

* Corresponding author.

E-mail address: j.pitt@imperial.ac.uk

To meet this challenge, this paper builds on the methodologies of Value-Sensitive Design [1] and Socially-Sensitive Design [2], and proposes a methodology of *SDG-Sensitive Design.* To begin with, Section III considers a range of different value-sensitive design perspectives on engineering a socio-technical system. Here, we consider a socio-technical system to be one which recognises the importance of interaction between people and technology in system design: crucially, in the Digital Society that technology includes components with Artificial Intelligence, as manifested by a software agent, 'smart' device, robot, and so on.

From each perspective, Section III.C derives a *set of sets* of design principles, which focus design of socio-technical systems on the potential interaction of human and artificial intelligence with respect to three aspects: firstly, decision-support with respect to the codification of deep social knowledge; secondly, visualisation of community contribution to successful collective action; and thirdly, systemic improvement with respect to the SDGs through impact assessment and measurement. In Section IV, an associated *meta-platform* is described, whose key features include transparency, generativity and reconfiguration through plugins, which can be used to encode the deep social knowledge encapsulated in the design principles, and to evaluate, measure and visualise the contribution to achieving one or more of the SDGs.

This methodology is illustrated through the design of two collective action apps for sustainability, each targeted at a specific societal problem; one (described in Section V) for encouraging plastic reuse and reducing plastic waste, and the other (presented in Section VI) for addressing redistribution of surplus food. However, as with the inter-connectedness of the SDGs, we conclude by arguing that the inter-connectedness of the Digital Society implies that *no app is an island*: system development cannot be undertaken in isolation from other systems, without concern for qualitative human values, or without considering its impact on achieving the SDGs.

II. Problems, Goals and Impacts

A. Wicked Problems

There are a host of societal problems that need to be addressed as part of the Digital Transformation to the Digital Society. This includes energy poverty, food insecurity, air quality, social justice and plastic reduction. However, all of these problems can be classified as *wicked* problems.

A wicked problem can be identified as a societal problem whose complexity and continually changing requirements is such that there is not necessarily a stopping rule or terminating condition, nor may there be a consistent set of criteria by which to evaluate immediately or ultimately such a condition. Additionally it may exhibit some or all of the other characteristics identified in [3].

The difficulty of finding digital solutions to wicked problems, like those listed above, is further exacerbated by (at least) five issues. Firstly, evaluation: for example, in the case of sustainability, there is no end state, by definition; so it can be asserted "has been sustained", and it can be claimed "is sustainable" , but neither of these imply "will be sustained (indefinitely)"). Secondly, dealing with unexpected, emergency and potentially catastrophic situations, for which the system may not even have been designed. Thirdly, polycentricity: there being multiple stakeholders with different and possibly conflicting objectives. Fourthly, the potential of technologies like Artificial Intelligence (AI) to be mis-used, for example by creating intrusive monitoring frameworks (e.g. surveillance capitalism [4]). And finally, but perhaps above all, satisficing (rather than satisfying) qualitative human values – i.e. we are dealing with a multi-criteria sub-optimisation problem with subjective, non-numeric data points (cf. [5]).

This implies that system design is not restricted to functionality: it also has to take into account values, and other systems affecting the same values. Our approach consists of applying a set of design principles to develop systems for a new platform for social coordination – and then try to identify which of the SDGs it might impact, and consider how to measure that impact.

B. Sustainable Development Goals (SDGs)

Adopted by all United Nations (UN) Member States in 2015, the UN SDGs are a set of 17 interconnected goals to address fundamental global challenges by 2030, and have therefore been described as "the blueprint to achieve a better and more sustainable future." Designed to be inclusive of all, these goals are a call for action by all countries to address challenges related to poverty, inequality, the climate crisis, peace and justice. Importantly, SDG 17 calls for "Partnership for the Goals" which specifically highlights the importance of achieving each goal in concert with the others. These integrated and universally accepted goals underpin a shared agenda, therefore enabling a collective action response that encourages innovation, in which the transition to the Digital Society will play a profound role. More often, global societal challenges are being addressed through technological innovation, thus are likely to be making an impact on achieving one or more of the SDGs. Therefore, when considering how digital innovation can aid in achieving one, or indeed any, of the SDGs, identifying and measuring the impact made is crucial. Commonly referred to as *impact measurement*, having a quantifiable understanding of exactly how anything – in this case digital innovation – is shaping our progression towards achieving the SDGs is fundamental to appreciating where we currently stand, and what more needs to be done.

C. Impact Measurement

As mentioned above, the SDGs are designed to be achieved in concert, so whilst it is crucial to measure the positive impacts made to achieving some of the SDGs, it is equally important to measure unintended (and potentially negative) impacts against others, thereby enabling a holistic and fully comprehensive overview of projects, policies, and plans, particularly concerning digital innovation. However, it is equally important to operationalise targets and take into account the significance and relevance of specific metrics [6] – not forgetting Goodhart's Law, that if any indicator or metric of some property of system is used instead as a target, then it ceases to function as a meaningful indicator of that property.

Beyond knowing how close (or not) a policy is to achieving its goals, impact measurement is both beneficial and important in many other respects. Firstly, it can aid with monitoring risks, as a desirable output does not necessarily lead to a desirable outcome. Secondly, impact measurement is crucial for business investment. According to [7], investors desire more detailed social and environmental performance data as this provides an improved understanding of non-financial returns, thus providing the opportunity for investment capital to be re-allocated accordingly. Thirdly, measuring impact enables accountability and transparency as it can be used as a metric to keep track of performance. Fourthly, impact measurement aids in better understanding social innovation. If social innovation is considered to be essentially experimental, we require impact measurement to inform the extent of success of such experiment. Finally, impact measurement offers a way to better communicate ideas, share views and add valuable contributions in a universally recognisable language, as the SDGs are stable and well-established.

However, there are a number of challenges to successful impact measurement, and despite being a noble expression of intention and ambition, ambiguity of interpretation generates a high level of complexity when attempting to measure impact. Primarily, the indirect and even long-term effects that are fundamental to generating a holistic view of an impact, are difficult to capture for a variety of reasons ranging from simply being unknown, to being abstract, qualitative and entirely theoretical. Furthermore, the SDGs are purposely designed with slight ambiguity, enabling each goal to be inclusive, interconnected, applicable to multiple stakeholders, and dynamic in order to absorb drastic global changes that may occur over the 15 years between their introduction in 2015, to their end goal in 2030.

As a result, there is simply no standardised, one-size fits all method of measuring impact. Instead, there is a pluralism of impact measurement models proposed and used by different organisations. Having many different measurement methods then poses challenges to making comparisons within wide scope evaluations of multiple ideas, projects or policies, for example. For smaller-scale projects, impacts (positive or indirect) may be smaller and therefore easily ignored (for example, small scale economic impacts). Moreover, for these small-scale or start-up projects, there is often limited financial resource, along with many competing strains placed on the budget, therefore rendering impact measurement – which will require additional resource – a near impossible task. Finally, a fundamental difficulty in measuring impact is often that it is limited in and by design: it cannot easily be 'bolted on' as an afterthought. Concern for SDGs (and values) has to be an integral part of the design process, and this depends on design principles.

III. Design Perspectives

In this section, the methodology of value-sensitive design is briefly reviewed, followed by our perspective(s) on the design of self-organising socio-technical systems with respect to a core set of critical human values: sustainability, socially-productive purposes, justice, legitimate governance, prosocial incentives, and personal identity. Based on this, we propose to enhance, or, complement the methodology of value-sensitive design (VSD) with a set of sets of inter-related design principles, which are applied depending on the perspective that is adopted.

A. Value-Sensitive Design (VSD)

In [1], it is suggested that VSD brings forward a "unique constellation of eight features", which includes: proactive influence of qualitative values on technological design from an early stage in the process; documenting values as "supra-functional" requirements with tests for compliance, system readiness and quality assurance; the iteration over and integration of conceptual, empirical and technical analysis and development; enlarging the scope of values beyond co-operation and participation (e.g. to self-actualisation and empowerment); distinguishing between usability and values with ethical significance; consideration of different classes of stakeholder, often observed in socio-technical systems; and building from the psychological proposition that values are universal (if possibly culturally relative).

However, we argue that socio-technical systems for solving wicked problems manifest many different values, and depending on the perspective one takes on the wicked problem that is being addressed, a different design approach is required.

B. The Socio-Technical Systems "Necker Cube"

We liken the problem to resolving a six-way Necker cube, because what is paramount depends, to a significant extent, on the perspective taken, which determines which face of the cube is 'on top', as illustrated in Fig. 1.

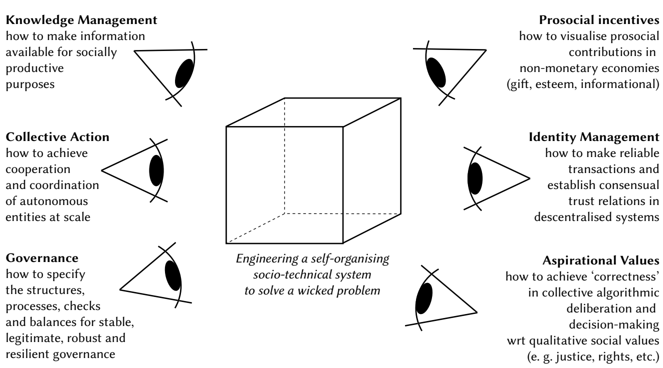


Fig. 1. The 6-Way "Necker Cube" for Value-Sensitive Design of Socio-Technical Systems.

Therefore, the design of a self-organising socio-technical system to address a wicked societal problem can be viewed as:

- a *collective action problem*: how provide the conditions for the evolution or emergence of cooperation and coordination of autonomous entities at scale;
- a *knowledge management problem*, how to make information available for socially productive purposes, from a diverse set of sources and expertise;
- a *legitimate governance problem*: how to determine the structures and processes for constitutional, collective and operational choice which avoid tyranny (as oligarchy, autocracy or majoritarian tyranny);
- a *prosocial incentive problem*, i.e. how to incentivise and visualise transactions in different types of non-monetary value-creation and exchange systems (e.g. esteem, gift, informational), to increase the social benefits of cooperation;
- an *aspirational values problem*, the ambition to achieve 'correctness' in collective algorithmic decision-making with respect to some set of shared and congruent values (aspirations); and
- an *identity management problem*, how to establish compartmentalised and consensual trust relations in decentralised systems, without compromising privacy.

C. Design Principles

In her pioneering work on self-governing institutions for sustainable common-pool resource (CPR) management [8], Ostrom observed that there were eight common features of the institution that determined whether or not the resource was sustained. She then turned to the issue of *supply*, and argued that if faced with a CPR collective action problem, instead of 'evolving' an institution with the necessary features, design one instead. The eight common features were then transformed into design principles. Effectively, these design principle represented deep social knowledge about the nature of self-governing institutions for sustain-able common-pool resource management.

In addition, in various works, we have attempted to identify similar findings from economic, political and social science on how the other problems (identified in the previous section) have been addressed in social systems, and we have tried similarly to distil this deep social knowledge into corresponding design principles. This has included:

- knowledge management principles derived in classical Athenian democracy [9], some of which are formalised in [10];
- principles of legitimate governance derived from a theory of basic democracy [11], and formalised as *democracy by design* [12];
- principles of prosocial incentives and social capital partially derived from anthropological studies of gift economies [13] and discussed as principles of axial (crytpo-)currency design [14]; and
- principles for aspirational values, in particular various different aspect of justice (e.g. distributive [15], retributive, procedural, and interactional).

The full set of sets of design principles is summarised in Table I. Note that the issue of identity management remains an open question, but design principles are being formulated, see e.g. [16].

However, even after applying all these principles in design, in practice there is also an *inertial problem*, caused by network growth as result of preferential attachment, the network effect (value increases non-linearly with scale), and the centralising tendency of the Internet at the application layer. This has inexorably led to the private ownership of the means of social coordination and information dissemination existing on an essentially publicly-built infrastructure. So there is a question of how to provide a viable alternative platform to the monopolist gatekeepers that have emerged as a result of the network effect at the application layer of the Internet.

We therefore need a platform which overcomes this inertial problem, supports multi-perspective VSD by *encoding* deep social knowledge (as captured by the design principles), and helps with impact measurement with respect to the SDGs. In fact, we need a *meta-platform*, a platform for generating platforms. A prototype of such a meta-platform is presented in the next section.

IV. PlatformOcean Meta-Platform

Experience with digital platforms for eLearning, eHealth, etc., and for other 'as-a-service' operational models, would suggest that there is

TABLE I. Design Principles for Socio-Technical Systems Collective Action

Collective Action
Clearly defined boundaries
Congruence between rules and environment
Collective-choice arrangements respect self-determination Monitoring, by self or appointed agencies
Graduated sanctions
Conflict resolution
Minimal recognition of right to self-organise
System of systems
Knowledge Management
Clearly defined boundaries
Lower transaction costs of knowledge exchange Agreement on common interest questions
Distinction between common interest and partial goods Common knowledge of procedural rules
Epistemic diversity
Recognition of expertise
Focal points for collective action
Legitimate Governance
Prevention rather than re-invention
Democracy is not an end-state, nor default Seamless transfer of power
No compromise on democratic processes Visibility, inclusivity, transparency, accountability Inter-dependence of diversity
Education in pro-social benefits
Procedural evaluation and reflexivity
Aspirational Values
Clearly defined roles and powers
Educate – populate – majoritate
Create conditions for cooperation
Evaluate costs of enactment vs. non-enactment Publicity implies parity
Popularity does not imply impartiality Diversity of sources and forces
Reflexivity
Transactional Values
Delimit purchasable goods
Identify purpose of currency
Closed loop (not convertible into fiat currency) Gratitude gift currency (not judgement currency) Issuance (initialisation of currency)
Right to mint
Determination of allocation decided by guild No debt, no credit, no interest

so single universal technological solution suitable for all sustainable-development collective-action problems. However, developing bespoke, and non-inter-operable platforms, leads to fragmentation, lack of re-use, loss of experience and ex-pertise, and so on.

Therefore, we are developing a meta-platform, which is being called PlatformOcean, which allows users to download an open source platform with a range of hosting options, instantiate that platform for particular requirements through mods and plug-ins, and make it accessible through a generic client and standard protocols.

The key design features of the PlatformOcean meta-platform, as illustrated in Fig. 2, are:

- *deep social knowledge*: the plug-in architecture also supports codification of the deep social knowledge captured by each of the five sets of design principles;
- *generativity* [17]: a tool to support the creation of new tools that were not envisaged by the designer of the original tool;
- *common-pool development*: the creation of plug-ins is a reflective process, as communities of developers use an instance of the platform to develop plug-ins for other instances;

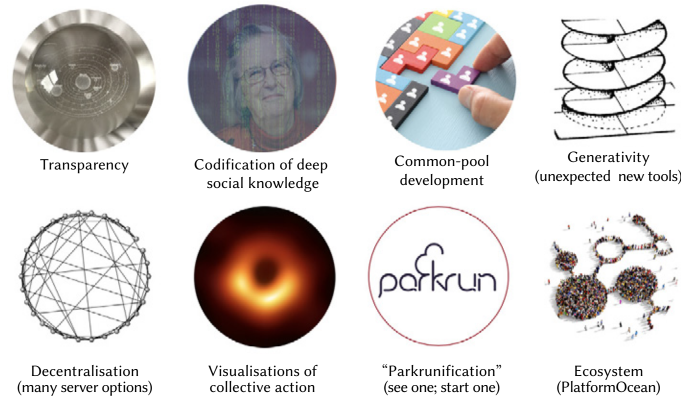


Fig. 2. PlatformOcean: Design Features.

- *visualisation*: for collective action, feedback to individuals and communities and how their small actions X contributed to a greater action Y which had significant impact Z;
- *parkrunification*: the generic platform can be used to create new downloadable instances of the platform, which others can use for faster customisation for related applications (we refer to this process as 'parkrunification' after the explosion of popularity of the parkrun phenomenon following the same process of observation and imitation);
- *server-side transparency*: the system architecture allows a range of options for self-hosting, with multi-purpose multi-function self-configuration implemented through plug-ins, supporting decentralisation;
- *client-side transparency*: each group or conversation in the client is with a different server, but open standard transfer protocols and programming interfaces provide seamless client-side integration, protect data, preserves privacy and prevent data leakage;
- *ecosystem*: the creation of a platform ecosystem supports sustainability through diversity and inhibits monopoly.

The PlatformOcean meta-platform provides a foundation for the codification and implementation of the design principles discussed in the previous section. The core idea is to provide communities with a fully customisable and self-hosted solution that best facilitates communication, whilst at the meta-level providing a toolset for collaboration in developing sustainable social media ecosystems. To achieve this aim, the meta-platform has been designed with a flexible plug-in architecture, inspired in part by other projects, for example: the Eclipse IDE, the computer game Minecraft, and Open Mustard Seed [18]. The overall architecture of PlatformOcean is illustrated in Fig. 3.

The platform is also designed to support client-side transparency. As Fig. 3 shows, each client conversation in a client-side app is interacting with one of the n platform instances, each of which can be distributed on a different platform with a different hosting option. Consequently, the server distribution appears seamless to the user.

The PlatformOcean meta-platform supports three additional features designed to overcome the inertial problem in the development and take-up of social platforms: self-customisation, generativity and re-use. The first of these key features is the flexible self-customisation and self-extension of individual PlatformOcean platform instances (see Fig. 3: UrbanRefill and UrbanForage platform instance are presented Section V and Section VI). Additionally, the open-standard nature of the communication protocol facilitates the development of custom clients, and increases interoperability and platform mobility. In this manner each platform instance, while derived from the same set of resources, acts in a way that is specialised to its user-base, and customised by its user-base.

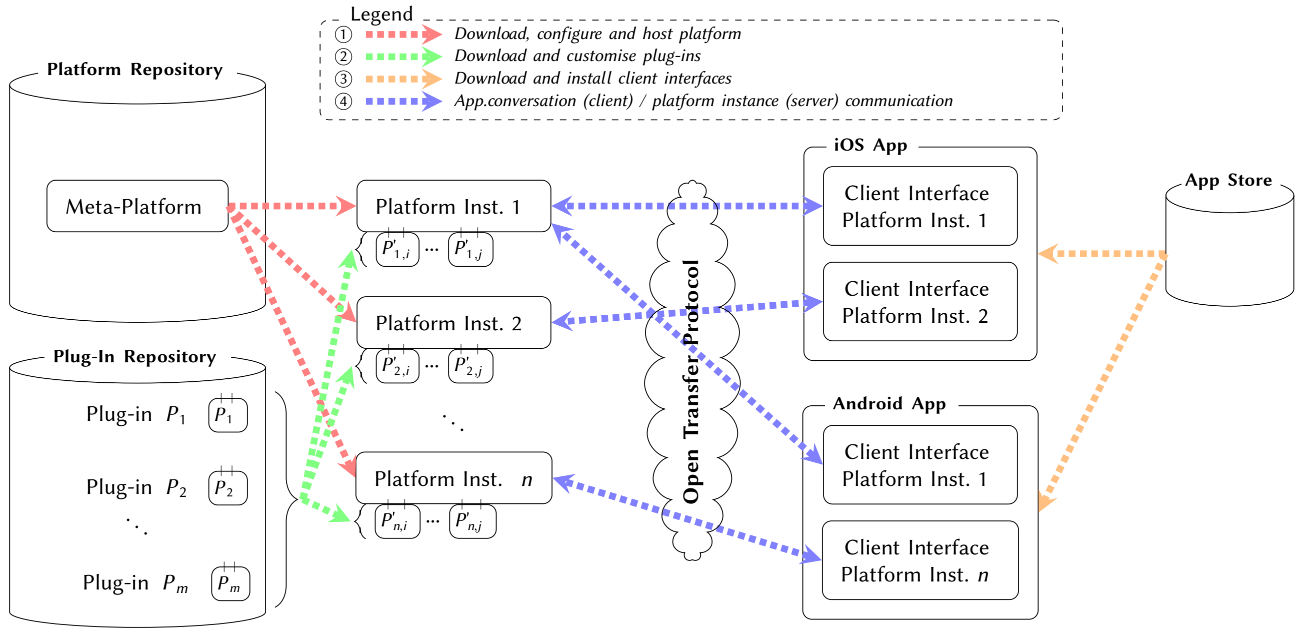


Fig. 3. PlatformOcean: Overall Architecture.

The plug-ins provide both additional functionality and the codification of deep social knowledge encapsulated by the design principles. Aspects of well-functioning governance systems – among others, the ability to implement deliberative assemblies based on rules of order [19], or allocate responsibilities fairly and define boundaries on member behaviour – can be replicated in server functionality using plug-ins. In the same way, plug-ins can also be designed to support the development of pro-social incentives, for example in the tracking and visualisation of progress towards goals over time or of individuals collaborating in a community to pool their achievements (as illustrated in the exemplars of the next two sections). Critically, we believe that re-usable plug-ins can also be developed for linking collective achievements to SDGs, and for linking platforms for different application which actually contribute towards SDGs in common. It is here that we envisage a particular role for AI components in a socio-technical system, for example each plug-in could be a communicating agent in a multi-agent system, or could provide intelligent decision support for deliberative assemblies, or as a real driver for equity [20].

The second feature ties strongly with the 'meta' component of the meta-platform. PlatformOcean is instilled with a notion of generativity. By sharing plug-ins and client designs, PlatformOcean aims to facilitate an over-arching social media eco-system, wherein disparate groups can support each other through the de-velopment of 'reusable parts'. A central plug-in and client repository (as illustrated in Fig. 3) will ideally provide a hub for shared development. This calls back to the concept of social capital: this central hub may have the added benefit of providing opportunities for previously unconnected groups to meet, 'build bridges' and collaborate on projects outside of the PlatformOcean context.

The final feature refers to the accessibility of the platform. The ease of setting up and self-hosting an instance of PlatformOcean is a key consideration. The server software can be installed, customised and run on hardware ranging from raspberry pi to personal computers with fairly minimal technical knowledge required. As technical solutions increase in complexity, the skills often required to set up and maintain them can often increase. Keeping the platform accessible to those with little technical background is imperative to its function as a sustainable solution. Moreover, the aim is to create open source archives and encourage collective development of plug-ins, i.e. this is a collective action problem in itself, and could be addressed by an instance of PlatformOcean itself (see Fig. 4).

The next two sections present the design of two such platform instances to address an environmental and a societal problems, respectively excess plastic waste (Section V) and unequal food distribution (Section VI).

V. Exemplar 1 - Plastic Waste Reduction

In this section, we describe the design of an app called UrbanRefill, aiming to reduce the use of plastic (and increase of plastic waste) by re-using liquid containers. The design of UrbanRefill applies the design principles of Section III.C with the target meta-platform of Section IV, addressing the following requirements:

- *functional requirements*: achieve the basic function of the application by reducing plastic consumption;
- *value-sensitive requirements*: enhancing sustainability, improving inter-connectivity and community capability for collective action;
- *SDG-sensitive requirements*: wider contribution to protecting the planet; preventing climate breakdown and; achieving the relevant SDGs.

We address each of these requirements in turn.

A. UrbanRefill: Application Design

The universal abundance of plastic has caused significant and wide scope damage to the planet, involving issues such as microplastic pollution and increased waste, leading to (sometimes irreversible) damage to numerous ecosystems. Additionally, our current plastic use fuels an unsustainable throwaway culture, in which the majority of products are deemed dispensable.

PlatformOcean's UrbanRefill application focuses on reducing single use plastics, by enabling the refilling of common household products, such as washing up liquid, hand soap and shampoo, as opposed to the purchase of new ones. In the UK at present, there are several types of refill scheme; however these have some limitations, ranging from failures to design practicality into the scheme, whilst also resulting in higher transport-related emissions (e.g. by using heavier materials such as glass), to an over-reliance on people's organisational skills.

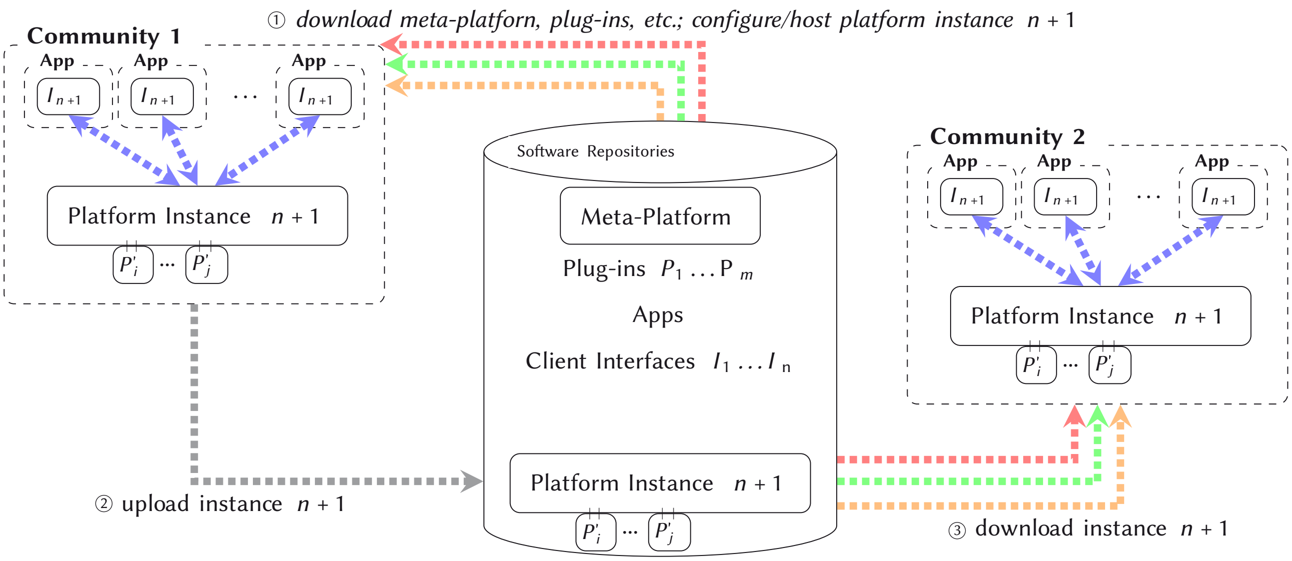


Fig. 4. PlatformOcean: Timeline of Re-use.

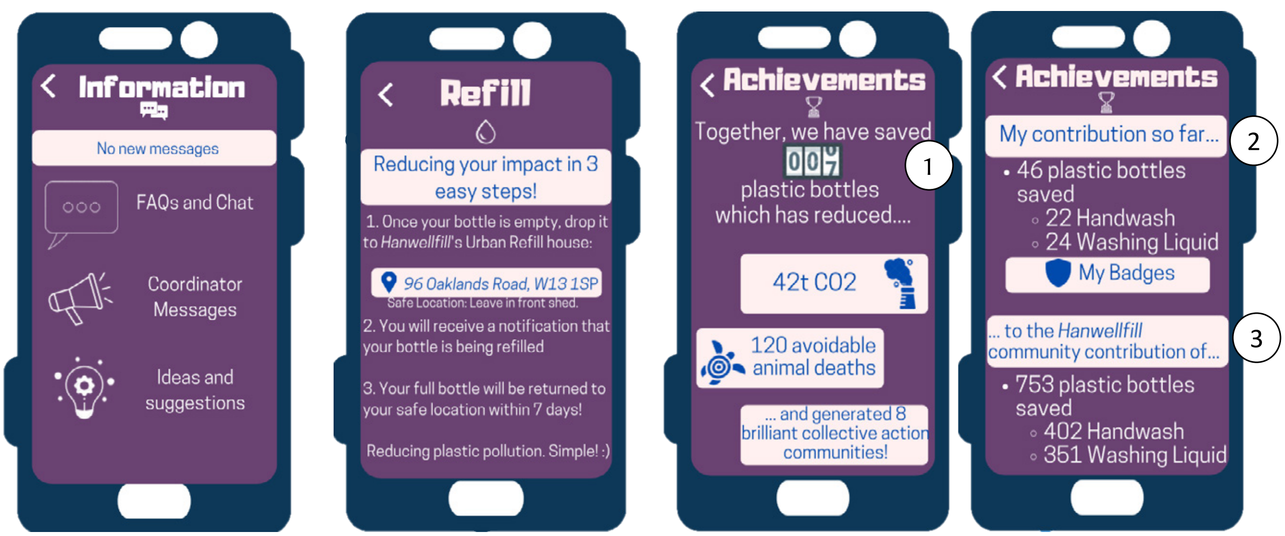


Fig. 5. UrbanRefill Design Template. Showing four screens (from left to right): Information which encourages community-wide active engagement, thus satisfying collective-choice arrangement principles, as well as visibility, inclusivity, transparency and accountability principles; Refill which provides Members with common knowledge of the rules; and Achievements, visualising community contribution to promoting values (e.g. sustainability) and to achieving SDGs.

UrbanRefill is distinct from other refill schemes in that it generates a community-led initiative to refill[1] . The premise is straightforward: within a given community setting, one Coordinator holds products in bulk, with *Members* dropping empty bottles off for the Coordinator to then refill and return within seven days. Importantly, for each product, there is a generic type of bottle which has a unique barcode and is associated to a particular Member. The app supports the management, coordination and logistical elements of the refill process in several ways:

- A Member's *Activity Status* provides information on the receipt of an empty bottle, as well as its return. Additionally, Members can choose to opt-in to push notifications for this information.
- The Coordinator has access to a bottle barcode scanning feature, which is used to scan empty bottles in, thereby marking them as *refill in progress*. The associated Member's bottle Activity Status is updated, and a Coordinator *Workflow* is generated, enabling the Coordinator to keep a track of the seven-day turnaround.
- In addition, the Coordinator has access to a map which reflects the location of Members and their bottle activity status, thus aiding the management and logistics of returns.
- Due to the genericity of the bottles for each product, prices and payments are pre-approved by Members and automatically taken upon the Coordinator marking a bottle as *returned*.

B. Applying the Design Principles

A quartet of screenshots of UrbanRefill are shown in Fig. 5, illustrating aspects of functional, value-sensitive and SDG-sensitive design. This section discusses each of these in turn.

1. Functional Design

With respect to the functioning of UrbanRefill (i.e. achieving the basic function of the app), there is the capability to ensure **congruence between the rules and the environment**, as the bottle drop off activities of both the Members and Coordinator can be individually configured according to the abilities of the people involved in the task. In addition, as the system is designed to take place within a local community setting, the **transaction costs** are lowered as these

[1] Currently the focus is on the aforementioned household goods, but variations of the application could enable expansion into many more products, as well as different ways of organising individuals within the community.

drop-off activities can be incorporated into tasks that were already occurring. Finally, with respect to the design principle of **publicity implies parity**, both Members and Coordinators have access to essential knowledge within the app, with the opportunity to gather additional information through the use of the *FAQ* or *Chat* systems.

2. Value-Sensitive Design

There are a number of design principles that have been utilised in the designing of UrbanRefill to address the value-sensitive requirements of the app. Firstly, there are **clearly defined boundaries** within the community – a participant is either a Member or a Coordinator – and the rules for those who have the right to appropriate from the common pool resource are correspondingly clearly defined, according to the design principle of **common knowledge of rules and their generation**.

Secondly, through the use of the Information page, **collective-choice arrangements** are encouraged, as members are able to provide feedback on the functioning of the system, whilst also contributing to the expansion and growth of the products that are available within their community.

Thirdly, **reflexivity** within the community is designed with a *Feelings* response that can be completed by both Members and Coordinators after each transaction. Aggregating this data within a community provides a visualisation of overall satisfaction, enabling introspection on the collective endeavour. This information could also prove useful in future variations with respect to dispute resolution (see below).

Finally, **visibility, inclusivity, transparency and accountability** are well-woven into the design of UrbanRefill, particularly within the Information page which aids transparency, inclusivity and active engagement throughout the community on past and current decisions. Transparency and visibility are also achieved as all Members are aware of the Coordinator being the primary decision-maker, with inclusivity also encouraged within this process. Additionally, designed into the basic premise of the app are well-established rules for the group, for example, regarding the mechanisms and frequency of refilling. The design of group achievements within the *Achievements* page will also aid in encouraging each community to work collectively.

3. SDG-Sensitive Design

By designing in a clearly established **common interest question** – achieving the goal of reducing plastic waste – UrbanRefill's Achievements page enables the aims of the ecosystem to be achieved. In doing so, positive contributions made towards SDGs are visible for all Members and Coordinators, which not only encourages use of the app but, crucially, allows for the im-pact of these positive contributions to be measured. By including the Achievements page within the design of UrbanRefill, impact measurement can occur. The contribution that UrbanRefill makes towards achieving the SDGs, and how this impact could be measured, is discussed in Section V.C.

4. Limitations

However, some of the identified principles are missing, either because they are not applicable, or because the user-centred design process could not anticipate the problems that could occur. For example, design principles relating to graduated sanctions and conflict resolution have not (yet) been included and are currently not applicable, as these 'errors' or misuses of the app are learned from trial and real-world implementation. In addition, the design of axial-crypto currencies (prosocial incentives) are not applicable, and design has been omitted, due to the fact that the current version of UrbanRefill uses only real-world financial transactions for real-world products.

It is, however, very possible that in other variations of the app such principles will need to be considered – for example in a case in which each Member harbours one of the products, with exchange done between Members on a product-equivalence basis rather than a financial exchange. Finally, principles of democracy by design (legitimate governance) are mostly absent from the current version of the app as it does not wholly function as a rules-based system, particularly with respect to the power balance between Members and Coordinators. Despite designing in both structures for organising the scheme and encouragement to produce an equal and fair system with less centralised authority (e.g. encouraging engagement of Members on decision making through the Coordinator Messages), the Coordinator does appear to control the majority of decisions. Therefore, as the system would scale, it might be necessary to introduce an 'Ombudsperson' role who would intervene in dispute resolution.

C. Impact on SDGs

Whilst, at present, there is not a focussed SDG designed to target the issues surrounding anthropogenic plastic generation and consumption specifically, given our ubiquitous and varied use of this material, the potential threats and present damage it causes, and its near "indestructible" nature (microplastics takes hundreds of years to degrade), the issue of plastic pollution spans several of the 17 SDGs. The predominantly relevant SDGs identified to plastic use are highlighted below, alongside a brief explanation as to how the use of UrbanRefill will contribute to the achievement of the specified goal, as well as, crucially, how this impact can be measured and reflected within the Achievements section of the app.

SDG 11: Sustainable Cities and Communities

Impact: building community networks that both allow for and promote sustainability, whilst also reducing plastic waste, which is often shipped from developed to developing countries;

Measurement: monitoring the number of Refill communities built on the platform can contribute to better understand the number of ongoing community-lead collective action initiatives.

SDG 12: Responsible Consumption and Production

Impact: generating a cultural and societal change to no longer accept the unsustainable norms of plastic use;

Measurement: tracking the number of plastic bottles 'saved' (i.e. not used because the bottle has been reused), and therefore the number of times individuals have chosen to reuse, rather than repurchase.

SDG 13: Climate Action

Impact: reducing plastic use lowers the emissions associated with plastic production processes, and with subsequent delivery of these goods to customers;

Measurement: estimating the total equivalent CO2 emission reduction as a result of the number of bottles 'saved'. This could also be disaggregated to display the information at individual, community and ecosystem levels.

SDG 14: Life Below Water

Impact: reducing plastic pollution in bodies of water, thereby lowering the consumption by fish, animals caught in waste materials, and microplastic chemical pollution;

Measurement: estimating the number of animal lives saved from death incurred due to plastic consumption or entanglement.

SDG 15: Life on Land

Impact: reducing plastic pollution on land and lowering the build-up of non-recyclable plastics, reducing its consumption by birds and other land animals;

Measurement: estimating the number of animal lives saved from death incurred due to plastic consumption, as well as the number of plastic bottles 'saved' and not contributing to waste build-up.

With regards to SDG 3 and 4, the use of UrbanRefill will undoubtedly contribute positively to these SDGs, however, no metric has yet been identified to accurately reflect the impact that the use of UrbanRefill has made towards achieving the goal.

SDG 3: Good Health and Wellbeing

Impact: by preventing climate change, wellbeing is likely to improve, generating or improving a community network can encourage active transport and improve wellbeing.

SDG 4: Clean Water and Sanitation

Impact: reducing microplastic pollution which affects all waterways including freshwater bodies, and thereby preventing chemical decomposition of plastics within freshwater.

Finally, it is important to note that, as mentioned in Section II, one of the challenges of interconnectedness and impact measure-ment is being aware of, and measuring unintended impact. It is an 'unknown unknown' that would need to be anticipated in any evaluation of UrbanRefill and the focal point of any re-design.

VI. Exemplar 2 – Surplus Food Redistribution

This section describe the design of a second app, called UrbanForage. The aim of UrbanForage is to decrease food waste through surplus food distribution, and again its design applies the principles of Section III. C and targets the meta-platform of Section IV, in particular with plug-ins delivering common functionality.

A. UrbanForage: Application Design

Food loss and waste has become one of the most pressing resource-use challenges at community, national and international scales. Globally, a third of all food produced for human consumption is either lost or wasted [21], resulting in the additional loss of all embodied inputs such as water, energy, labour, land and capital. Food waste accounts for 8% of global anthropogenic greenhouse gas emissions and costs the global economy $940 billion annually [22], [23]. Not only are the environmental and economic impacts significant, there are also huge ethical consequences, as 1 in 9 people remain undernourished in a world where excess food is generated on a daily basis [22]. The redistribution and reuse of surplus food is a means to reduce the impacts of food waste, which is arguably a product of the current unsustainable food supply chain. However, as well as seeking systemic change through political channels, we can also try to apply pressure through bottom-up behavioural change.

In the UK, surplus food redistribution and reuse charities are having a profound impact at both local and national levels. Collectively, these charities work across all sectors of the food system, in addition to multiple community level initiatives and digital applications. However, at least three limitations can be identified. Firstly, not all surplus food donations are either environmentally or economically viable for collection by surplus food charities. Secondly, user retention for relevant applications is not consistent, thus labour to do the necessary work is not always guaranteed. Finally, the independent and fragmented nature of current community level efforts reduces the true impact potential of collective community action [24].

The UrbanForage application brings together four potential stakeholders:

- Coordinator: a charity such as City Harvest[2] or The Felix Project[3] who acts a broker people the other three stakeholders;
- Surplus food donors: organisations with excess food, e.g. supermarkets, restaurants, wholesalers, caterers, etc.;
- Beneficiaries: organisations providing nourishment to vulner-able people, such as homeless shelters, women's refuges, care centres, children's breakfast clubs, etc.;
- Volunteers: individuals providing pick-up and delivery.

In particular, UrbanForage addresses the first limitation, that some pick-ups and drop-offs are uneconomic or non-environmental to use a van, so we try to use volunteers who walk or cycle, and lower transaction costs by aligning the delivery route with a journey they would have taken anyway, for example as part of their commute. In this way UrbanFor-age enhances the capacity of redistribution charities through increasing stakeholder self-organisation.

B. Applying the Design Principles

Despite the apparent differences in functionality, many of the issues addressed by UrbanForage are the same as in UrbanRefill; therefore generic plug-ins can be used to provide this functionality: a notable example is the Badges and Impacts pages. Moreover, the client interface has much in common, so unsurprisingly then perhaps, the look-and-feel of both applications is similar, see Fig. 6.

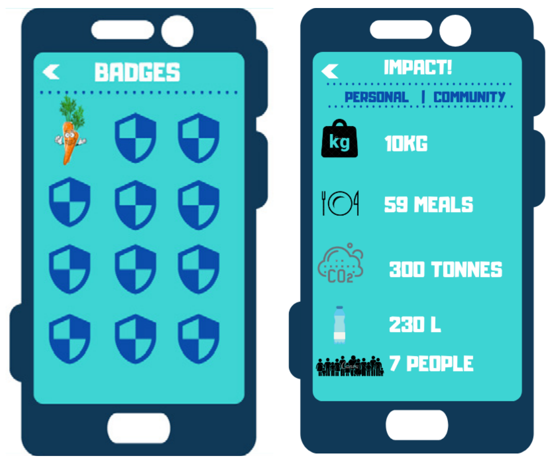


Fig. 6. UrbanForage: Badges and Impact Screens.

However, there are key divergences as well. One notable divergence is in the monitoring and sanctions, which as noted above, in UrbanRefill, was left undesigned, but in UrbanForage the volunteers need to be vetted by the coordinating charity (a legal responsibility) and moreover their performance of delivery tasks *has* to meet an approved standard (food delivered too late is still wasted; some food types deteriorate, and so on). Therefore, the system of monitoring and sanctions has to be very explicit, but for this purpose, a bespoke plug-in can be designed and implemented, and integrated like any other plug-in.

C. Impact on SDGs

The current unsustainable and resource-intensive food system results in vast amounts of food loss and waste. Due to the interconnected nature of the food system, its impacts –environmental, economic, social and political – are complex, having both direct and indirect impacts on a number of SDGs. Several of these are in common with the UrbanRefill application, but the relevant SDGs identified to food loss and waste are highlighted below. We give a brief explanation as to how the use of UrbanForage will contribute to the achievement of the specified goal, as well as, crucially, how this is impact can be measured and reflected within the Achievements page of the app.

[2] www.cityharvest.org.uk/

[3] thefelixproject.org

SDG 2: Zero Hunger

Impact: the redistribution and reuse of surplus food facilitates meal generation for the most vulnerable, as well as adds to food bank resources, helping to mitigate against hunger and food insecurity;

Measurement: monitoring the number of meals provided can provide figures as to how many vulnerable people are receiving regular meals.

SDG 12: Responsible Consumption and Production

Impact: redistributing and reusing surplus food directly de-creases the quantity of food waste;

Measurement: monitoring the net amount of surplus food that is reused will provide data as to how much food waste is prevented.

SDG 13: Climate Action

Impact: a reduction in food waste directly decreases emissions associated with embodied input (e.g. land, water, energy), as well as emissions generated in landfill;

Measurement: estimating equivalent water, CO2 emission and net weight of food resource saved; and aggregating this data across multiple local charities and food distribution schemes into national statistics.

VII. Summary and Conclusions

No man is an island, entire of itself

– John Donne, 1624

In summary, the basic problem addressed by this paper is engineering socio-technical systems to address wicked societal problems and contribute to meeting the UN Sustainable Development Goals. It presented six different perspective on the problem and, for five of them proposed a set of design principles. Based on the reconfigurable plug-in architecture of the meta-platform (PlatformOcean), it was proposed to use AI in plug-ins, firstly to codify the deep social knowledge encapsulated by the principles to support organisational coherence with respect to the SDGs, and secondly to facilitate the collection of data for measuring and assess-ing impact on the SDGs.

In this sense, this work could be considered to be an instantiation of value-sensitive design called *Sustainable Development Goal-Sensitive design*. We have applied this design methodology applications for two pressing societal problems, excess plastic waste and unequal food distribution. In both cases, it could be seen how individual data and actions could be aggregated and processed by AI plug-ins in order to assist with self-organisation and sustainability but also to visualise individual contributions to the collective and make recommendations.

As we transition further towards the Digital Society and progress towards reaching the SDGs' 2030 target, there is a unique opportunity – and a profound responsibility – to consider what sort of digital future we want to shape.

In particular, we have argued that socio-technical platforms for social coordination are not standalone systems, and have to be designed and deployed with respect to each other as part of a "platform ecosystem", with respect to qualitative human values, and the impacts on the SDGs. The particular role of AI anticipated here would be variously in the coordination and sustainability of the ecosystem, in the codification of deep social knowledge that helps realise human values, and in the oblique measurement of the impact on SDGs (see, for example, [25]).

In measuring these impacts, we are able to account for the wider implications of our digital innovations, and by doing so, we are provided with immense prospects: mitigating anthropogenic climate change, reducing inequality and poverty, and promoting peace and justice. Through measuring the impact of our plans, projects and policies against the UN SDGs, we are offered the ability to assess our behaviour against the universally agreed goals that form the framework of our future.

PlatformOcean, and the two exemplars shown within this paper, have been designed not just with functionality in mind, but also considering both the intended positive and unintended impacts towards the SDGs. In particular, the design of plug-ins, and the use of computational intelligence (AI) within those plug-ins, could provide the basis for effective human-machine interaction and cooperation in socio-technical systems: on the one hand through the codification of deep social knowledge and self-organisation for sustainability, for example; and on the other hand measurement, impact assessment and feedback which contributes to the achievement of (or trade-off between) SDGs.

Future work must now strengthen the bridge the gap between digital innovation and achieving the SDGs, and using AI to measure impact is an important first step. Everyone and everything is interconnected: we can no longer hide from the wider impacts that our individual actions, or the algorithms that we develop, have on society, or the planet [26]. The Digital Society is everyone's future – but as everyone and everything is interconnected, that interconnection should be leveraged for the common interest of the collective, not for the personal profit of a privileged few.

Acknowledgment

We are grateful to the Association for Information Technology Trust (AITT) for support for project PlatformOcean. We also thank the organisers of the ECAI2020 AI4Eq Workshop, and the anonymous reviewers of various submissions for their many encouraging comments and helpful recommendations.

References

[1] B. Friedman, P. Kahn, A. Borning, "Value sensitive design and information systems," in *The Handbook of Information and Computer Ethics*, K. Himma, H. Tavani Eds., Wiley, 2008, pp. 69–101.

[2] K. Bellman, J. Botev, H. Hildmann, P. Lewis, S. Marsh, J. Pitt, I. Scholtes, S. Tomforde, "Socially-sensitive systems design: Exploring social potential," *IEEE Tech. Soc. Mag.*, vol. 36, no. 3, pp. 72–80, 2017.

[3] H. Rittel, M. Webber, "Dilemmas in a general theory of planning," *Policy Sciences*, vol. 4, no. 2, pp. 155–169, 1973.

[4] S. Zuboff, "Big other: surveillance capitalism and the prospects of an information civilization," *Journal of Information Technology*, vol. 30, pp. 75–89, 2015.

[5] H. Simon, "Rational choice and the structure of the environment," *Psychological Review*, vol. 63, no. 2, pp. 129–138, 1956.

[6] T. Hák, S. Janoušková, B. Moldan, "Sustainable development goals: A need for relevant indicators," *Ecological Indicators*, vol. 60, pp. 565 – 573, 2016.

[7] Common Fund for Commodities, "The sustainable develop-ment goals and impact measurement – a CFC journey." Annual Report, 2017.

[8] E. Ostrom, *Governing the Commons*. Cambridge Univ. Press, 1990.

[9] J. Ober, *Democracy and Knowledge*. Princeton Univ. Press, 2008.

[10] D. B. Kurka, J. Pitt, J. Ober, "Knowledge management for self-organised resource allocation," *TAAS*, vol. 14, no. 1, pp. 1:1–1:41, 2019.

[11] J. Ober, *Demopolis: Democracy before liberalism in theory and practice*. Cambridge Univ. Press, 2017.

[12] J. Pitt, J. Ober, "Democracy by design: Basic democracy and the self-organisation of collective governance," in *IEEE SASO*, 2018, pp. 20–29.

[13] B. Malinowski, *Argonauts of the Western Pacific: An account of native enterprise and adventure in the Archipelagoes of Melanesian New Guinea*. Routledge and Kegan Paul, 1922.

[14] J. Pitt, J. H. Clippinger, C. Sørensen, "Values, axial curren-cies, and computational axiology: Digital currencies can do more than buy stuff," *IEEE Technol. Soc. Mag.*, vol. 37, no. 3, pp. 56–63, 2018.

[15] N. Rescher, *Distributive Justice*. Bobbs-Merrill, 1966.

[16] A. Rychwalska, G. Goodell, M. Roszczynska-Kurasinska, "Data management for platform-mediated public services: Challenges and best practices," *CoRR*, vol. abs/1909.07143, 2019.

[17] J. Zittrain, The Future of the Internet – And How to Stop It. New Haven, CT: Yale University Press, 2008.

[18] T. Hardjono, P. Deegan, J. Clippinger, "Social use cases for the ID3 open mustard seed platform," *Technology & Society Magazine*, vol. 33, no. 3, pp. 48–54, 2014.

[19] S. Robert, H. Robert, W. Evans, D. Honemann, T. Balch, *Robert's Rules of Order, Newly Revised*, 10th ed. Perseus Publishing, 2000.

[20] A. Manjarrés, A. Pickin, M. A. Artaso, E. Gibbons, "Ai4eq: for a true global village not for global pillage," *Technology & Society Magazine*, vol. 40, no. 1, 2021.

[21] Food and Agriculture Organisation of the UN, "Global food losses and food waste – extent, causes and prevention." http://www.fao.org/3/mb060e/mb060e00.pdf.

[22] Food and Agriculture Organisation of the UN, "Food wastage footprint and climate change." http://www.fao.org/3/a-bb144e.pdf.

[23] C. Hanson, P. Mitchell, *The Business Case for Reducing Food Loss and Waste*. Washington, DC: Champions 12.3, 2017.

[24] E. Facchini, E. Lacovidou, J. Gronow, N. Voulvoulis, "Food flows in the united kingdom: The potential of surplus food redistribution to reduce waste," *Journal of the Air & Waste Management Association*, vol. 68, no. 9, pp. 887–899, 2018.

[25] D. Schiff, B. Rakova, A. Ayesh, A. Fanti, M. Lennon, "Explaining the principles to practices gap in AI," *Technology & Society Magazine*, vol. 40, no. 1, 2021.

[26] A. Luccioni, Y. Bengio, "On the morality of artificial intelligence," *IEEE Tech. Soc. Mag.*, vol. 39, no. 1, pp. 16–25, 2020.

Steph Pitt

After graduating with a BSc in Natural Science from the University of Bath (UK) in 2019, Steph worked on air quality research in North East Asia as an Environmental Affairs Intern at the United Nations ESCAP office in South Korea. Subsequently, she worked on improving social coordination within local plastic reduction and air quality movements as part of the PlatformOcean project at Imperial College London. Currently, Steph is studying for a Master's in Public Health Science at Karolinska Insitutet in Sweden, with a focus particular interest on the intersection between public health interventions and environmental sustainability.

Marlína van Meelis Lacey

Marlína van Meelis Lacey graduated with a BSc in Environmental Geography and International Development from The University of East Anglia (UK). She has worked in Cambodia for two NGOs: The Human Resource and Rural Economic Development Organisation (HURREDO) and Youth Coalition for Unity and Development (YCUD), and for the Raleigh Interna-tional Water, Sanitation and Hygiene (WASH) programme based in India. She has also worked as a drone pilot for an AI start-up and on the PlatformOcean project at Imperial College London.

Ed Scaife

Ed Scaife graduated with an MEng ACGI in Electronic and Information Engineering from Imperial College London (UK) in 2020. He was the principal developer of the prototype for the sustainable social media meta-platform PlatformOcean. Outside of this, his main academic interests are Computer Vision, Artificial Intelligence and Human-Centred Robotics. He currently works as a software and systems engineer in the automotive industry.

Jeremy Pitt

Jeremy Pitt is Professor of Intelligent and Self-Organising Systems in the Depart-ment of Electrical and Electronic Engi-neering at Imperial College London (UK). His research interests focus on developing formal models of social processes using computational logic, and their application in self-organising multi-agent systems for engineering cyber-physical and socio-technical systems. He has been an investigator on more than 30 national and European research projects and has published more than 150 articles in journals and conferences. He is a Fellow of the BCS and a Fellow of the IET, a member of the IEEE, and since 2018 has been Editor-in-Chief of IEEE Technology and Society Magazine.

Assessing Lexical-Semantic Regularities in Portuguese Word Embeddings

Hugo Gonçalo Oliveira[1], Tiago Sousa[2], Ana Alves[3]

[1] CISUC, Department of Informatics Engineering, Faculty of Sciences and Technology, University of Coimbra (Portugal)
[2] ISEC, Polytechnic Institute of Coimbra (Portugal)
[3] CISUC & ISEC, Polytechnic Institute of Coimbra (Portugal)

Received 23 November 2020 | Accepted 28 December 2020 | Published 19 February 2021

Abstract

Models of word embeddings are often assessed when solving syntactic and semantic analogies. Among the latter, we are interested in relations that one would find in lexical-semantic knowledge bases like WordNet, also covered by some analogy test sets for English. Briefly, this paper aims to study how well pretrained Portuguese word embeddings capture such relations. For this purpose, we created a new test, dubbed TALES, with an exclusive focus on Portuguese lexical-semantic relations, acquired from lexical resources. With TALES, we analyse the performance of methods previously used for solving analogies, on different models of Portuguese word embeddings. Accuracies were clearly below the state of the art in analogies of other kinds, which shows that TALES is a challenging test, mainly due to the nature of lexical-semantic relations, i.e., there are many instances sharing the same argument, thus allowing for several correct answers, sometimes too many to be all included in the dataset. We further inspect the results of the best performing combination of method and model to find that some acceptable answers had been considered incorrect. This was mainly due to the lack of coverage by the source lexical resources and suggests that word embeddings may be a useful source of information for enriching those resources, something we also discuss.

Keywords

Natural Language Processing, Computational Semantics, Word Embeddings, Lexical Semantics, Analogy.

DOI: 10.9781/ijimai.2021.02.006

I. Introduction

Two main approaches have been followed for representing the words of a language according to their semantics: lexical-semantic knowledge bases (LKBs), such as wordnets [1]; and distributional models, like word embeddings. The former organise words and their meanings, often connected by explicit relations, such as Hypernymy or Part-of, and may include additional lexicographic information (part-of-speech, gloss). On the other hand, the latter follow the distributional hypothesis [2], which says that words that occur in the same contexts tend to convey similar meanings, and represent words as vectors of numeric features, according to the contexts they are found in large corpora. On distributional models, since 2013 the trend was to use efficient methods that learn dense-vector representations of words, like word2vec [3] or GloVe [4]. Besides their utility for computing word similarity, e.g., with the cosine similarity of the vector representations, such models are known for preserving several linguistic regularities, and have shown very interesting results when solving analogies of the kind "*what is to b as a* is to a*"? (e.g., what is to Portugal as Paris is to France?). So much that both previous tasks are extensively used for assessing word embeddings in different languages.

Popular analogy test sets cover syntactic and semantic relations of different types, from word inflections and derivations, to word knowledge relations like capital-country. Yet, we are interested in studying relations between word meanings that one would find, implicitly, in a language dictionary or, explicitly, in a LKB. Given that they connect general-language words according to their meanings, we refer to them as lexical-semantic relations. More precisely, our goal with this work is twofold. We aim to:

- Assess how lexical-semantic relations are preserved by Portuguese word embeddings;
- Analyse to what extent analogy solving methods could be useful for enriching LKBs.

Towards our goal, we needed an analogy test targeting lexical-semantic relations in Portuguese, which we created as described in this paper. It was baptised as *Teste para Analogias Léxico-Semânticas* (TALES, in English, Test for Lexical-Semantic Analogies) and is exclusively focused on these relations. Although some analogy tests already cover lexical-semantic relations [5], [6], they are for English and, while they could have been translated to Portuguese, as the Google Analogy Test was [7], we decided to create a new test from scratch, because different languages represent different socio-cultural realities, they do not cover exactly the same part of the lexicon and, even where they seem to be common, several concepts are lexicalised differently [8]. This is important because, besides assessing word embeddings, TALES can provide training data for relation discovery in word embeddings, potentially useful for augmenting Portuguese LKBs, such as Portuguese wordnets [9].

* Corresponding author.

E-mail address: hroliv@dei.uc.pt

TALES follows a similar format to the English BATS test [5] and covers different types of lexical-semantic relation, with the same number of entries, 50, for each, which makes it a balanced test. The entries of TALES were selected based on their presence in several Portuguese lexical resources and their frequency in a corpus. We attempted at solving the lexical-semantic analogies of TALES by applying classic and more recent analogy solving methods [10] to pretrained word embeddings available for Portuguese. This included static word embeddings (word2vec and GloVe) but also static representations obtained from recent BERT [11] neural language models. As it happens for the lexical-semantic relations in BATS, accuracies are low, even if some relations are more challenging than others. However, in opposition to some relation types, namely syntactic and world knowledge relations, several entries in TALES have many acceptable answers (e.g., a hypernym generally has several hyponyms). And even though the adopted BATS format enables the inclusion of several answers, in many cases they are too many and it is just not possible to get them all from the lexical resources. Therefore, incorrect answers may include relations that are just missing, which makes word embeddings, potentially, a useful source of information for enriching those resources. Having this in mind, we analyse some of the results obtained and discuss this possibility. Indeed, missing examples were found for every relation type covered by TALES. Even if, for some types, these were a minority of cases, for others they represented almost 20%of the answers considered incorrect.

The remainder of the paper is structured as follows: next Section (II) overviews related work on available test sets for assessing word embeddings, in English and Portuguese, as well as some work on the automatic creation and enrichment of LKBs; Section III describes the creation of TALES, including all the decisions taken in the process, and shows examples of its contents; Section IV describes the models of pretrained word embeddings used and the analogy solving methods applied in our experimentation; Section V reports on the performance of solving the lexical-semantic analogies of TALES with word embeddings, using the different methods, also looking at the performance per-relation; before concluding, Section VI starts by analysing incorrect answers that would be acceptable and discusses the utility of word embeddings when it comes to enriching structured lexical resources.

II. Related Work

The quality of word embeddings is typically assessed with two kinds of test: word similarity and analogy. The former contain pairs of words and a score proportional to their semantic similarity. Given two words, scoring their semantic similarity becomes a matter of computing the cosine of their vectors. The correlation between the computed scores for all the pairs in the test and the ground-truth scores may then be measured for evaluation, i.e., the higher the correlation, the better the performance.

Popular tests of this kind, for English, include WordSim-353 [12] and SimLex-999 [13]. WordSim-353 contains 353 word pairs and their relatedness score (0-10), based on the judgement of 13 to 16 human judges. Due to the known differences between similarity and relatedness, WordSim-353 was later [14] manually split into similar and related pairs. For this purpose, semantic relations between the words of the pair were identified, and pairs were split into: similar (synonyms, antonyms, identical, or hyponym-hyperonym); related (meronym-holonym); none of the previous relations but average similarity higher than 5; unrelated (remaining pairs). SimLex-999 contains 999 word pairs (666 noun-noun, 222 verb-verb, 111 adjective-adjective) and their similarity score, based on the opinion of ≈50 judges. This is the only test where judges were specifically instructed to differentiate between similarity and relatedness and rate regarding the former only. Its authors thus claim that it targets genuine similarity.

Anyway, despite the post-annotations in WordSim-353, relatedness scores are only a number that represents a strength, but tells nothing about the actual relation between the words or concepts they denote. To go further in distributional models, one may resort to analogies, i.e., look for pairs of words that are related similarly to a known pair of words or a set of pairs. When presenting word2vec, evaluation used what became known as the Google Analogy Test (GAT) [3]. It has analogies of the kind *a is to* a^* *as b is to* b^*, split between nine syntactic (e.g., adjective to adverb, opposite, comparative, verb tenses) and five semantic categories (e.g., capital-country, currency, male-female), with 20–70 unique example pairs per category, which may be combined in 8,869 semantic and 10,675 syntactic questions.

BATS [5] is a broader alternative to GAT, balanced between four types of relation – grammatical inflections, word-formation, lexical-semantic and world-knowledge relations –, with 10 categories of each type and 50 word pairs per category (overall 2,000 unique word pairs). Moreover, BATS enables more than one possible answer for each question, which makes sense for some relation types (e.g., a hypernym will have more than one hyponym). The lexical-semantic relations in BATS were acquired from Princeton WordNet [1], a LKB where word senses are grouped in synonym sets, and semantic relations are established between the latter.

Experiments using BATS have shown that some categories are more challenging than others, and lexical-semantic relations are among those with lower accuracy. This also motivated the experimentation with alternative methods that consider more than one example for solving analogies, namely 3CosAvg and LRCos (see sub-section B of Section V).

DiffVec [6] is another dataset for evaluating word embeddings. It covers 15 relation categories, including both grammatical (8) and lexical-semantic relations (7), obtained from several sources. Specifically, lexical-semantic relations were obtained from SemEval-2012 task 2 [15] and from the BLESS dataset [16]. With 12,458 questions in total, it is larger than GAT and, although covering less categories, also larger than BATS, but imbalanced.

Performance on analogy tests is typically measured with accuracy, i.e., the proportion of answers that match the expected word. Though, some researchers also assessed this task in a retrieval or classification scenario [17], i.e., quantifying how many correct answers could be retrieved. For this purpose, measures like precision, recall, or Mean Average Precision (MAP) were used.

For assessing Portuguese word embeddings, some of the previous tests were translated to Portuguese [7], namely WordSim-353, SimLex-999 and GAT. Several approaches were tested for answering WordSim-353 and SimLex-999 [18], including knowledge and distributional approaches. GAT has been used for assessing Portuguese Word Embeddings [19] and, more recently, was translated to the BATS format [20], which enabled the application of alternative methods for analogy solving.

Another related dataset for Portuguese is B^2SG [21], which targets semantic relations, but has a different structure. It is similar to the Test Of English as a Foreign Language (TOEFL), but based on the Portuguese part of BabelNet [22], and was partially evaluated by humans. B^2SG contains frequent Portuguese nouns and verbs (target), each followed by four candidates, from which only one is related, and is organised in six files: two for synonymy, two for hypernymy, and two for antonymy, between nouns and between verbs, respectively. An important difference to the analogy tests is that B^2SG narrows the possible answers to the four candidates.

Back to the analogy tests, we believe that, besides assessing word embeddings, they can be useful for developing models of relation discovery in the embedding space, especially considering lexical-semantic analogies, which often have more than one acceptable answer. More precisely, models trained in analogy tests could be useful for creating or enriching knowledge bases. The goal would be similar to earlier attempts for extracting relations from dictionaries [23], or from raw corpora, having in mind the enrichment of LKBs like WordNet [1], and tackled with handcrafted patterns [24], or patterns learned with weakly-supervised approaches, for extracting hypernymy [25] and other relations . The latter approaches would start with known seeds, which could be acquired from WordNet itself. An alternative way of enriching LKBs, which are focused on lexical knowledge, is to extend them with world knowledge, e.g., by linking them with Wikipedia, as in the BabelNet project [22]. For Portuguese, on this scope, Onto.PT is a wordnet [27] that combines information in existing thesauri with relations extracted from several Portuguese dictionaries [28].

III. Creating the TALES Test Set

In order to assess to what extent lexical-semantic relations are preserved in Portuguese word embeddings, we first needed a benchmark. For this purpose, we created a test set, dubbed TALES, that could be used in a similar way to other popular analogy test sets. This section describes the most important decisions taken in the creation of this test, starting with the adopted data format, target relations, and ending with decisions specifically concerning some relation types.

A. Data Format

We opted to represent TALES in a format similar to BATS, where included files have entries like those in Fig. 1. Specifically, for each covered relation, there would be a file where each row corresponds to an entry and has two-columns: one with a word, to be used in the formulation of a question (*b*), and another with one or more words, to be used as the target answers (b^*). We recall that an analogy can be formulated as 'what is to *b* as a^* is to *a*', for which the answer is b^*. Considering the BATS entries in Fig. 1, possible questions would be: *what is to cat as reptile is to rattlesnake?* (i.e., Hypernym-of cat), or *what is to citrus as turtleneck is to sweater?* (i.e., Hyponym-of citrus). We also note that, besides direct relations, BATS includes inherited relations in the possible answers, such as the inherited hypernyms in the first entries in Fig. 1.

As it happens in BATS, but not in GAT, when there is more than one possible answer, they are all included in the second column, split by '/'. This is relevant, especially in the context of lexical-semantic relations. For instance, a hypernym should have several hyponymys, or an object might have several parts. Also, as in BATS, we split the test into different files, one for each relation covered. Each file has the same number of entries, 50, which means that TALES is balanced between all of the relations covered.

B. Target Relations

For selecting the relation types to include in TALES, we initially targeted the more common types in wordnets, also included in BATS [5], namely Hypernymy, Meronymy, Synonymy and Antonymy. We then looked at relations of those and other types in a large set of relations extracted from ten lexical resources for Portuguese [29], covering both the European and the Brazilian variant[1], and at the number of instances of each kind in more than one resource. The number of resources that a relation instance is found in, hereafter *r*, can be seen as an indicator of its consensus, utility and, indirectly, of its quality, i.e., given that most of the exploited resources had some automatic step in their creation, *r* can also be used for avoiding incorrect relations.

When looking at available relations and how they were organised, we first decided to split synonymy in three types – Synonymy_n, between nouns, Synonymy_v, between verbs, and Synonymy_adj, between adjectives – and Hypernymy in two – Hypernymy_n, between nouns, and Hypernymy_v, between verbs. We further decided to use Antonymy and Meronymy, though only one type of each: Antonymy between adjectives, for being the most representative, and Part-of for Meronymy, because it was the only type for which there were enough instances (see sub-section C). Finally, we also found enough Purpose-of relation instances and decided to included this type as well.

C. Instance Selection

Once we had decided on target relations, we wanted to select the most consensual 50 instances of each selected type. These would be the 50 instances of each type with highest *r*. Yet, in most cases there would be ties, i.e., more than 50 instances had the same *r*. So, we also ranked instances by the frequency of their first argument (first column, to be used as *b*) in CETEMPúblico [38], a Portuguese corpus of news. As corpus frequency is an indicator of the commonality / usage frequency of words, it is also relevant for selecting words to include. Therefore, we only considered instances where the first argument occurred at least 100 times in CETEMPúblico[2]. After this, not enough Member-of and Material-of relations were left, which is the main reason for our test covering only Part-of, whereas BATS covers three types of Meronymy, the same as in WordNet [1]: Part, Member and Substance.

Despite being strict with the first relation argument, we dropped the frequency constraints for the second argument (second column), which we recall could be more than one, and relaxed the *r* constraint for all but the first word. For the remaining words, the only constraint was that they occur in a relation of the target type with the first argument, in at least two resources ($r = 2$). Since some of the lexical resources

[1] These resources were PAPEL [28], Dicionário Aberto [30], Wiktionary.PT [31], TeP [32], OpenThesaurus.PT, OpenWordNet-PT [33], PULO [34], WordNet.Br [35], Port4Nooj [36] and ConceptNet [37].

[2] CETEMPúblico was used only for ranking and filtering, based on the first argument of each relation instance, while all words still came from the lexical resources. In fact, we did not use CETEMPúblico directly, only the frequency lists available from AC/DC [39].

cat	feline/beast/animal/organism/fauna/placental/ carnivore/chordate/felid/eutherian/mammal/mammalian/...
hawk	raptor/bird/vertebrate/creature/beast/being/animal/organism/fauna/chordate/animate_being/craniate/...
rattlesnake	snake/reptile/pit_viper/serpent/ophidian
church	chapel/abbey/basilica/cathedral/duomo/kirk
citrus	lemon/orange/lime/mandarin/tangerine/yuzu
sweater	turtleneck/cardigan/pullover/slipover/turtle/polo-neck

Fig. 1. Example entries in a BATS files for 4_Lexicographic_semantics, namely *L01 [hypernyms - animals]* (first three lines) and *L03 [hyponyms - misc]* (last three lines).

considered included relations extracted from dictionaries, possibly not so common, and others were created automatically, setting $r = 2$ minimises the number of incorrect or unuseful relations. At the same time, this may contribute to lower Mean Average Precision with some models (see examples in Section VI).

D. Non-symmetrical Relations

With initial experiments, we noticed that, in non-symmetrical relations, the challenge was different, depending on whether we were using direct (e.g., vehicle Hypernymy-of car) or inverse relations (car Hyponymy-of vehicle). This is mainly due to the fact that, in some directions, it is more common to have many possible answers. As mentioned earlier, a hypernym will have several hyponyms, but a hyponym will often have a single (direct) hypernym. Or, something can be part of different things (e.g., blade part-of knife, axe, sower) or have different parts (e.g., parts of the body). Therefore, for each non-symmetrical relation, we created two different files, one with direct and another with inverse relations. In the latter, the order of the arguments was switched in the original relation set, which then went through the automatic creation process, including the application of the aforementioned constraints to the argument that then became the first. Since the switch was made in the original relation set, the instances in the file of direct relations are not necessarily the inverse of those in the direct.

E. Hypernymy and Concreteness

After Synonymy, Hypernymy_n is the second relation for which we had more instances, so we decided to further split them into more coherent sets. In BATS, there is a file for Hypernymy, another for its inverse, Hyponymy, and a third file for Hypernymy between animals only. For TALES, we did not create a file for a single class, but looked at another property of words: concreteness, i.e., the degree to which words refer to objects, persons, places, or things that can be experienced by the senses [40]. So, we split the Hypernymy relations, direct and inverse, roughly into concrete (+concrete) and not concrete / abstract (-concrete). Concreteness values were obtained from the Minho Word Pool [41], where 3,800 Portuguese words have assigned values of concreteness and imageability, between 1 (minimum) and 7 (maximum). In this case, we empirically set that concrete words would have a minimum concreteness value of 6 (covering e.g., house, ball, money), whereas abstract would have 4.5 or less (covering e.g., age, space, energy). Again, to maximise the number of acceptable answers, this constraint was only applied to the first argument. Still, it is expectable that concrete concepts do relate with more concrete concepts and less concrete with less concrete concepts.

F. Test Set Characterisation

Table I characterises TALES, the resulting test. It lists the relation types covered and their direction (D for direct, I, for inverse), the minimum *r* (higher for relations for which there were more instances) applied to the first-column argument, and examples of included relations, in Portuguese, with a rough English translation. As in BATS, for entries with more than one acceptable answer, the second argument has each possible answer split by '/'.

TABLE I. Characterisation of the Generated Lexical-semantic Relations Test

Relation		***r***	**Examples**
Synonym-of_n		7	(*local, sítio*) (*proposta, alvitre/sugestão/proposição*) (location, site), (proposal, suggestion/proposition)
Synonym-of_v		8	(*existir, viver/durar/...*) (*ouvir, perceber/entender/escutar/...*) (exist, live/last), (listen, feel/understand)
Synonym-of_adj		7	(provisório, provisional/temporário) (rural, rústico/pastoril/...) (provisional, temporary), (rural, rustic/pastoral)
Antonym-of_adj		5	(*estreito, largo*) (*velho, jovem/novo/moço*) (narrow, wide), (old, young/new/lad)
Hypernym-of_n (+concrete)	D	4	(*fruto, morango/ameixa/...*) (*veículo, jipe/monovolume/...*) (fruit, strawberry/plum), (vehicle, jeep/minivan)
	I	4	(*carro, veículo*) (*perna, suporte/segmento/membro/apoio*) (car, vehicle), (leg, support/segment/member)
Hypernym-of_n (-concrete)	D	4	(*regra, restrição/lei/etiqueta/...*) (*questão, pergunta/problema/...*) (rule, restriction/law/etiquette), (query, question/problem)
	I	4	(*futuro, tempo*) (*orgulho, satisfação/sentimento*) (future, time), (pride, satisfaction/feeling)
Hypernym-of_v	D	3	(*vir, chegar/desembarcar/cair*) (*contar, relatar/somar*) (come, arrive/land/fall), (count, report/sum)
	I	3	(*querer, ordenar/exigir*) (*pagar, subornar/dar/corromper*) (want, order/demand), (pay, bribe/give/pervert)
Part-of	D	2	(*mês, ano*) (*sala, casa/prédio/domicílio/edifício/habitação/...*) (month, year), (room, house/building/home)
	I	2	(*água, oxigénio/hidrogénio*) (*palavra, sílaba*) (water, oxygen/hydrogen), (word, syllable)
Purpose-of	D	3	(*levantar, guindaste*) (*desenhar, lapiseira/caneta/lápis/sombra/...*) (rise, crane), (draw, pencil/pen/shadow)
	I	3	(*lixa, polir*) (*fogão, aquecer/cozinhar*) (sandpaper, polish), (cooker, heat/cook)

As nothing was done to avoid semantic ambiguity, it is common to mix different senses of the same word, some of them figurative. Yet, we do not see this as a problem. First, static word embeddings (e.g., word2vec, GloVe) also have a single vector per word, thus ignoring word senses. Second, in most cases, there are several acceptable answers, which might apply for different senses of the first argument. Such an example is the word *perna* (leg), for which four hypernyms are possible: *suporte/apoio*, related with the 'support' meaning, and *membro/segmento*, related to the 'limb' meaning.

IV. Experimentation Setup

TALES can be used for assessing Portuguese word embeddings, specifically, their ability to capture lexical-semantic relations. For this purpose, we first used three pretrained models of static word embeddings for Portuguese, where four methods were applied to solve TALES. Moreover, to embrace recent trends, we decided to test as well embeddings produced by pretrained neural language models, namely BERT [11]. For loading the embeddings and performing the tests, we used the Vecto[3] package, which supports analogy tests in the previously described BATS format, adopted by TALES, and includes the implementation of different analogy solving methods. This section describes the models and methods used in our experimentation in more detail.

A. Models of Word Embeddings

The analogy solving methods were first applied to three pretrained models of static word embeddings, all with 300 dimensions, but covering different learning algorithms, namely GloVe, word2vec CBOW and word2vec SKIP-GRAM. These models are part of NILC embeddings [19], a set of pretrained word embeddings for Portuguese, freely available for download[4].

However, the current trend in language representation are neural language models, like BERT [11], which rely on Transformer neural networks for encoding words and longer sequences in meaningful embedding vectors. An important difference towards static word embeddings is that BERT provides contextual word representations, meaning that, depending on its surrounding context, the same word might be represented by different vectors. Since, like any analogy test, the entries of TALES lack context and do not handle different senses of the same word, we were unsure whether we could take advantage of the contextual features of the previous models. Yet, recent work has showed that contextualised representations in a given layer (apparently, the lower, the better) can outperform static word embeddings in analogy solving [42]. Therefore, we decided to apply the analogy solving methods also to word representations by two BERT models covering Portuguese, namely: the Multilingual Cased BERT-Base model by Google[5], which includes Portuguese among 104 languages; and BERTimbau-Large, pretrained exclusively for Portuguese [43]. The former has 12 layers and encodes word sequences in 768-size vectors, the size of its hidden layers, while the version of BERTimbau used has 24 layers and encodes word sequences in 1,024-sized vectors.

Since we were dependent on Vecto for running the tests, and Vecto is only prepared to deal with static word embeddings, we had to convert BERT representations to a static format. With the help of the bert-as-a-service[6] tool, this conversion was made in three steps: (i) running through all the entries in the vocabulary of each BERT model, which includes words and subwords (word pieces); (ii) retrieve their representation in a selected layer of the model; (iii) use the resulting vector as the static representation of the vocabulary entry. BERT provides contextualised representations, but there is no context in the questions of analogy tests, therefore, we simply tested representations obtained from different layers and present results for: the first and the second, which should be less context-specific [42], [44]; and also the second to last, because the last layer is too close to the target functions during pretraining and may be biased to those targets. We also note that, with this adaptation, we might not be taking the most out of BERT. The main issue is related to the vocabulary coverage, which we are limiting to the entries in BERT's vocabulary file – 119,547 for BERT-ML and 29,794 for BERTimbau –, when we know that BERT relies on the WordPiece tokeniser and represents several words with the combination of two or more entries (subwords). At the same time, many subwords are used for obtaining inflections, which are scarce in the target lexical resources and thus in TALES. We leave alternative approaches on handling BERT for consideration in future work.

B. Analogy Solving Methods

In order to solve TALES, four different methods were applied to the selected models of word embeddings, all with implementation available in Vecto. For each method, Vecto outputs a JSON report with information on each question, including a ranked list of candidate answers, a summary of the experimentation setup, and the accuracy of the test, computed from the first answer of each rank.

The first method, Similar-to-B (eq. 1), is often used for retrieving similar words, based on the cosine similarity of their vectors. Though not exactly an analogy-solving method, due to its simplicity, it has been used as a baseline [45] for this purpose. In fact, achieving the best accuracy with Similar-to-B means that more complex analogy solving methods are not doing any good.

$$b^* = \underset{w \in V}{\operatorname{argmax}} \cos(b, w) \tag{1}$$

The second method, vector offset [3], was originally used for solving analogies with word2vec, and later became also known as 3CosAdd (eq. 2). It formulates the analogy as *a is to* a^* *as b is to* b^*, where b^* has to be inferred from a, a^* and b.

$$b^* = \underset{w \in V}{\operatorname{argmax}} \cos(w, a^* - a + b) \tag{2}$$

Instead of considering only the word b (Similar-to-B) or this word plus a single pair of analogously-related words (a, a^*), the remaining two methods, both proposed by Drozd et al. [10], try to make the most out of the full test set. 3CosAvg computes the average offset between words in position a and respective words in position a^*, in a set of relations of the target type (eq. 3). The answer, b^*, must maximise the cosine with the vector resulting from summing the average offset to b.

$$b^* = \underset{w \in V}{\operatorname{argmax}} \cos(w, b + avg\_offset) \tag{3}$$

The final method tested is LRCos (eq. 4), which considers the probability that a word w belongs to the same class as other words in position a^*, as well as the similarity between w and b, measured with the cosine. Although any classification algorithm could be used for this, the default implementation of LRCos, used by us, relies on logistic regression for computing the likelihood of a word belonging to the class of words a^*.

$$b^* = \underset{w \in V}{\operatorname{argmax}} P(w \in target\_class) * cos(w, b) \tag{4}$$

We also experimented with other methods available for this purpose, namely 3CosMul and PairDirection, but concluded that they would not add much, and so left their results out of this paper. Specifically, accuracy of PairDirection was often 0 or very close.

[3] https://github.com/vecto-ai

[4] http://nilc.icmc.usp.br/embeddings

[5] https://github.com/google-research/bert

[6] https://github.com/hanxiao/bert-as-service

V. Experimentation Results

We tackled the challenge of solving the questions in TALES by applying the four methods described in sub-section B of Section IV — Similar-to-B (SIM), 3CosAdd (3CAD), 3CosAvg (3CAV), LRCos (LRC) — to the five models of word embeddings introduced earlier — GloVe, word2vec CBOW, word2vec SKIP-GRAM, BERT-ML, BERTimbau.

This section reports on the results of this experimentation. Besides revealing the accuracy of different methods in different models of embeddings, for different relations, performed experiments provide useful insights on the potential of this approach for discovering new relations, which is further discussed in section VI. To help us reach some conclusions, we first look at the overall performance of different configurations, measured with the accuracy and MAP@10, and then at the performance per relation.

A. Overall Accuracy

Table II has the overall performance of each method in the static word embeddings, considering all the 14 relations, only the symmetrical (synonymy and antonymy), and only the non-symmetrical, in terms of accuracy and MAP@10. Tables III and IV have similar information, respectively for the representations obtained from three different layers of the two BERT models used.

Given that TALES is balanced between the 14 relations, each in a different file with 50 entries, these are averages of the performance for each relation. Accuracy is given by the proportion of entries (*b*) for which the first answer given (*b**) was correct (i.e., it was one of the words in the second column of the entry for *b*). The MAP@10 considered not only the first answer, but the top-10 answers given by Vecto for each question.

TABLE II. Performance of Static Word Embedding Models Through Different Methods in TALES

	GloVe				**word2vec-CBOW**				**word2vec-SKIP**			
	SIM	**3CAD**	**3CAV**	**LRC**	**SIM**	**3CAD**	**3CAV**	**LRC**	**SIM**	**3CAD**	**3CAV**	**LRC**
						Accuracy						
Symmetrical	0.19	0.08	0.19	0.14	**0.22**	0.10	**0.22**	0.15	**0.22**	0.10	**0.22**	0.14
Non-Symmetrical	0.07	0.04	0.08	**0.12**	0.08	0.04	0.09	0.07	0.07	0.04	0.09	0.09
All	0.10	0.05	0.11	**0.13**	0.12	0.05	0.12	0.10	0.11	0.06	0.12	0.10
						MAP@10						
Symmetrical	0.28	0.14	**0.29**	0.21	**0.29**	0.15	0.28	0.20	0.28	0.15	0.26	0.20
Non-Symmetrical	0.16	0.08	0.16	**0.18**	0.11	0.06	0.12	0.11	0.12	0.07	0.13	0.13
All	0.19	0.10	**0.20**	0.19	0.16	0.09	0.17	0.14	0.16	0.09	0.17	0.15

TABLE III. Performance of Word Embeddings From Different Layers of BERT-ML Through Different Methods in TALES

	Layer 1				**Layer 2**				**Layer 11**			
	SIM	**3CAD**	**3CAV**	**LRC**	**SIM**	**3CAD**	**3CAV**	**LRC**	**SIM**	**3CAD**	**3CAV**	**LRC**
						Accuracy						
Symmetrical	0.02	0.00	0.02	0.01	0.02	0.00	0.02	0.01	0.02	0.00	0.02	0.01
Non-Symmetrical	0.00	0.00	0.00	0.02	0.00	0.00	0.00	0.02	0.00	0.00	0.00	0.02
All	0.01	0.00	0.01	0.02	0.01	0.00	0.01	0.02	0.01	0.00	0.01	0.02
						MAP@10						
Symmetrical	0.03	0.00	0.03	0.02	0.03	0.00	0.02	0.01	0.03	0.00	0.03	0.01
Non-Symmetrical	0.01	0.00	0.01	0.04	0.01	0.00	0.01	0.04	0.01	0.00	0.01	0.03
All	0.02	0.00	0.02	0.03	0.02	0.00	0.02	0.03	0.01	0.00	0.02	0.02

TABLE IV. Performance of Word Embeddings From Different Layers of BERTimbau Through Different Methods in TALE.

	Layer 1				**Layer 2**				**Layer 23**			
	SIM	**3CAD**	**3CAV**	**LRC**	**SIM**	**3CAD**	**3CAV**	**LRC**	**SIM**	**3CAD**	**3CAV**	**LRC**
						Accuracy						
Symmetrical	0.03	0.00	0.03	0.03	0.03	0.01	0.03	0.02	0.05	0.01	0.05	0.03
Non-Symmetrical	0.00	0.00	0.01	0.05	0.00	0.00	0.01	0.03	0.05	0.01	0.05	0.03
All	0.01	0.00	0.01	0.04	0.01	0.00	0.01	0.03	0.05	0.01	0.05	0.03
						MAP@10						
Symmetrical	0.07	0.02	0.07	0.03	0.07	0.02	0.07	0.03	0.08	0.02	0.08	0.03
Non-Symmetrical	0.02	0.01	0.02	0.07	0.02	0.01	0.02	0.06	0.08	0.02	0.08	0.04
All	0.03	0.01	0.04	0.06	0.03	0.01	0.04	0.05	0.08	0.02	0.08	0.04

When the word b is not in the embeddings vocabulary, the question is not answered. We consider these cases the same as giving an incorrect answer. While this would not have much impact on the comparison of GloVe and both word2vec models, which were learned from the same corpus and are expected to cover the same vocabulary, it is not the case of the BERT embeddings. So, this is required for making comparison fairer. We also note that, following its definition, the figures for 3CosAdd imply not 50 but 2,450 questions (50×49), because they are based on averages of using each of the 50 entry pairs as $b : b^*$ when each of the remaining 49 entries is used as $a : a^*$.

All methods were used with default parameters of the Vecto implementation. This means that for LRCos, the logistic regression classifier was trained with 49 positive pairs (one from each entry, i.e., a and the first a^*, except the target one) and 49 negative pairs (each with two arguments from different entries, i.e., a is from an entry and a^* is from another, meaning that they are probably not related, at least not as the positive examples).

The main conclusion is that TALES is a very challenging test. Accuracies are way under the best figures for syntactic and semantic analogies using the same embeddings (i.e., between 40 and 60% [19], [20]). Yet, a similar situation happens for English, on the BATS dataset [10], where best accuracies for lexical-semantic relations are always below 30%, with the single exception for the opposites with GloVe.

Another important conclusion is that we could not improve the performance with the embeddings obtained from BERT. Accuracy is so low that the differences between different layers are minimal if any. As mentioned in sub-section A of Section IV, context is not used for this task, and we thus could not take advantage of this feature of BERT. Yet, the main negative impact should result from limiting the word coverage to the entries in BERT's vocabulary. Still, even when questions with out-of-vocabulary (OOV) words are ignored, these accuracies are still significantly below the best with the static word embeddings (e.g., highest accuracy would be 0.12, achieved with Similar-to-B in any layer of BERT-ML for the symmetrical relations, followed by 0.09 with LRCos in the first layer of BERT-PT for non-symmetrical relations).

Considering all relations, the method+model configuration with the best accuracy was LRCos+GloVe (13%), but by the minimal margin of a single percentage point. On the other hand, for the symmetrical relations, the highest accuracies are achieved with 3CosAvg and with the Similar-to-B baseline in both word2vec models. This happens because both synonymy and antonymy occur between similar concepts, for which this baseline is already a good estimation. For synonymy, we can say that there are no benefits of using more sophisticated methods. This result is an important contribution to the overall accuracy of Similar-to-B. On the other hand, for non-symmetrical relations, LRCos+GloVe is not only the most accurate configuration but also the only with an average accuracy higher than 10% in this scenario, suggesting that, despite its limitations, LRCos suits this kind of relation better. At least when applied to GloVe, because in word2vec LRCos performs better for the symmetrical relations.

We note that the method originally applied for solving analogies in word2vec [3], 3CosAdd, is generally the one with worst performance, worse than Similar-to-B. This is also a consequence of how accuracy is computed for this method, which predicts b^* from a single pair $a : a^*$. Although this might work well for some relations, for the target ones, results show that it normally does not.

Together with other pretrained models, the static models used here have previously been used for solving analogies of different types, in Portuguese, with 3CosAdd [19] and, more recently, also 3CosAvg and LRCos [20]. For those attempts using 3CosAdd, it was always clear that GloVe was the most accurate model for semantic analogies. On syntactic analogies, it was generally outperformed by fastText-SKIP, which deals better with regular word terminations. Yet, for attempts using LRCos, the best method, GloVe was the best model for both semantic and syntactic analogies [20]. We note that, for all of those analogies, relations are not symmetrical. Therefore, even if, in our work, the selection of the best method and model could raise some doubts, based on previous work, we can say that the LRCos+GloVe combination is the best option for solving analogies.

This is also consistent with related research for English [6], [10], [17], where GloVe is often used for this purpose, and the methods that use more instances as reference (3CosAvg and LRCos) perform better than those that try to solve the analogy based on a single instance (3CosAdd). Yet, even if the previous works for English considered the synonymy and antonymy relations, they did not include the Similar-to-B baseline in their comparison. According to our experiments, that baseline could perform better than the other methods, thus constituting an exception in the preference for LRCos, especially if the embeddings are learned with word2vec.

B. Overall MAP@10

Although accuracy has been extensively used by others for assessing word embeddings in analogy solving [3], [10], this is a limited metric, because it does not discriminate between methods that still rank the correct answer high, and were thus closer to be correct, and methods that gave it a lower rank. This is especially important when tests include questions with more than one acceptable answer. For this task, ranking can be considered by adopting retrieval-based measures like precision and recall, with a threshold on the similarity score, or the Mean Average Precision (MAP) [17]. Therefore, towards a different perspective on evaluation, we also computed the MAP@10. This also had in mind the future exploitation of the methods used for improving TALES or, better, lexical resources in general, with new relations discovered (see Section VI).

As expected, MAP is higher than accuracy but, for most relations, it is not substantially higher. This means that, even if not that many, there are indeed correct answers ranked between second and tenth. Nevertheless, MAP scores support the idea that GloVe is a consistent model, not only for non-symmetrical analogies, using LRCos, but also for the symmetrical, using 3CosAvg or simply Similar-to-B. And it is for the non-symmetrical where differences towards other models are more clear.

C. Per-Relation Performance

Tables V, VI and VII present the MAP@10 for each relation with each method+model configuration, respectively for the static word embeddings, for different layers of the BERT-ML model, and for different layers of the BERT-PT model. Results make it clear that some relations pose different challenges than others. For instance, following the discussion in sub-section A, the Similar-to-B baseline outperformed all the other methods for Synonym-of. Though, when applied to different models, it also becomes clear that Similar-to-B is not so good for Antonymy, as this method is outperformed by 3CosAvg and LRCos in all the static word embeddings. This helps us narrow down the exceptions where Similar-to-B is enough for symmetrical relations to only synonymy. In fact, their high overall performance for symmetrical relations, in Table II, was influenced by the presence of three types of synonymy and only one of antonymy.

The best MAP for synonymy between nouns (0.27) and between verbs (0.37) was achieved in word2vec-CBOW, though the latter was tied with word2vec-SKIP, always with Similar-to-B. Between adjectives, the best MAP (0.25) was in GloVe, this time tied with Similar-to-B and 3CosAvg. For Antonym-of, the best MAPs resulted from applying the LRCos method to word2vec-SKIP (0.30) and to GloVe (0.29).

TABLE V. MAP@10 for Different Relations, with Static Word Embedding Models and Different Methods

Relation		GloVe				word2vec-CBOW				word2vec-SKIP			
		SIM	3CAD	3CAV	LRC	SIM	3CAD	3CAV	LRC	SIM	3CAD	3CAV	LRC
Synonym-of_n		0.23	0.12	0.25	0.13	0.27	0.15	0.26	0.08	0.25	0.14	0.25	0.13
Synonym-of_v		0.34	0.15	0.33	0.27	0.37	0.19	0.34	0.26	0.37	0.18	0.33	0.23
Synonym-of_adj		**0.25**	0.11	**0.25**	0.13	0.23	0.13	0.24	0.15	0.22	0.10	0.19	0.10
Antonym-of_adj		0.25	0.16	0.27	0.29	0.24	0.14	0.25	0.26	0.22	0.17	0.24	**0.30**
Hypernym-of_n	D	**0.20**	0.07	0.17	0.07	0.19	0.08	0.19	0.05	0.15	0.07	0.15	0.06
(+concrete)	I	0.18	0.15	0.25	**0.29**	0.15	0.09	0.20	0.19	0.14	0.09	0.16	0.22
Hypernym-of_n	D	**0.19**	0.07	0.16	0.12	0.17	0.08	0.17	0.07	0.18	0.08	0.17	0.13
(-concrete)	I	0.11	0.07	0.10	0.16	0.07	0.04	0.06	0.08	0.10	0.07	0.11	0.12
Hypernym-of_v	D	0.17	0.09	0.14	0.16	0.20	0.12	0.17	0.21	**0.22**	0.11	0.19	0.11
	I	0.21	0.12	0.16	**0.25**	0.20	0.13	0.20	0.24	0.22	0.12	0.20	0.21
Part-of	D	0.10	0.05	0.09	**0.16**	0.02	0.02	0.02	0.04	0.02	0.02	0.02	0.03
	I	0.09	0.05	**0.10**	0.08	0.05	0.04	0.05	0.02	0.05	0.04	0.05	0.01
Purpose-of	D	0.11	0.05	0.12	**0.13**	0.00	0.00	0.03	0.07	0.02	0.02	0.04	0.12
	I	0.11	0.13	0.25	**0.35**	0.00	0.02	0.06	0.10	0.02	0.06	0.15	0.18

TABLE VI. MAP@10 for Different Relations, with Word Embeddings from Different Layers of BERT ML and Different Methods

Relation		Layer 1				Layer 2				Layer 11			
		SIM	3CAD	3CAV	LRC	SIM	3CAD	3CAV	LRC	SIM	3CAD	3CAV	LRC
Synonym-of_n		0.01	0.00	0.01	0.01	0.01	0.00	0.01	0.00	0.01	0.00	0.02	0.00
Synonym-of_v		0.05	0.00	0.04	0.04	0.05	0.00	0.03	0.04	0.04	0.00	0.03	0.04
Synonym-of_adj		0.03	0.01	0.04	0.00	0.03	0.01	0.04	0.00	0.03	0.00	0.03	0.00
Antonym-of_adj		0.02	0.00	0.02	0.01	0.03	0.00	0.02	0.01	0.03	0.00	0.04	0.00
Hypernym-of_n	D	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.00	0.01	0.00
(+concrete)	I	0.01	0.00	0.01	0.09	0.01	0.00	0.02	0.09	0.00	0.00	0.02	0.09
Hypernym-of_n	D	0.01	0.00	0.01	0.02	0.01	0.00	0.01	0.02	0.01	0.00	0.01	0.00
(-concrete)	I	0.03	0.01	0.03	0.16	0.04	0.01	0.04	0.15	0.00	0.01	0.01	0.11
Hypernym-of_v	D	0.04	0.00	0.04	0.00	0.04	0.00	0.04	0.00	0.02	0.00	0.02	0.01
	I	0.04	0.00	0.04	0.08	0.03	0.00	0.04	0.06	0.03	0.00	0.03	0.02
Part-of	D	0.01	0.00	0.00	0.01	0.00	0.00	0.00	0.01	0.00	0.00	0.00	0.00
	I	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Purpose-of	D	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	I	0.00	0.00	0.00	0.02	0.00	0.00	0.00	0.03	0.00	0.00	0.00	0.03

TABLE VII. MAP@10 for Different Relations, with Word Embeddings from Different Layers of BERTimbau and Different Methods

Relation		Layer 1				Layer 2				Layer 23			
		SIM	3CAD	3CAV	LRC	SIM	3CAD	3CAV	LRC	SIM	3CAD	3CAV	LRC
Synonym-of_n		0.05	0.02	0.05	0.06	0.06	0.02	0.05	0.04	0.09	0.04	0.08	0.04
Synonym-of_v		0.02	0.00	0.02	0.06	0.03	0.00	0.02	0.07	0.08	0.01	0.10	0.06
Synonym-of_adj		0.08	0.02	0.08	0.00	0.07	0.02	0.07	0.00	0.05	0.01	0.04	0.00
Antonym-of_adj		0.11	0.03	0.11	0.00	0.11	0.03	0.12	0.00	0.12	0.04	0.12	0.01
Hypernym-of_n	D	0.03	0.00	0.03	0.00	0.03	0.00	0.04	0.00	0.04	0.00	0.03	0.00
(+concrete)	I	0.02	0.01	0.02	0.12	0.01	0.01	0.02	0.09	0.08	0.03	0.08	0.07
Hypernym-of_n	D	0.02	0.00	0.01	0.00	0.02	0.00	0.02	0.01	0.12	0.03	0.09	0.00
(-concrete)	I	0.03	0.02	0.04	0.19	0.03	0.03	0.04	0.17	0.15	0.08	0.17	0.16
Hypernym-of_v	D	0.01	0.00	0.01	0.05	0.01	0.00	0.01	0.03	0.13	0.02	0.09	0.01
	I	0.02	0.01	0.02	0.16	0.02	0.01	0.03	0.13	0.16	0.05	0.15	0.05
Part-of	D	0.00	0.00	0.00	0.06	0.01	0.00	0.00	0.04	0.01	0.01	0.01	0.02
	I	0.02	0.01	0.02	0.00	0.02	0.00	0.02	0.01	0.04	0.01	0.04	0.00
Purpose-of	D	0.02	0.00	0.02	0.03	0.02	0.00	0.02	0.02	0.03	0.00	0.05	0.03
	I	0.04	0.01	0.07	0.09	0.04	0.01	0.05	0.06	0.02	0.02	0.08	0.04

For five out of the 10 non-symmetrical relations, the best MAP was achieved by LRCos in GloVe, confirming that this configuration is a good choice for relations of this kind. This happened to the inverse of Hypernym-of, between concrete nouns (0.29) and verbs (0.25), to Part-of (0.16), and Purpose-of, in both direct (0.15) and inverse direction (0.35).

The best MAP for the direct Hypernym-of was again for the Similar-to-B, with GloVe for nouns, and word2vec-SKIP for verbs. To some extent, the performance of this baseline in these relations is explained by the high similarity in hypernym-hyponym pairs, as it happens for synonymy, i.e., hyponyms are very similar to their hypernyms, only more specific than synonyms. Yet, LRCos performs much better on the inverse direction than on the direct, suggesting that it is more difficult to find hyponyms (b^*) given their hypernym, when compared to the other way round. This reflects the higher number of hyponyms, especially when considering indirect hyponyms. Even though, in most cases, there was more than one acceptable answer, a list of hyponyms can be so extensive that some will frequently be missing (see Section VI). Besides the large number, the heterogeneity of the hyponyms also contributes to this low result, making it harder to learn a good representation of the hyponym class with the logistic regression classifier. This further explains, at least partially, why differences between the direct and inverse Hypernymy-of are smaller for 3CosAvg, which does not rely on a classifier, in some cases with better performance for the direct relation. This again suggests that different configurations are better suited for different goals.

Still on Hypernym-of, performance is generally better when it is between concrete concepts than for those more abstract. This should be due to the nature of abstract nouns, with which one cannot interact directly, making it also difficult to generalise the contexts they occur in general text and, for LRCos, to represent their class. Here, the main surprise was the performance of the BERTimbau embeddings, namely the first-layer encoding, where LRCos achieved the best MAP overall for the inverse Hypernym-of between abstract words (0.19), suggesting that the aforementioned contexts are better captured by BERT. Yet, what also contributed to this interesting performance is that this is the relation with the lower proportion of questions with OOV words, only two out of the 50. For the same relation in BERTimbau, we also highlight the MAP of the 3CosAvg in layer 23 (0.17). In fact, for all methods but LRCos the MAP of all relations improves for the upper layer. On the other hand, for LRCos, it decreases for most relations in that layer, which would be more in agreement with previous work [42], [44]. Still in layer 23 of BERTimbau, Similar-to-B achieved a MAP above 10% for five different relations: Antonym-of (0.12) and all direct and inverse hypernymy relations between verbs and abstract nouns. BERT-ML performed much worse, with a single MAP above 10%, namely the inverse Hypernym-of between abstract words, with LRCos, in any layer (0.16, 0.15, 0.11).

For Part-of, performance was the poorest of all relations. In GloVe, the MAP with LRCos for the direct relations (0.16) was twice the same measure for the inverse (0.08). With 3CosAvg, the latter was slightly higher (0.10). And even though it is the second relation for which there are less questions with OOV words in the BERT models, 17 and nine, respectively for BERT-ML and BERTimbau, it is far from achieving the second best MAP. Similarly to Hypernym-of, this might be affected by the fact that an object might have several parts and it may be a part of different objects. Yet, in this case, the low MAP is also affected by other issues (see Section VI).

On the other hand, one of the highest MAPs in the test was achieved for Purpose-of in the inverse direction (0.35). Not only its accuracy was high with LRCos in GloVe, but it was also considerably higher than the baselines, and contrasting with the lower performance in word2vec-CBOW. This stresses that LRCos is well-suited for different kinds of semantic relations. Moreover, although not included in similar tests for English, this suggests that it would be interesting to include the Used-For relation (inverse of Purpose-of) in such benchmarks.

VI. On the Utility of Word Embeddings for Relation Discovery

As discussed in the previous section, when compared to syntactic or word knowledge analogies, solving lexical-semantic analogies from word embeddings is a challenging task. This happens for many reasons. For instance, lexical-semantic relations typically include very frequent words in language, which would result in better representations, if it were not for ambiguity also being higher, i.e., a significant number of these words has more senses than, for instance, names of cities and countries. Another reason is a great number of questions with many possible answers, in opposition, e.g., to syntactic or capital-country relations, for which there is a single answer. In fact, for some cases, there are so many possible answers that they simply cannot all be covered by the dataset, which, in our case, means that they are not in the source lexical resources as well (at least in more than one). Therefore, inspecting the answers by the different method+model configurations and identifying typical issues may, on the one hand, lead to a handful of fixes in future versions of the dataset (i.e., inclusion of more possible answers) and, potentially more interesting, result in important conclusions and suggestions regarding the utility of this kind of approach for enriching lexical resources.

In this section, we first look at the proportion of answers, in a sample, that were automatically considered incorrect, but were still acceptable. Then, we focus on some relation types to find typical issues and confirm that, despite missing from TALES, most acceptable answers constitute good candidates for enriching the lexical resources exploited in its creation. The main conclusion here is that word embeddings should be seen as useful sources of lexical-semantic information that, with analogy solving methods, might be ready for enriching structured lexical resources. Of course, given the still great amount of incorrect answers, the discovered relations should be seen as suggestions, to be considered, or not, for inclusion in the lexical resources, also depending on criteria set by the resource creators.

A. Acceptable Incorrect Answers

For better insights on the achieved performance and typical issues, we inspected the results of the configuration with the highest accuracy, LRCos+GloVe. Towards a more systematic analysis: (i) we focused on the top-10 answers, the same considered when computing MAP@10, for a sample of 15 randomly selected questions of each relation, totalling 210 questions and 2,100 answers; (ii) out of those considered to be incorrect, we manually identified those that were still acceptable, roughly meaning that the relation would make sense. For the target sample, Table VIII shows the proportion of correct answers side-by-side the proportion of answers that we considered acceptable, including one example for each relation type. In this sample, 259 answers (≈12.3%) automatically marked as incorrect were considered acceptable by us, which corresponds to more than twice the number of answers automatically marked as correct in the same sample (111, ≈5.3%).

We recall that incorrect answers are those that did not meet our criteria for inclusion in TALES, i.e., they correspond to relations that are not in any of the exploited lexical resources or they are in a single one, so confidence on them is low. More than suggesting their addition to future versions of TALES, these answers highlight that the exploited lexical resources are limited in terms of coverage, and that this kind of approach can be useful for enriching such resources. An exception

TABLE VIII. Proportion of Acceptable Relation Instances Considered Incorrect in the Answers of LRCos+GloVe

Relation		Correct	Acceptable	Example (PT)	(EN)
Synonym-of_n		6.0%	6.7%	(*luta, combate*)	(fight, combat)
Synonym-of_v		6.7%	8.7%	(*voltar, retornar*)	(come back, return)
Synonym-of_adj		5.3%	6.7%	(*antigo, primitivo*)	(ancient, primitive)
Antonym-of_adj		4.7%	6.0%	(*legítimo, ilegal*)	(legitimate, illegal)
Hypernym-of_n	D	2.7%	10.7%	(*sala, auditório*)	(room, auditorium)
(+concrete)	I	8.7%	16.7%	(*edifício, prédio*)	(edifice, building)
Hypernym-of_n	D	4.0%	18.7%	(*regra, analogia*)	(rule, analogy)
(-concrete)	I	4.7%	7.3%	(*memória, lembrança*)	(memory, reminder)
Hypernymy-of_v	D	5.3%	13.3%	(*receber, acolher*)	(receive, welcome)
	I	10.7%	16.0%	(*mostrar, demonstrar*)	(show, demonstrate)
Part-of	D	3.3%	17.3%	(*porta, armário*)	(door, wardrobe)
	I	4.0%	13.3%	(*humano, cérebro*)	(human, brain)
Purpose-of	D	5.3%	17.3%	(*aquecer, forno*)	(heat, hoven)
	I	2.7%	14.7%	(*camisola, vestir*)	(sweater, wear)

regards a minority of acceptable answers that are the plural (2.7% of the acceptable) or feminine form (1.1% of the acceptable) and are thus not expected to be found in the lexical resources used, because their entries are typically lemmatised. This happens, for instance, in Part-of relations, with *segundos* (seconds) Part-of *minuto* (minute); *minutos* (minutes) Part-of *hora* (hour); or *alunos* (students) Part-of *escola* (school).

Another situation regards transitive relations (e.g., Hypernymy and Part-of), because some lexical resources only make direct connections explicit, not indirect (e.g., inherited hypernyms). This also depends on the taxonomy adopted by the lexical resource and is much noisier in resources extracted from dictionaries. Even though BATS includes relations inherited through transitivity (see e.g., Fig. 1), we did not consider them in the creation of TALES, both due to the aforementioned issue and to the lack of information on word senses, in some resources.

We note that, for each relation, the proportion of acceptable incorrect answers is not correlated with the proportion of correct answers (Pearson coefficient is 0.08). The former is higher for all relations, but this difference ranges from 0.7 points (Synonym-of_n) to 14.7 (direct Hypernym-of_n abstract). On the other hand, the proportion of acceptable incorrect answers is related to the lack of coverage of the instance by TALES, and thus, indirectly, by the lexical resources. By manual inspection, we confirmed that the average number of possible answers, i.e., words related in the target way, is an important contribution to the proportion of acceptable answers not in TALES. For Antonym-of, the relation for which this number is lower, as well as for the other symmetrical relations, the coverage of the lexical resources is not as low as for the other relations. Even in a broad interpretation of antonymy and synonymy, the set of antonyms and synonyms is not as large as for other relations. This is also the case of the inverse Hypernym-of_n for abstract words, but not for the remaining non-symmetrical relations. In fact, the universe of instances of the non-symmetrical relations is considerably larger. As mentioned earlier, a hypernym has several hyponyms, but an object might also have many parts or be used for different purposes. This number increases if inherited relations are considered, namely for hypernymy (e.g., animal Hypernym-of mammal Hypernym-of dog) and part-of (e.g., minute Part-of hour Part-of day Part-of month, ...).

B. Typical Issues

A deeper error analysis was made for the relation with a lower MAP in TALES (the inverse Part-of) and for those with a higher proportion of acceptable answers. Yet, recalling the recurrently given example of hypernymy – a concept might have a huge number of hyponyms – we first focus on the inverse Hypernym-of relation.

As expected, they can be so many that TALES does not cover all possible hyponyms of most Hypernym-of entries. For instance, it includes five types of *escola* (school) but not others given by LRCos+GloVe as an answer, namely *preparatória* (preparatory), *conservatório* (conservatory), *secundária* (secondary) or *liceu* (high school). This happens because none of the aforementioned connections are in any of the lexical resources used. Some, in fact, can be used just as modifiers of *escola*, often appearing together (e.g., *escola preparatória* or *escola secundária*), but they can also be used alone, with the same meaning. Another example is the word *jornal* (newspaper), for which the first answer was *semanário* (weekly newspaper), not accepted because, despite being correct, the instance *jornal* Hypernym-of *semanário* was found in a single lexical resource, and thus not included in TALES. Other issues are related to the presence of world knowledge, much of which not included in dictionaries and LKBs. This happens, for instance, for the word *moeda* (currency), with the first answer 'ecu', the former European currency, precursor of the euro, not in the source lexical resources. The word 'euro' came in second, but is also not in TALES, again because it was in a single lexical resource. A second example of this kind occurred for *automóvel* (car), for which many answers were brands of cars, starting with *fiat*, followed by *volkswagen* (rank #4), *renault* (#5), *bmw* (#6) and *audi* (#7).

Considering the inverse Part-of (Has-Part) relation, for which MAP was very low, we came to the conclusion that the test for this relation includes several difficult entries. Some have multiple senses that can be significantly different, such as *ser* (to be / living being), *câmara* (camera, chamber), or *programa* (program, show). Others refer to abstract concepts, like *todo* (whole), *mundo* (world), *espaço* (space), *organização* (organization), *vida* (life) or *coisa* (thing). On the one hand, the issue of ambiguity is minimised by the presence of several acceptable answers. On the other hand, ambiguous words are used in different contexts, making the relations less obvious in the embedding space. Vagueness could possibly be minimised if, as we did for Hypernym-of, we split concrete and abstract nouns, but available Part-of instances are not enough for this.

Two other issues were noted regarding the confusion of this relation with:

- Hyponymy, i.e., some answers were hyponyms of *b* and not part. For instance, for *homem* (man), answers included *rapaz* (boy), *jovem* (young) and *garoto* (kid); for *casa* (house), *apartamento* (apartment) and *mansão* (mansion); or, for *mês* (month), names of months, like *abril* (April), *maio* (May), *março* (March) and *fevereiro* (February).

- Its inverse, i.e., some answers were not the parts, but the whole of *b*. For instance, for *dia* (day), answers included *semana* (week) and *mês* (month); for *palavra* (word), *expressão* (expression) and *frase* (sentence); or, for *texto* (text), *documento* (document) and *comentário* (comment).

Looking at the relations for which more acceptable answers were found, they include again many cases for which there is a large set of acceptable answers, and not all are in TALES. Examples of such answers include: words for which *sentimento* (feeling) is a hypernym, namely *otimismo* (optimism) and *ansiedade* (anxiety); words for which *porta* (door) is a part, namely *prédio* (building), *casa* (house), *armário* (wardrobe) or *banheiro* (bathroom); or words for which *cozinhar* (to cook) is a purpose-of, namely *forno* (hoven), *molho* (sauce) or *caldo* (broth). These examples also show that, despite acceptable answers, not all are the most obvious and their inclusion would probably require better-defined criteria. We would say that it is virtually impossible to name all possible feelings, all things which have a door, or everything used for cooking, which shows why a lexical resource will never be fully complete regarding some relations. Nevertheless, we believe to have shown that this can be minimised by exploiting word embeddings learned from large corpora.

VII. Concluding Remarks

Towards better insights on how lexical-semantic relations are preserved in pretrained models of word embeddings for Portuguese, we have presented the following contributions:

- TALES, a new analogy-like test covering 14 types of lexical-semantic relations, created automatically with information in Portuguese lexical resources;
- An evaluation covering four different analogy solving methods in TALES, when applied to five pretrained models of Portuguese word embeddings, including static word embeddings as well as embeddings obtained from BERT models;
- An analysis of the obtained results, having in mind the application of the adopted methods for relation discovery in word embeddings and their utility for enriching lexical resources.

TALES is freely available from https://github.com/NLP-CISUC/PT-LexicalSemantics, for anyone willing to use it. As we have shown, it is a challenging test, for which high performances will require better solving methods or different models of word embeddings. Interested researchers may also want to assess other models for Portuguese or alternative ways of exploiting the models used here. According to our experiments, better results are achieved with static word embeddings than with BERT. However, the performance of the latter can most certainly be improved, if this model is used differently. To leverage on Vecto, the platform we used for loading the embeddings and running the tests, we had to get static word representations from BERT, based on its vocabulary file, which makes it impossible to get representations for OOV words and, more importantly, to words obtained from a combination of subwords. While context does not seem to be important in this kind of text, recent work for English has shown that BERT models can still outperform static word embeddings when solving analogies [42]. For better analysing if this is also the case for Portuguese, in the future, we will study alternative ways of handling BERT. In order to keep using Vecto, one possibility would be to include not only a representation for each entry in BERT's vocabulary, but also for all the words in TALES. However, if we just do this, results would probably be positively biased, due to less confusion. Another possibility is to include the encodings of words in a large representative list, starting, for instance, with the vocabulary of the static word embeddings. We should also look at previous work on using BERT for solving lexical tasks (e.g., [44]).

Future experiments may also include alternative analogy solving methods. While we did not get improvements with 3CosMul and PairDistance [46], more recent methods, like the Translation and the Regression Model [17], are not included in Vecto, and were thus not tested.

Based on the analysis of incorrect answers, namely on the proportion of acceptable answers, we are looking forward to using this kind of approach for suggesting new relation instances to Portuguese lexical resources and thus contributing to their semi-automatic enrichment. If focused on a single lexical resource, it is perhaps advisable to use a new test obtained exclusively from its relations, to better capture the criteria followed in its creation. For some LKBs, we could possibly leverage on the word sense organisation and, if desired, include inherited relations in the test. After this, it should be a matter of going through all the incorrect answers and consider their addition to the LKB or not. As we have shown, even though many may be definitely incorrect, some might be acceptable instances that are simply missing from the resource. This will, or course, contribute to better structured lexical resources, with higher coverage.

Acknowledgment

This is an extended version of a paper describing the creation of the TALES dataset and its usage for assessing three models of static word embeddings [47], presented at the ECAI 2020 Workshop on Hybrid Intelligence for Natural Language Processing Tasks. Besides some more detailed discussions, the main additions are the inclusion of the BERT models in our experiment and a more thorough inspection of the results, in order to complement our discussion on the utility of this kind of approach for enriching computational lexical resources.

We would like to thank the reviewers for their useful comments that lead to a better version of this manuscript.

References

[1] C. Fellbaum Ed., *WordNet: An Electronic Lexical Database (Language, Speech, and Communication)*. The MIT Press, 1998.

[2] Z. Harris, "Distributional structure," *Word*, vol. 10, no. 2-3, pp. 1456–1162, 1954.

[3] T. Mikolov, K. Chen, G. Corrado, J. Dean, "Efficient estimation of word representations in vector space," in *Proceedings of the Workshop track of ICLR*, 2013.

[4] J. Pennington, R. Socher, C. D. Manning, "GloVe: Global vectors for word representation," in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*, EMNLP 2014, 2014, pp. 1532–1543, ACL.

[5] A. Gladkova, A. Drozd, S. Matsuoka, "Analogy-based detection of morphological and semantic relations with word embeddings: what works and what doesn't.," in *Proceedings of the NAACL 2016 Student Research Workshop*, 2016, pp. 8–15, ACL.

[6] E. Vylomova, L. Rimell, T. Cohn, T. Baldwin, "Take and took, gaggle and goose, book and read: Evaluating the utility of vector differences for lexical relation learning," in *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics* (*Volume 1: Long Papers*), ACL 2016, 2016, pp. 1671–1682, ACL.

[7] A. Querido, R. Carvalho, J. Rodrigues, M. Garcia, J. Silva, C. Correia, N. Rendeiro, R. Pereira, M. Campos, A. Branco, "LX-LR4DistSemEval: a collection of language resources for the evaluation of distributional semantic models of Portuguese," *Revista da Associação Portuguesa de Linguística*, vol. 3, no. 3, pp. 265–283, 2017.

[8] G. Hirst, "Ontology and the lexicon," in *Handbook on Ontologies*, S.

Staab, R. Studer Eds., International Handbooks on Information Systems, Springer, 2004, pp. 209–230.

[9] V. de Paiva, L. Real, H. Gonçalo Oliveira, A. Rademaker, C. Freitas, A. Simões, "An overview of Portuguese wordnets," in *Proceedings of 8th Global WordNet Conference*, GWC'16, Bucharest, Romania, 2016, pp. 74–81.

[10] A. Drozd, A. Gladkova, S. Matsuoka, "Word embeddings, analogies, and machine learning: Beyond king - man + woman = queen," in *Proceedings the 26th International Conference on Computational Linguistics: Technical papers COLING 2016*, COLING 2016, 2016, pp. 3519–3530.

[11] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, NAACL-HLT 2019, 2019, pp. 4171–4186, ACL.

[12] L. Finkelstein, E. Gabrilovich, Y. Matias, E. Rivlin, Z. Solan, G. Wolfman, E. Ruppin, "Placing search in context: The concept revisited," *ACM Trans. Inf. Syst.*, vol. 20, pp. 116–131, Jan. 2002.

[13] F. Hill, R. Reichart, A. Korhonen, "Simlex-999: Evaluating semantic models with genuine similarity estimation," *Computational Linguistics*, vol. 41, no. 4, pp. 665–695, 2015.

[14] E. Agirre, E. Alfonseca, K. Hall, J. Kravalova, M. Paşca, A. Soroa, "A study on similarity and relatedness using distributional and WordNet-based approaches," in *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, NAACL '09, 2009, pp. 19–27, ACL.

[15] D. Jurgens, S. Mohammad, P. Turney, K. Holyoak, "SemEval-2012 task 2: Measuring degrees of relational similarity," in **SEM 2012: The 1st Joint Conference on Lexical and Computational Semantics – Vol 1: Proc. of main conference and shared task, Vol 2: Proc. of 6th (SemEval 2012)*, 2012, pp. 356–364, ACL.

[16] M. Baroni, A. Lenci, "How we BLESSed distributional semantic evaluation," in *Proceedings of GEMS 2011 Workshop on GEometrical Models of Natural Language Semantics*, Edinburgh, UK, 2011, pp. 1–10, ACL.

[17] Z. Bouraoui, S. Jameel, S. Schockaert, "Relation induction in word embeddings revisited," in *Proceedings of the 27th International Conference on Computational Linguistics*, COLING 2018, Santa Fe, New Mexico, USA, Aug. 2018, pp. 1627–1637, ACL.

[18] H. Gonçalo Oliveira, "Distributional and Knowledge-Based Approaches for Computing Portuguese Word Similarity," *Information*, vol. 9, no. 2, 2018.

[19] N. S. Hartmann, E. R. Fonseca, C. D. Shulby, M. V. Treviso, J. S. Rodrigues, S. M. Aluísio, "Portuguese word embeddings: Evaluating on word analogies and natural language tasks," in Proceedings 11th Brazilian Symposium in Information and Human Language Technology *(STIL 2017)*, 2017.

[20] T. Sousa, H. G. Oliveira, A. Alves, "Exploring different methods for solving analogies with portuguese word embeddings," in *Proceedings 9th Symposium on Languages, Applications and Technologies, SLATE 2020, July 13-14, 2020, School of Technology, Polytechnic Institute of Cávado and Ave, Portugal*, vol. 83 of OASIcs, 2020, pp. 9:1–9:14, Schloss Dagstuhl - Leibniz-Zentrum für Informatik.

[21] R. Wilkens, L. Zilio, E. Ferreira, A. Villavicencio, "B2SG: a TOEFL-like task for Portuguese," in *Proceedings of the Tenth International Conference on Language Resources and Evaluation*, LREC 2016, Paris, France, 2016, ELRA.

[22] R. Navigli, S. P. Ponzetto, "BabelNet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network," *Artificial Intelligence*, vol. 193, pp. 217–250, 2012.

[23] N. Calzolari, L. Pecchia, A. Zampolli, "Working on the italian machine dictionary: a semantic approach," in *COLING 1973 Volume 2: Computational And Mathematical Linguistics: Proceedings of the International Conference on Computational Linguistics*, 1973, Association for Computational Linguistics.

[24] M. A. Hearst, "Automatic acquisition of hyponyms from large text corpora," in *Proceedings of 14th Conference on Computational Linguistics*, COLING 92, Morristown, NJ, USA, 1992, pp. 539–545, Association for Computational Linguistics.

[25] R. Snow, D. Jurafsky, A. Ng, "Learning syntactic patterns for automatic hypernym discovery," *Advances in neural information processing systems*, vol. 17, pp. 1297–1304, 2005.

[26] P. Pantel, M. Pennacchiotti, "Espresso: Leveraging generic patterns for automatically harvesting semantic relations," in *Procs of 21st International Conference on Computational Linguistics and 44th annual meeting of the Association for Computational Linguistics*, Sydney, Australia, 2006, pp. 113–120, ACL Press.

[27] H. Gonçalo Oliveira, P. Gomes, "ECO and Onto.PT: A flexible approach for creating a Portuguese wordnet automatically," *Language Resources and Evaluation*, vol. 48, no. 2, pp. 373–393, 2014.

[28] H. Gonçalo Oliveira, D. Santos, P. Gomes, N. Seco, "PAPEL: A dictionary-based lexical ontology for Portuguese," in *Proceedings of Computational Processing of the Portuguese Language - 8th International Conference (PROPOR 2008)*, vol. 5190 of LNCS/LNAI, Aveiro, Portugal, September 2008, pp. 31–40, Springer.

[29] H. Gonçalo Oliveira, "A survey on Portuguese lexical knowledge bases: Contents, comparison and combination," *Information*, vol. 9, no. 2, 2018.

[30] A. Simões, Á. I. Sanromán, J. J. Almeida, "Dicionário-Aberto: A source of resources for the Portuguese language processing," in *Proceedings of Computational Processing of the Portuguese Language, 10th International Conference (PROPOR 2012), Coimbra Portugal*, vol. 7243 of LNCS, April 2012, pp. 121–127, Springer.

[31] L. Anton Pérez, H. Gonçalo Oliveira, P. Gomes, "Extracting lexical-semantic knowledge from the Portuguese Wiktionary," in *Proceedings of the 15th Portuguese Conference on Artificial Intelligence*, EPIA 2011, Lisbon, Portugal, October 2011, pp. 703–717, APPIA.

[32] E. G. Maziero, T. A. S. Pardo, A. D. Felippo, B. C. Dias-da-Silva, "A Base de Dados Lexical e a Interface Web do TeP 2.0 - Thesaurus Eletrônico para o Português do Brasil," in *VI Workshop em Tecnologia da Informação e Linguagem Humana*, TIL, 2008, pp. 390–392.

[33] V. de Paiva, A. Rademaker, G. de Melo, "OpenWordNet-PT: An Open Brazilian WordNet for Reasoning," in *Proceedings of 24th International Conference on Computational Linguistics*, COLING (Demo Paper), 2012.

[34] A. Simões, X. G. Guinovart, "Bootstrapping a Portuguese wordnet from Galician, Spanish and English wordnets," in *Advances in Speech and Language Technologies for Iberian Languages, Proceedings of 2nd International Conference, IberSPEECH 2014, Las Palmas de Gran Canaria, Spain*, vol. 8854 of LNCS, 2014, pp. 239–248, Springer.

[35] B. C. Dias-da-Silva, "Wordnet.Br: An exercise of human language technology research," in *Proceedings of 3rd International WordNet Conference (GWC)*, GWC 2006, South Jeju Island, Korea, January 2006, pp. 301–303.

[36] A. Barreiro, "Port4NooJ: an open source, ontology-driven portuguese linguistic system with applications in machine translation," in *Proceedings of the 2008 International NooJ Conference (NooJ'08)*, Budapest, Hungary, 2010, Newcastle-upon-Tyne: Cambridge Scholars Publishing.

[37] R. Speer, J. Chin, C. Havasi, "Conceptnet 5.5: An open multilingual graph of general knowledge," in *Proceedings of Thirty-First Conference on Artificial Intelligence (AAAI)*, 2017, pp. 4444–4451.

[38] P. A. Rocha, D. Santos, "CETEMPúblico: Um corpus de grandes dimensões de linguagem jornalística portuguesa," in *V Encontro para o processamento computacional da língua portuguesa escrita e falada (PROPOR 2000),* São Paulo, 19-22 de Novembro 2000, pp. 131–140, ICMC/USP.

[39] D. Santos, E. Bick, "Providing Internet access to Portuguese corpora: the AC/DC project," in *Proceedings of 2nd International Conference on Language Resources and Evaluation*, LREC 2000, 2000, pp. 205–210.

[40] A. Paivio, J. C. Yuille, S. A. Madigan, "Concreteness, imagery, and meaningfulness values for 925 nouns," *Journal of Experimental Psychology monograph supplement*, vol. 76, no. 1, pp. 1–25, 1968.

[41] A. P. Soares, A. S. Costa, J. Machado, M. Comesaña, H. M. Oliveira, "The Minho Word Pool: Norms for imageability, concreteness, and subjective frequency for 3,800 Portuguese words," *Behavior Research Methods*, vol. 49, no. 3, pp. 1065–1081, 2017.

[42] K. Ethayarajh, "How contextual are contextualized word representations? comparing the geometry of bert, elmo, and gpt-2 embeddings," in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 2019, pp. 55–65.

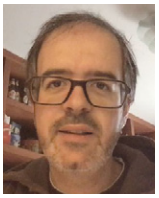
[43] F. Souza, R. Nogueira, R. Lotufo, "Bertimbau: Pretrained bert models for brazilian portuguese," in *Proceedings of the Brazilian Conference on Intelligent Systems (BRACIS 2020)*, vol. 12319 of LNCS, Cham, 2020, pp. 403–417, Springer.

[44] I. Vulić, E. M. Ponti, R. Litschko, G. Glavaš, A. Korhonen, "Probing pretrained language models for lexical semantics," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2020, pp. 7222–7240.

[45] T. Linzen, "Issues in evaluating semantic spaces using word analogies," in *Proceedings of the 1st Workshop on Evaluating Vector-Space Representations for NLP*, Berlin, Germany, Aug. 2016, pp. 13–18, ACL.

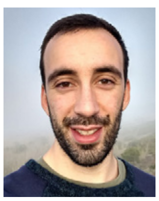
[46] O. Levy, Y. Goldberg, "Linguistic regularities in sparse and explicit word representations," in *Proceedings of 18th Conference on Computational Natural Language Learning*, CoNLL 2014, 2014, pp. 171–180, ACL.

[47] H. Gonçalo Oliveira, T. Sousa, A. Alves, "TALES: Test set of Portuguese lexical-semantic relations for assessing word embeddings," in *Proceedings of the ECAI 2020 Workshop on Hybrid Intelligence for Natural Language Processing Tasks (HI4NLP 2020)*, vol. 2693 of CEUR Workshop Proceedings, 2020, pp. 41–47, CEUR-WS.org.



Hugo Gonçalo Oliveira

Hugo Gonçalo Oliveira is an Assistant Professor at the Department of Informatics Engineering of the University of Coimbra and a researcher of the Center of Informatics and Systems of the same university (CISUC), in the Cognitive and Media Systems lab. His main research interests lie in Artificial Intelligence and sub-domains of Computational Creativity (CC) and Natural Language Processing (NLP), with a focus on the Portuguese language. He has developed some prototypes for generating creative text in Portuguese, such as Tra-la-Lyrics and PoeTryMe, and participated in the FET European projects ConCreTe and PROSECCO. His research also led to the development of several lexical resources for Portuguese, such as PAPEL and Onto.PT, the main scope of his PhD thesis (2013), which was awarded as the best in the computational processing of Portuguese (2011-2014). His current research also covers topics like Semantic Textual Similarity and Conversational Agents, following his leading role in two national-funded projects on the latter (AIA, FLOWANCE). He is the author of more than 100 peer-reviewed scientific papers, has participated in national-funded projects (InfoCrowds, REMINDS, Socialite), in the organisation of scientific conferences and workshops (CC-NLG 2017-19, SLATE 2019, PROPOR 2018) and shared tasks (Second HAREM, TweetMT, ASSIN-2), and is a regular member of the program committee of some of the main scientific events on CC (ICCC) and NLP (ACL, EMNLP, COLING).



Tiago Sousa

Tiago Sousa has a degree in Informatics Engineering from Institute of Engineering of Coimbra and is a student of the Master in Informatics and Systems at the Institute of Engineering of Coimbra, where his main research focus is around Portuguese word embeddings and their applications to analogy solving and relation discovery. Since 2016, he also works at Present Technologies as a software developer.

Ana Alves

Ana Alves is, since 2007, a member of the Ambient Intelligence Laboratory (AmILab) of the Cognitive and Media Systems (CMS) group, integrated in the Center for Informatics and Systems of the University of Coimbra (CISUC). Her research is dedicated mainly to the way urban spaces are organized and how people use them by influencing patterns of mobility and land use analysis. The information to support this analysis is mined primarily from the Web and social networks and consists mostly of textual data. Learning to interconnect and represent this information is another research challenge she is devoted to. She is also an Assistant Professor at Polytechnic Institute of Coimbra (IPC), Coimbra Institute of Engineering (ISEC), since 2000, lecturing courses on programming, operating systems and ubiquitous computing. She holds a Ph.D. in Science and Information Technology since 2012, MSc. in Informatics and Systems in 2004 (pre-bolonha), and a 5-year Bachelor degree in Informatics Engineering in 2000 from the University of Coimbra. She is a member of scientific associations such as: Portuguese Artificial Intelligence Association (APPIA - Associação Portuguesa Para a Inteligência Artificial); Association for Computational Linguistics (ACL); and Registered researcher at the Foundation for Science and Technology (Fundação para a Ciência e Tecnologia).

The Semantics of History. Interdisciplinary Categories and Methods for Digital Historical Research

Esther Travé Allepuz[1], Pablo del Fresno Bernal[2], Alfred Mauri Martí[3], Sonia Medina Gordo[1] *

[1] Universitat de Barcelona, Barcelona (Spain)
[2] Sistemes de Gestió de Patrimoni SCCL, Barcelona (Spain)
[3] Centre d'Estudis Martorellencs, Martorell (Spain)

Received 4 November 2020 | Accepted 21 December 2020 | Published 10 February 2021

unir
LA UNIVERSIDAD EN INTERNET

Abstract

This paper aims at introducing and discussing the data modelling and labelling methods for interdisciplinary and digital research in History developed and used by the authors. Our approach suggests the development of a conceptual framework for interdisciplinary research in history as a much-needed strategy to ensure that historians use all vestiges from the past regardless of their origin or support for the construction of historical discourse. By labelling Units of Topography and Actors in a wide range of historical sources and exploiting the obtained data, we use the Monastery of Sant Genís de Rocafort (Martorell, Spain) as a lab example of our method. This should lead researchers to the development of an integrated historical discourse maximizing the potential of interdisciplinary and fair research and minimizing the risks of bias.

Keywords

Unit of Topography, Actor, Integrated History, Ontology, Database, Records Management, Landscape Archaeology.

DOI: 10.9781/ijimai.2021.02.002

I. Introduction

Historical Science is a discipline devoted to the analysis and comprehension of the past for a better understanding of the present, and hence a potential forecasting of the future. Therefore, space and time are unavoidably the main scenarios of research in History, and the alternation between permanence and change is its main object of study. This analysis of the so-called historical time [1], in which different entities exist and events occur transforming the reality of the past into something new and different, leads historians to construct the past in a narrative form.

The fact that data related to time and change are present in any written, material or immaterial vestige of the past boosts –and even forces– interdisciplinary research in History. Despite this, traditional approaches have frequently focused on written vestiges preferably, while disregarding other sources of information such as archaeology, iconography, literature, and a wide range of social sciences and humanities, which have been often considered as complementary at their best. In doing so, the different vestiges of the past do not *integrate* within a single discourse, but originate parallel discourses that might incur into contradiction and potentially lead to endless academic debates.

The spatial and material turns in History [2]–[6] have challenged these old-fashioned approaches, and have led to a more accurate construction of the past. A new interdisciplinary research framework –in which the difference between SSH and STEM blur into a new paradigm of digital and FAIR science– challenges History, and requires an effort from different disciplines in order to explore common languages and codes for the construction of the past. In this process of science going digital, common and exchangeable units of information are required, despite the specificity of different areas of expertise. Within this domain, some experiences on NLP have been developed [7]–[8] and most of them are related to textual sources [9]–[10]. Despite this, the strengthening of concept-based and relation-based corpora for the development of NLP in digital humanities [11] focusing on the ontological approach to historical data suggests a challenging scenario for historians. Scholars dealing with historical science should face the need of rethinking methodologies and the way to use ICT in order to solve wider and more complex research questions and to take our investigation beyond the 20th Century historical issues. One of our objectives is to provide ontological reflections about data and data management in order to produce richer historical relates, as long as they integrate as many vestiges of the past as possible.

Our contribution aims at offering a methodological proposal and practical application our teamwork has developed within the last years, as an extended version of [12]. Arising from landscape archaeology and the study of the material vestiges of the medieval period, our research methodology deals with data labelling and records management, and nowadays it has overcome the archaeological domain to integrate all vestiges of the past regardless of their nature or origin to strengthen historical research in the digital domain. We will develop these issues within an updated state of the art in section II. In the following sections, we will describe the method according to the labelling categories we propose and the resulting data modelling. We will use a practical case of study –the medieval monastery of Sant Genís de Rocafort, Martorell, Spain– as an illustrative example.

* Corresponding author.

E-mail addresses: esther.trave@ub.edu (Esther Travé Allepuz), pdfsgp@gmail.com (Pablo del Fresno Bernal), bnn@heraclit.net (Alfred Mauri Martí), smedingo12@alumnes.ub.edu (Sonia Medina Gordo).

II. Landscape Archaeology and Integrated History

History has not been the most enthusiastic discipline to join the so-called Digital Humanities so far, and data managing strategies have been widely challenged in our domain [13]. Far from being overwhelmed by the unknowns of this digital turn, a few exceptions deal with different ways of representing historical information [14]–[15] and the building of a semantic definition for historical ontology [16]–[18]. Recent experiences focus on quantitative data analyses [19] and, predominantly, on written historical texts [20]–[21]; and some of them struggle to find the best ways to deal with bias [22] and uncertainty [23]. Despite this, a normalized user-friendly code to exploit vestiges of different nature and support is still missing and historical knowledge seems to be restricted to its written apparel.

Our team developed a first proposal to identify minimum units of information and label them as Units of Topography, Units of Stratigraphy and Actors, as will be defined below, in the framework of Landscape Archaeology. The initial goal was to integrate vestiges both from written or material sources in the archaeological analysis of medieval landscape and, therefore, to explore landscape as a historical construct from a holistic perspective. More than twenty years later, our research information system has gone far beyond landscape archaeology or the medieval period to become a solid proposal for historical research, understanding Historical Science as a FAIR [24] construction of the past. It follows the principles stated by far-reaching research projects in this field such as the International Research on Permanent Authentic Records in Electronic Systems (InterPARES) [25] (pp. 6–7) or Parthenos [26]. This FAIR-ness is achieved when ensuring the reliability and traceability of the research process, and when integrating as many sources of information as possible, even those that had been traditionally disregarded until recently as marginal or non-significant.

This is possible due to the development of ICT in the past decades. Recent advances allow for a significant speedup of data gathering and exploitation processes of much larger datasets, which opens a brand-new field for historical research in which new and more complex questions can be asked to past vestiges. Ensuring the existence of clear and unambiguous definitions of the ontology-mediated elements that identify units of information and their relations [27] is an underpinning issue to this regard. The following section deals with it.

III. Labelling Categories and Data Modelling

As defined by K. Thibodeau [28] (p. 7), an Entity is something that existed and an Event is something that happened or was done. Entities and Events have a relationship of involvement, as every event involves at least one entity that might be the participant in the event, its observer, the mechanism for the event to happen, or the object altered by the event itself. In terms of data-labelling, the categories Unit of Topography and Actor, as defined by A. Mauri [29] (p. 45), and their relations, provide the unique and univocal identifiers for historical facts regardless of their link to permanence (Entity) or change (Event), or the nature and support of the vestige. Units of Topography as we use them are, in fact, a wider conception of archaeological Units of Stratigraphy [30], which overcomes their materiality and turns them into a broader concept to identify any entity or action existing or occurring at a particular time, notwithstanding its presence or absence in the archaeological record.

The following definitions apply to each one of these categories:

- Unit of Topography (UT): It is the evidence of an action or situation that can be located in space and time, regardless of the specificity of the information source and its biotic, non-biotic or anthropic attributes. Each UT has a specific location and date. Location can be expressed as a UTM coordinate or as an administrative delimitation that might have changed through time.
- Unit of Stratigraphy (US): It is the material evidence of an action occurred in the past, representing an archaeological aspect of the cycle of time. They are of universal character and can be found on any archaeological site in the world [30] (p. 42). As a reflection of materiality, graphic and cartographic representations are essential attributes of these units.
- Actor: It is the individual or corporative, active or passive, protagonist of an action identified as a UT. If being an individual, its attributes are their name, gender, religion, citizenship, date of birth and death, etc. Different individual actors gathered for a given period of time with a particular purpose and under determinate conditions can act as corporative actors.

As the US category is contained in the definition of a UT and we might consider them as equals at some point, Table I summarizes their differences and ontological specificities.

As shown in Fig. 1, several types of relationships can be set between UT/US and Ac. A UT can include, link or delimitate another UT. Hence, Inclusion, Delimitation and Link are classes of the UT-UT relation. An Actor always plays an active or passive role within a UT, so Role is the only class of Ac-UT. Actors can relate to other actors through familial, political, social or economic Ac-Ac relationships. The materiality of US implies that the only possible relation between US is physical contact. When interpreting the archaeological register, we can group several US into activities and assemble these activities into groups of activities

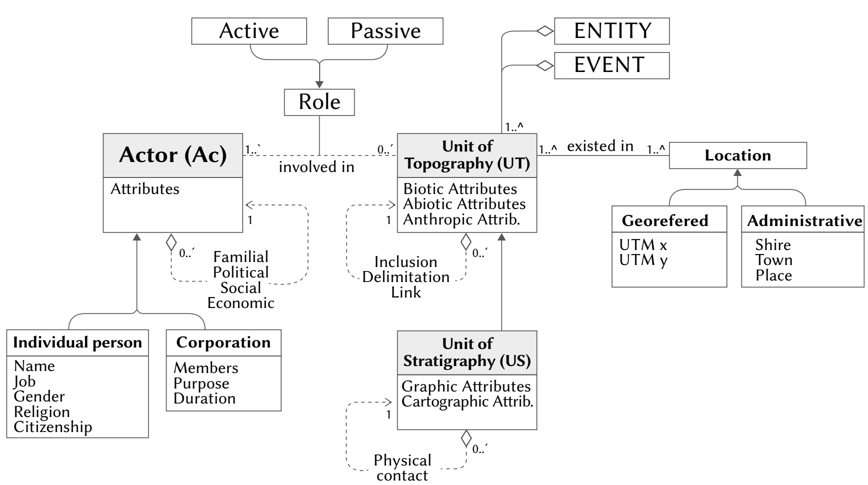


Fig. 1. UML diagram of ontological concepts –UT, US, Ac and their relations– for integrated historical research.

according to [31]. The relation and equivalence between UT, as shown in Table I, and US is then a matter of scale: UT can be equivalent to US in the archaeological record, or we can define UT during the data interpretation process. Anyhow, the UT scale is the one allowing for an interdisciplinary approach in History.

TABLE I. Units of Information (UT/US) Ontological Summary

Attribute		US	UT
Source of obtention		Material	Written, material, visual, sound…
Materiality		Essential	Non-essential
Informative dimensions	Spatial	Essential	Essential
	Descriptive	Essential	Essential
	Graphic	Essential	Non-essential
	Cartographic	Essential	Non-essential
	Temporal	Essential	Essential
Possible relations	US-US/UT-UT	Physical contact	Inclusion, delimitation, link
	Ac-US/Ac-UT	Non-possible*	Non-essential
	US-UT	US = UT US + US = UT	US = UT UT + UT = UT
Involvement in event		Essential (altered)	Non-essential

* The material register does not inform about Ac.

IV. A Case Study: Results at Sant Genís De Rocafort

A. Site and Materials

We have selected a Catalonian monument –the monastery of Sant Genís de Rocafort– located at the town of Martorell (Barcelona) in order to put into practice our labelling proposal and demonstrate the validity of our information system from an interdisciplinary perspective. The Lord of Castellvell and his wife founded the monastery of Sant Genís in 1042 as the ruling centre of a small Priory [32].

Since then, the monastery, located on the top of a hill, became an emblematic element of Martorell's landscape, and down on the foothill the ancient temple of Santa Margarida became the parish church of the Priory, which had a flourishing period in 11th – 13th Cent AD, even though the monastic community never included more than four or five monks. The monastery became dependent of larger monasteries within a Benedictine network, probably in the 13th Cent. Since then, the monastery of Sant Miquel de Cruïlles (Girona, Spain) [33] had the patronage over Sant Genís, while being itself dependent from the Piedmontese Saint Michael's Abbey (San Michele della Chiusa, Italy). In the late middle age, Sant Genís started its slow decadence until 1534, when it became a secular Priory owned by a Barcelonese merchant named Joan Bolet. He slightly refurbished the buildings of Sant Genís and Santa Margarida, and the Priory endured until the 19th Cent AD. After some years in private hands, the Priory became a public property of Martorell's Town Council in 1967.

Historical research at this site has benefited from the proposed system of management information. Architectural analysis of the building, restoration works, and archaeological fieldwork provide a new research perspective for the construction of an integrated historical discourse built from the written evidences kept in archival records and thoroughly analysed in the past [34].

B. Practical Examples of Data Labelling

Accordingly, we selected some written, graphic or material vestiges of Sant Genís' past and analysed them by identifying and labelling the US, UT and Ac informed within at different stages. Managing information throughout these lab examples in an integrated form provides a synthesis for the method's development. The following subsections deal with each type of selected sources, labelling them appropriately, and Table II and Table III summarize the information gathered so that it can be furtherly discussed.

TABLE II. Simplified Ac Dataset Gathered from Selected Examples.

Ac	Name	Attributes	Related UT/Ac
01	Bonus	Lord of Castellvell	Ac02, 03; UT01
02	Guilielmus		Ac01
03	Sicardis		Ac01; UT01
04	Clement VII	Pope	UT20, 21
05	Joan Bolet	Barcelonese merchant	Ac06; UT07, 22, 23
06	Simó Capellades	Priest, Prior	Ac05; UT23, 25

TABLE III. Simplified UT Dataset Gathered from Selected Examples

ut	Brief Description	Related UT/US/Ac	Attributes	Date
01	Donation	Ac01, 03	Event (property)	1042
02	Romanesque Church	UT07, 25, 26, 27, 28, 29, 30, 31*; **US5001 – 5024**	Building	12th Cent
03	County of Barcelona	UT07	Political entity	
04	Castellvell	Ac01; UT05, 06, 07	Town	
05	Martorell	UT04, 06, 07	Town	
06	Priory border	UT07, 08 – 18	Border	
07	Priory of Sant Genís	UT04, 05, 06, 22, 25	Religious entity	
08	Congostell	UT07	Place	
09	Mountain range	UT07, 10, 11	Mountain range	
10	Rosanes	UT07, 11	Place	
11	Rocafort	UT02, 07, 10, 25	Hill	
12	Montgoi	UT07	Hill	
13	Vena	UT07	Place	
14	Grau	UT07	Place	
15	Torrent	UT07	Place	
16	Lloreda	UT07	Place	
17	Torrent of Lloreda	UT07	Waterflow	
18	Anoia River	UT07	Waterflow	
19	Donation letter	Ac01, 03; UT01	Document	1042
20	Papal bull	Ac04, UT21	Document	1534
21	Concession	Ac04, UT20, 22	Event (gift)	1534
22	Patronage	Ac05, UT07, 22	Entity (property)	1534
23	Refurbishment	Ac05, UT02, 25, 31*; **US247, 248, 249, 5025, 5027, 5031, 5033, 5068 – 5070**	Building transformation	16th Cent
24	St Miquel de Cruïlles	UT25	Religious entity	
25	Monastery of St Genís	UT07, 24, 25, 31*; **US5026, 5030, 5032**	Religious entity	
26	Partial collapse of Romanesque building	UT02; **=US5034**	Destruction	1448
27	Apse walls	UT02, 26, 28		
28	Apse demolition	UT02, 27; **=US5086**	Destruction	1928
29	Abandonment	UT02, 07, 25; **US5019, 5023, 5077, 5080, 5083**	Enduring event	19th Cent
30	Restoration	UT02; **US5035 – 5039, 5079, 5082, 5085**	Building transformation	2014
31	Archaeological fieldwork	UT02, 25, 26, 28, 29; **US247, 248, 249**	Event (Research)	Since 2010

* Units of Stratigraphy (US) with the labelling format **000** are related to the Romanesque church or the monastery buildings and thus included in UT30.

1. Archival Sources or Written Primary Files

A copy of the donation document of 1042 kept at the Diocesan Archive of Barcelona [32] and published in [34] (p. 139-143) informs about the location and borders of the Priory. We can label texts and identify the units of information contained and their attributes by following this code: <**UT00**> <**Ac00**> <Att-UT00> <Att-Ac00> <Date-UT00>. We also label relations in accordance with concepts related: <Ac00-UT00> <Ac00-Ac00> <UT00-UT00>. A semicolon separates different UT, Ac, Attributes, or Relations identified through the same word or syntagmatic expression.

Sit omnibus notum quod ego <**Ac01** *Bonus*>, <Ac01-Ac02 *filius*> <**Ac02** *Guilielmi*>, <Att-Ac01 *dominus Castri Vetuli*>, *et* <Ac01-Ac03 *uxor mea*> <**Ac03** *Sicardis*> *pariter in unum in nomine Domini* <**UT01** *donatores*> *Domino Deo et sancto Genesio martiri Christi, cuius* <**UT02** *ecclesia*> *sita est in* <**UT03** *comitatu barchinonensi*>, <UT02-UT04; UT02-UT05 *intra terminos*> <**UT04** *Castri Vetuli*> <UT04-UT05 *de*> <**UT05** *Martorello*>. *[...]* <UT06-UT07 *Habent autem*> <**UT06***terminum*> <**UT07** *hec omnia*> *ab ortu solis in ipso* <**UT08** *Coangustello*>, *atque ascendendo in sumitate* <**UT09** *serre*> *pergit per sumitatem illius* <**UT09** *serre*>, *que est* <Att-UT09 *intra* <**UT10** *Rodanes*> <UT10-UT11 *et*> <**UT22** *Rocam fortem*>>, *usque in* <**UT12** *muntem Gaudii*>. *Et inde pergit usque ad* <**UT13** *Bennam*>, *secundum quod hiemali tempore sive pluviali decurrunt aque ab oriente contra septentrionem. A meridie in ipsa Bennam, atque pergit inde usque ad ipsum* <**UT14** *Gradum*>, *et descendit in ipsum* <**UT15** *torrentem*> *qui discurrit. Ab occidio in ipsa* <**UT16** *Laureta*>, *et inde pergit per ipsam* <**UT17** *rieram de Laureta*> *usque in medium* <**UT18** *flumen Annole*>. *A circio similiter in medium* <**UT18** *flumen Anolle*> *et inde pergit per medium ipsius fluminis usque in ipsum* <**UT08** *Coangustellum*>. *[...]* <Date-UT01; Date-UT19 *Facta* <**UT19** *carta donationis*> *huius sex idus aprilis anno undecino regni Henrici Regis*>. [26], [29] (p. 140)

Fig. 2. Aerial view of Sant Genís de Rocafort after fieldwork carried out in 2020 (a), and photographic vestiges from the beginning of the 20th Century (b-c) [35]–[36]. Interpreted relations between US and UT are recorded in Table III.

2. Photographic Vestiges and Architectural Analysis

Architectural analysis from an archaeological perspective aims at identifying the building phases and further transformations in architectural heritage. Photographic vestiges usually allow for the identification of constructive elements that have disappeared nowadays. We identified and labelled US in Fig. 2 accordingly:

3. Archaeological Fieldwork

Because archaeology is a destructive process, building a precise and detailed archaeological record following a clearly stated protocol has been the commitment of archaeological science for a very long time. US have been the main unit of information since 1980 [30].

In Fig. 3, we selected just three US for labelling as a lab example amongst the entire archaeological record at this site. The stratigraphic method for archaeological excavation and register is widely accepted among scholars in this domain and our UT/Ac labelling strategy is built in accordance to the archaeological method, as stated above, due to the authors' archaeological background. Therefore, we could not avoid selecting archaeological examples for building an integrated historical narrative, as the archaeological record is concomitant with our proposal for information management, even though the excavation results in Sant Genís are much wider than shown in this paper.

4. Bibliographic Reflections

In all scientific production and in any form of Past Construction in particular, the so-called state-of-the-art –or past reflections, in Thibodeau's terms [28]– are valuable sources of information that must be considered in terms of data labelling and management. The same labelling method proposed for archival sources or written primary files works for secondary information as well.

In Fig. 4 we show an example of data labelling within an excerpt from a published piece of research about the Priory of Sant Genís [34]. That is the summary and study of a rich documentary assemblage informing about the priory from archival sources, and providing the historical framework and state-of-the-art before the archaeological excavation started.

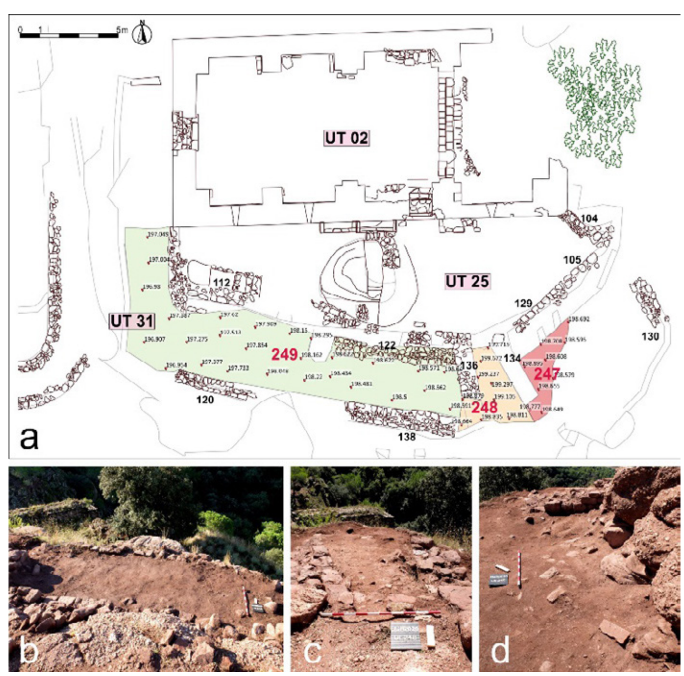


Fig. 3. Cartographic (a) and photographic representations of US 249, 248 and 247 (b-d). As per architectural analysis, US-UT relations are recorded in Table III.

Fig. 4. Labelling example of a short excerpt from the bibliographic reference of a previously published study [34].

UT/Ac data labelling is operational both on original texts when possible –even if annotated by hand, as in the example on Fig. 4– and also on transcriptions and translations. In section I above, we labelled on the transcribed Latin version of the Priory's foundation and below, the translation to the excerpt in Fig. 4 is labelled in the English language. Actors labelled in both examples are summarized in Table II.

*In <Date-UT20 1534, April 24th>, <Att-Ac04 Pope> <**Ac04** Clement VII> announced in his <**UT20** papal bull> to <**Ac05** Joan Bolet> his approval of the initiative to restore the buildings of the Priory and to increase its rents. He also <**UT21** allowed> Bolet to exercise his <**UT22** patronage> <UT22-UT07 upon the> <**UT07** Priory>, <Att-UT22 assuming all financial responsibility as an owner>. The sole <Att-UT22 exception to this absolute control was to use the Priory for non-religious purposes>, as stated in the <**UT19** donation text of 1042>, which was strictly forbidden. Receiving this papal bull was an achievement for <**Ac05** Joan Bolet>, <Ac05-UT23 who had begun> the <**UT23** arrangement of refurbishing work> at the <**UT07** Priory> buildings some time earlier, with the collaboration of neighbouring clerks and bishop's encouragement. The Benedictine abbot of <**UT24** Sant Miquel de Cruïlles> was tolerant with Bolet's plans, and the <Att-Ac00 <Ac06-UT25 prior of> <**UT25** Sant Genís>> <**Ac06** Simó Capellades> was an enthusiastic and indispensable <Ac06-UT23 collaborator> <Ac06-UT05 of the> <Att-Ac05 Barcelonese merchant> as well.* [34] (p. 79)

5. Landscape Analysis and Cartographic Sources

The proposal introduced in this paper offers a useful tool for landscape archaeology as well, and we can label cartographic sources similarly. We must bear in mind that the concept of Unit of Topography –in accordance with the definition proposed– implies a location and date as main attributes for further exploitation. UT/US gathered from selected examples and included in Table III have a precise location expressed as UTM coordinates that have been used to produce and label the maps shown in Fig. 5.

To that extent, Geographic Information Systems (GIS) have a great potential for data processing and exploitation, and they provide a useful tool for the landscape approach to historical knowledge. Providing topographic and chronologic attributes for past entities and events is a conceptual requirement for a spatial turn [37] in History, but also for a general scientific procedure of Past Construction under normalized terms and categories.

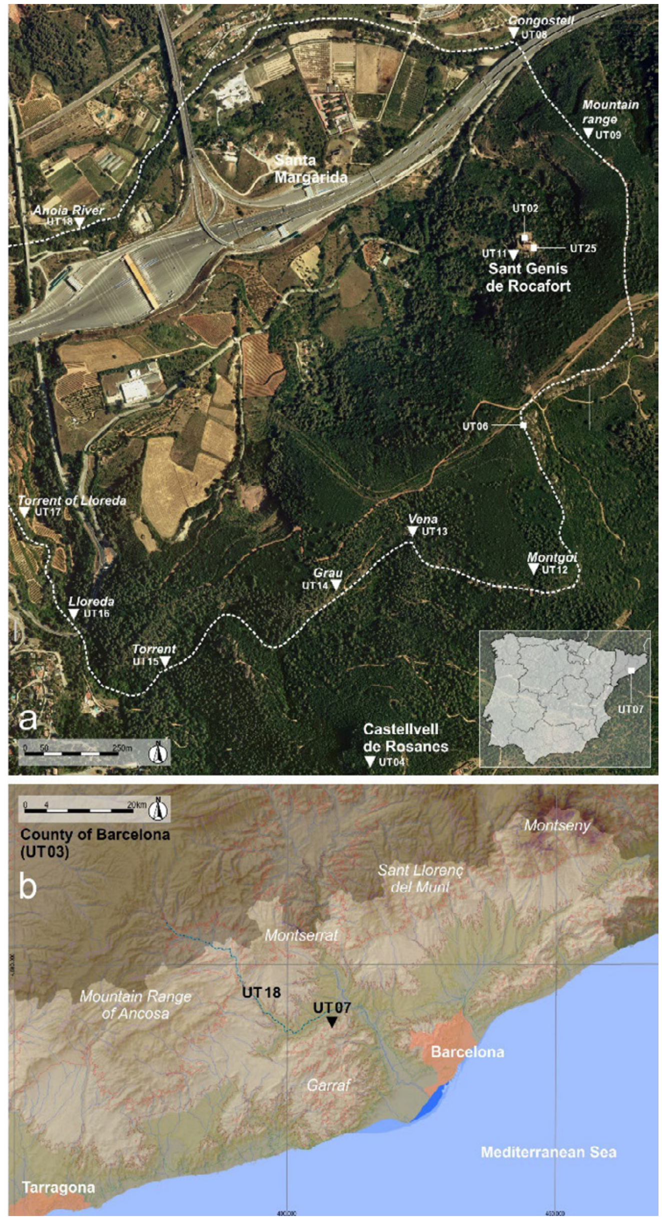


Fig. 5. Aerial view [38] of the Priory of Sant Genís with UT labelling according to the toponymy mentioned in the foundation document (a). Below, UT identification on a general map [29] of the Medieval County of Barcelona (b).

C. Data Exploitation

Cartographic representations and data exploitation by means of GIS technology are both a way to process data and a final representation which may be used for publication and dissemination. Anyway, our proposal –originally thought from the domain of landscape archaeology– goes beyond spatial representations and can focus on temporal sequences or relational data interpretation as well.

Although we summarized data gathered from our lab example in a couple of simple tables, these form part of an ontology-mediated database in which UT, Ac and the relations between them are collected in separate tables. The diagram shown in Fig. 6 represents the main components of a database storing the crucial units of information, their attributes and relations. When keeping this structure, databases can adapt to the needs of a particular research project and show variable interfaces and self-search exploitations, but they will always be interchangeable and potentially interconnected, as far as they share a common data modelling [39]–[40], as shown in Fig. 6.

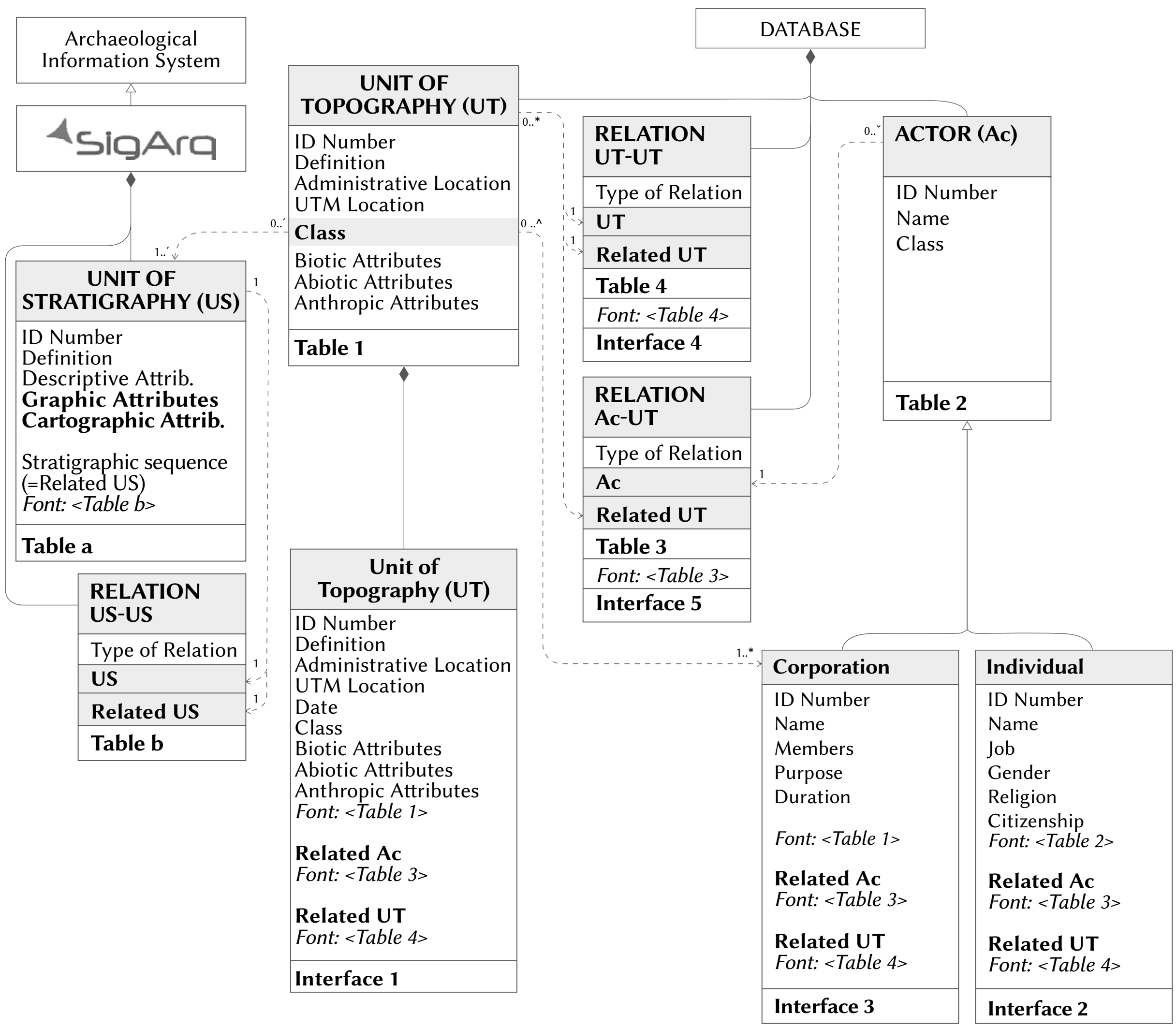


Fig. 6. UML diagram of generic database components.

Fig. 7 represents some interfaces of the current applications used nowadays by our team. One of these is a GIS-based information system created for archaeological purposes [41]–[42]. All of them share the same categories –UT/US/Ac– and hence they allow interdisciplinary research beyond the particular expertise of anyone of us.

When gathering and storing data in the form of tidy-structured tables with variables in columns and observations in rows [43], and according to identified US, UT and Ac regardless of the nature or support of the vestige, multiple representations are possible. Flux diagrams and matrices can visually establish the temporal sequence of activities and their permanence or transformation. Fig. 8 shows an extract of a historical Harris-like [30] matrix created for the Priory of Sant Genís the Rocafort including the Ac, UT exemplified throughout this paper. The archaeological US Harris-matrix routinely developed in archaeology is also included.

This is an example of data exploitation and representation as an interdisciplinary historical matrix, including the archaeological results within the historical discourse arising from written evidence and explaining the Priory's past in a richer construction. Notice how the material vestiges of the buildings' refurbishment in the 16th Century were positively identified within the archaeological register combined with the architectural analysis. In such a representation, Actors mentioned in written vestiges can be assigned to phases and located visually within the corresponding period.

V. Discussion: Towards an Integrated History

The most striking point of using Unit of Topography and Actor as ontological concepts of Historical semantics is that they allow for a truly interdisciplinary research. Unfortunately, today historical science understood as a whole still lacks a common code for data integration within its discourse. The methodological particularities of each method make sense as far as they follow specific goals and socially determined functions. This should not be a problem for creating an integrated construction of the past, as far as they share a common system for information management and exchange, which –unfortunately– has not happened yet [44] (p. 41-42).

In a context of FAIR research, the aim to create an integrated historical discourse is a challenge that historians should face with a sense of urgency. Nevertheless, interdisciplinarity in history does not mean –or should not mean– juxtaposing different past constructions arising from each discipline (history, archaeology, literature, iconography, archival science, linguistics, law, and SSH in general), but creating an interdisciplinary narrative joining the efforts of many different scholars. Sharing a common system for information management and exchange allows us to monitor the research process from the beginning and to locate information precisely, which makes it findable and accessible to colleagues from diverse expertise domains. This is the best way to deal with bias and uncertainty.

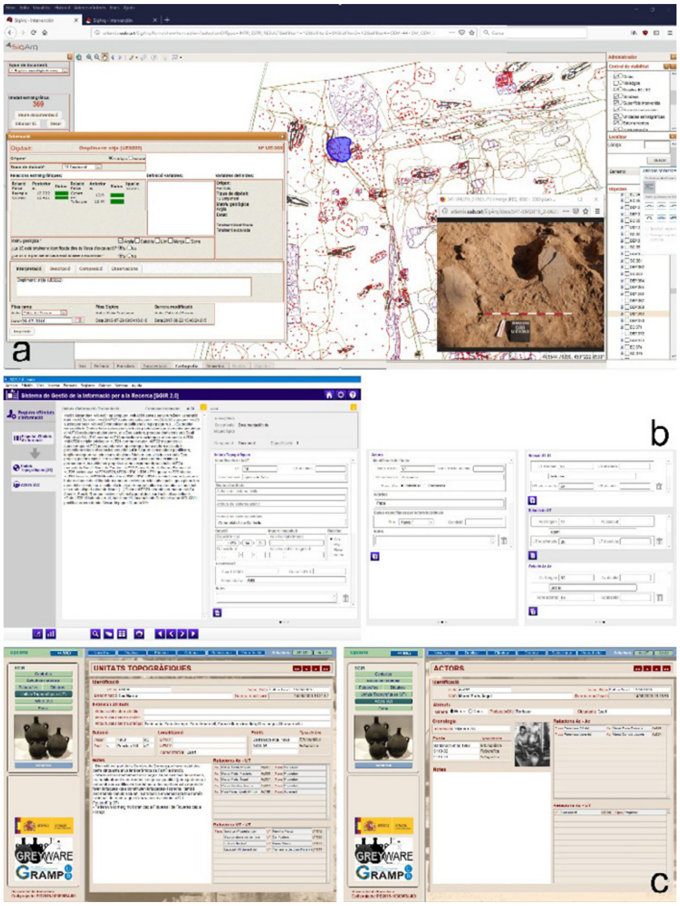


Fig. 7. Examples of record management tools developed by our team. Interface of SigArq software (a); SGIR 2.0 database screen view with forms for UT (left), Ac (centre) and Relation –UT-UT, Ac-UT and AC-AC– (right) (b), and Greyware database interfaces for UT (bottom left) and Ac (bottom right), showing in both cases relations between them as an automatized search from relation tables (c).

Our proposal suggests widening the scope of the archaeological method as a response to this challenge. As defined in [30] (p. 42), the concept of US includes any kind of action leaving a material imprint and identified within spatiotemporal coordinates, no matter if it is positive –adding materials– or negative –removing them. Therefore, during the archaeological fieldwork we register positive US when stratigraphic accumulation of materials occurs and we identify negative US in holes, broken structures or eroded layers.

This concept has proved to be wide enough to be adopted by other archaeology-related disciplines such as architectural analyses of buildings and material heritage studies [45] (p. 79). Since materiality –and, therefore, its cartographic and graphic informative dimension– is the main feature of US, could we define a similar concept equivalent to this unit of information but delinked from its material component? Yes, we could. Units of topography provide this univocal identification of entities and events in the past, with spatiotemporal coordinates and relations between them. Archaeology does not inform about actors, but many other sources of historical information do. Therefore, the proposal of UT/Ac gathering is an adequate compromise solution in order to develop an ontology for past construction in which entities and events are identified through non-ambiguous parameters.

Successful data labelling strategies (TEI) are limited to written sources and hence increasingly used in literature and language studies [46]. They might be successfully applied to the written examples we provided, but they fail in labelling iconographic or photographic vestiges. Textual encoding and labelling tools have a great potential

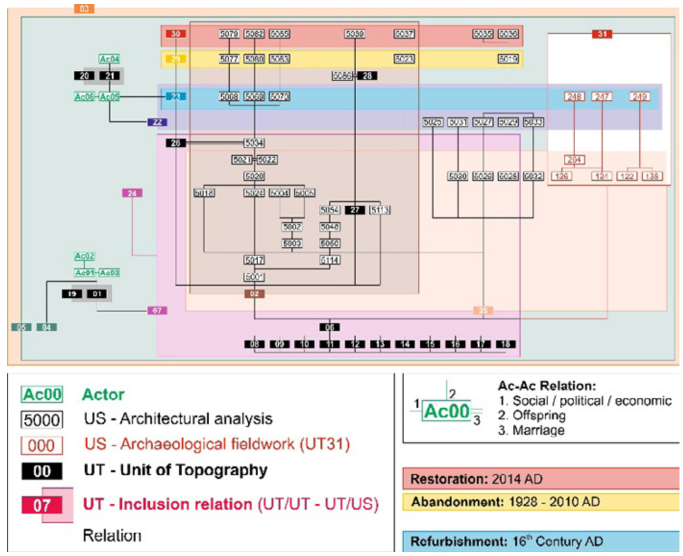


Fig. 8. Excerpt of the Historical matrix created for the archaeological site of Sant Genís de Rocafort. Only UT, Ac and US mentioned throughout this paper as a lab example for the method proposed and summarized in Table II and Table III are included.

for written vestiges from the past, but historical science needs an additional category for data labelling regardless of the origin or support of the vestige. Our labelling proposal implies looking for UT and Ac in too many different shapes and supports –even if textual sources are the most abundant. This actually implies more interpretative knowledge on the historians' part, as it is not possible to detect these data units through mere automatic data labelling applications yet.

Furthermore, UT/Ac identification allows for multiple readings of past vestiges, which can be as exhaustive as required in a particular research project. Data gathered in the examples provided throughout this paper have been enough to demonstrate the validity of the method. We have attempted to find a balance between a theoretical demonstration and a practical case of study, but the historical sources informing about the Priory of Sant Genís de Rocafort are much wider and the historical matrix arising from them is far more complex. This also shows how the historian can read, analyse and interpret past vestiges to a desired level, according to their interest or domain of expertise, and how future experts dealing with the same vestiges can then generate new knowledge *building upon* previous reflections, but not *disregarding* them.

In recent years, data modelling and database construction in the terms described in this paper have allowed us to develop integrated approaches [29], [47] and software [42] overcoming the traditional inconveniences arising from the fragmentation of sources of information. Interactive multimedia and artificial intelligence have a great potential to automatize research processes and have proved to be novel and useful in the domain of SSH. Research projects in the field of History, Archaeology or Archival Science can benefit significantly from shared and transdisciplinary approaches to the past when using a common code. We consider the dialectics between US/UT and Ac as useful categories for data modelling, according to the semantics of Entity and Event as major ontological concepts in historical science [28], [48].

VI. Conclusion

Historical science is a wide discipline that has to consider all the sources of information available, which implies several other disciplines taking part in this process. Archaeology, linguistics, literature, and many others provide valuable data to contribute significantly to the construction of the past. Historians should not add some of these data to a main discourse arising only from written sources but integrate all this information within interdisciplinary processes of data gathering and exploitation.

The NLP community has presented several attempts to process historical knowledge, according to two underpinning –one theoretical and one practical/methodological– ideas: the definition of event and data extraction through text labelling. In doing so initiatives share the common limitation of not having a precise and shared definition of event amongst the academy, and the data extraction procedure being limited to written sources. Furthermore, no attempt has been made to find a domain-specific definition of event combining the historical perspective and ongoing research in the NLP field.

Hybrid intelligence would be, to our perception, a challenging field to explore the possibilities of historical knowledge to become digital and interdisciplinary, and to develop appropriate UT/Ac recognition patterns. NLP systems might be focussed on finding and tagging event-meaningful concepts in written sources, even including archaeological excavation documents and text-supported records, but the heterogeneity of supports and formats for historical vestiges are much wider than these.

The concept of Unit of Stratigraphy, broadly used in archaeology, provides a useful characterization of actions in the past according to their materiality. Widening this idea, the categories Unit of Topography and Actor, as described throughout this paper, provide single and univocal semantic concepts to identify entities and events. Building databases according to these categories is a valuable strategy that integrates knowledge both from SSH and STEM to the historical domain, and made information systems interoperable, ensuring the traceability of the entire research process. While there is a range of opportunities of automatizing processes in terms of text labelling by tagging Units of Topography and Actors, there is still a need for trained and experienced historians who decide the level at which data have to be recorded. Anyway, there is a considerable potential in terms of data exploitation and visualization, in which ICT in general should definitely contribute.

Acknowledgments

This study is part of the current research tasks carried out by the Medieval and Postmedieval Archaeology Research Group GRAMP.-UB (2017-SGR-833-GRC) at the University of Barcelona, to which the authors belong, and it is included in our Landscape Archaeology research line. Theoretical development, conceptualization, and applied methods are part of funded research projects developed by the Centre d'Estudis Martorellencs (CLT009/18/00036 – DGPC/exp. 35) and the University of Barcelona (GREYWARE PID2019-103896RJ-I00). The authors wish to thank Ms Noemí Travé for language edition and review.

References

[1] J. Arostegui. *La investigación histórica: teoría y método (Historical research: Theory and method)*, Barcelona, Spain: Crítica, 1995.

[2] A. Torre. "Un «tournant spatial» en histoire? Paysages, regards, ressources (A spatial turn in history ? Landscapes, visions, resources)", *Annales. Histoire, Sciences Sociales*, vol. 2008, no. 5, pp. 1127-1144, 2018, [On-line] Available: https://www.cairn-int.info/article-E_ANNA_635_1127--a-spatial-turn-in-history.htm

[3] R. White. *What is Spatial History? Spatial History Lab: Working Paper*, Standford, USA: Standford University, 2010, [On-line] Available: https://web.stanford.edu/group/spatialhistory/cgi-bin/site/pub.php?id=29

[4] D. Cosgrove. "Landscape and Landschaft. Lecture delivered at the «Spatial Turn in History» Symposium", *German Historical Institute Bulletin*, vol. 35 no. Fall, pp. 57-71, 2004. [On-line] Available: https://is.cuni.cz/studium/predmety/index.php?do=download&did=130419&kod=JMB067

[5] F. Trentmann. "Materiality in the Future of History: Things, Practices and Politics", *Journal of British Studies*, vol. 48, pp. 283-307, 2009, doi: 10.1086/596123.

[6] D. Hicks. "The material-cultural turn: event and effect" in *The Oxford Handbook of Material Culture Studies*, Hicks, D.; Beaudry, M. C. Eds. Oxford, UK: Oxford University Pres, 2010, pp. 25-98. doi: 10.1093/oxfordhb/9780199218714.013.0002.

[7] E. Hinrichs, M. Hinrichs, S. Kübler, T. Trippel. "Language technology for digital humanities: introduction to the special issue", *Languaje Resources & Evaluation*, vol. 53, pp. 559-563, 2019.

[8] D.W. Oard. "A Whirlwind Tour of Automated Language Processing for the Humanities and Social Sciences" in *Working Together or Apart: Promoting Digital Scholarship*, Washington DC, USA: Council on Library and Information Resources, 2009.

[9] J. Kuhn. "Computational text analysis within the Humanities: How to combine working practices from the contributing fields", *Language Resources & Evaluation*, vol. 53, pp. 565-602, 2019.

[10] M. Piotrowski. *Natural Language Processing for Historical Texts. Synthesis Lectures on Human Language Technologies*, San Rafael, California, USA: Morgan & Claypool Publishers, 2012. [On-line] Available: https://doi.org/10.2200/S00436ED1V01Y201207HLT017

[11] P. Ruiz. *Concept-based and relation-based corpus navigation: applications of natural language processing in digital humanities. Linguistics.* PhD Dissertation. Paris, France: PSL Research University, 2017. [On-line] Available: https://tel.archives-ouvertes.fr/tel-01575167v2

[12] E. Travé, P. Del Fresno, A. Mauri, S. Medina. "The Semantics of historical knowledge. Labelling Strategies for Interdisciplinary and Digital Research in History", in *HI4NLP 2020 Hybrid Intelligence for Natural Language Processing Tasks 2020. Proceedings of the Workshop on Hybrid Intelligence for Natural Language Processing Tasks (HI4NLP 2020)co-located with 24th European Conference on Artificial Intelligence (ECAI 2020) Santiago de Compostela, Spain, August 29, 2020*, P. Gamallo, M. García, P. Martín-Rodilla, M. Pereira-Fariña, Eds. Santiago de Compostela, Spain: CEUR, 2020, pp. 17-21, [On-line] Available: http://ceur-ws.org/Vol-2693/paper2.pdf

[13] J. Delve, R. G. Healey. "Is there a role for data warehousing technology in historical research?" in *Humanities, Computers and Cultural Heritage. Proceedings of the XVI International Conference of the Association of Arts and Sciences*, Royal Netherlands Academy of Arts and Sciences Ed. Amsterdam, Netherlands: KNAW, 2005, pp. 106-111. [On-line] Available: https://dans.knaw.nl/nl/over/organisatie-beleid/publicaties/DANShumanitiescomputersandculturalheritageUK.pdf

[14] A. Ivanovs, A. Varfolomeyev. "Editing an exploratory analysis of medieval documents by means of XML technologies" in *Humanities, Computers and Cultural Heritage*, Netherlands: KNAW, 2005, pp. 155-160.

[15] F. de Jong, H. Rode, D. Hiemstra. "Temporal language models for the disclosure of historical text", in *Humanities, Computers and Cultural Heritage*, Netherlands: KNAW, 2005, pp. 161-168.

[16] S. Eriksson, H. Skånes. "Addressing semantics and historical data heterogeneities in cross-temporal landscape analysis", *Agriculture, Ecosystems and Environment*, vol. 139, pp. 516-521, 2010, doi: 10.1016/j.agee.2010.09.011.

[17] G. Nagypál. "History ontology building: The technical view", in *Humanities, Computers and Cultural Heritage,* Netherlands: KNAW, 2005, pp. 207-214.

[18] B. Tomaszewski. "Local Model Semantics, Categories, and External Representation: Towards a Model for Geo-historical Context", in *Modeling and Using Context. 6th International and Interdisciplinary Conference, CONTEXT 2007, Roskilde, Denmark, August 20-24, 2007, Proceedings. Lecture Notes in Artificial Intelligence*, vol. 4635, B. Kokinov, D. C. Richardson, Th. R. Roth-Berghofer, L. Vieu, Eds. Berlin / Heidelberg, Germany: Springer, 2007, pp. [On-line] Available: http://context-07.ruc.dk/Context2007DocCons.pdf

[19] H. Berger. 'Microhistory and quantitative data analysis', In *Humanities, Computers and Cultural Heritage,* Netherlands: KNAW, 2005, pp. 39-42.

[20] A. Antonacopoulos, D. Karatzas. "Semantics-Based Content Extraction in Typewritten Historical Documents", in *Proceedings of the 8th International Conference on Document Analysis and Recognition (ICDAR2005)*, IEEE Computer Society Ed. Washington DC, United States: IEEE Digital

Library, 2005, pp. 48-53, doi: 10.1109/ICDAR.2005.215.

[21] M. Heller, G. Vogeler. "Modern information retrieval technology for historical documents", in *Humanities, Computers and Cultural Heritage,* Netherlands: KNAW, 2005, pp. 143-148.

[22] C. Coffin. "The voices of History: Theorising the interpersonal semantics of historical discourses", *Text*, vol. 22, no. 4, pp. 503-508, 2002, doi: 10.1515/text.2002.020.

[23] G. Nagypál, B. Motik. "A Fuzzy Model for Representing Uncertain, Subjective, and Vague Temporal Knowledge in Ontologies", in *On The Move to Meaningful Internet Systems 2003: CoopIS, DOA, and ODBASE. OTM 2003. Lecture Notes in Computer Science*, vol. 2888, R. Meersman, Z. Tari, D. C. Schmidt, Eds. Berlin / Heidelberg, Germany: Springer, 2003, pp. 906-923, doi: 10.1007/978-3-540-39964-3_57.

[24] M. D. Wilkinson et al. "The FAIR Guiding Principles for scientific data management and stewardship", *Scientific Data*, vol. **3**, paper 160018, 2016, doi: 10.1038/sdata.2016.18

[25] G. Gänser, G. Michetti. *InterPARES Trust Project. Ontology of Functional Activities for Archival Systems*; TR05-Final Report-20180526, 2018. [On-line] Available: https://interparestrust.org

[26] Parthenos. *Guidelines to FAIRify Data Management and Make Data Reusable.* [On-line] Available: https://zenodo.org/record/2668479

[27] A. Mosca. "Ontology-mediated Data Integration and Access in Research and Innovation Policy", in *Proceedings of the Joint Ontology Workshops 2017 (JOWO 2017), Episode 3: The Tyrolean Autumn of Ontology, Bozen-Bolzano, Italy, 21–23 September 2017,* S. Borgo, O. Kutz, F. Loebe, F. Neuhaus, Eds. Bozen-Bolzano, Italy: CEUR, 2017, [On-line] Available: http://ceur-ws.org/Vol-2050/invited4-Mosca.pdf

[28] K. Thibodeau, "The Construction of the Past: Towards a Theory for Knowing the Past", *Information*, vol. 10, paper 332, 2019, doi: 10.3390/info10110332.

[29] A. Mauri. *La configuració del paisatge medieval: el comtat de Barcelona fins al segle XI (The formation of medieval landscape: The county of Barcelona until the 11th Cent). PhD Dissertation.* Barcelona, Spain: Universitat de Barcelona, 2006. [On-line] Available: https://www.tesisenred.net/handle/10803/2065

[30] E. C. Harris. *Principles of archaeological stratigraphy (Second edition)*, San Diego, USA: Academic Press, 1989, [On-line] Available: https://www.sciencedirect.com/book/9780123266514/principles-of-archaeological-stratigraphy

[31] A. Carandini. *Historias en la Tierra: Manual de Excavación Arqueológica (Stories from the earth. Excavation Handbook)*, Barcelona, Spain: Crítica, 1997.

[32] *Foundation Act of the Priory of Sant Genís de Rocafort.* Barcelonese Diocesan Archive, Vol. *Dotaliarum*, no. 5, pp. 31-32.

[33] E. Mallorquí, *El monestir de Sant Miquel de Cruïlles a l'edat mitjana. Estudi i edició dels documents. 1035-1300 (The monastery of Sant Miquel de Cruïlles in the middle age. Study and edition of documents. 1035-1300)*, Girona, Spain: Ajuntament de la Bisbal d'Empordà, 2000.

[34] J. Baucells. *El Priorat de Sant Genís de Rocafort (Martorell) (The Priory of Sant Genís de Rocafort)*, Centre d'Estudis Martorellencs. Ed. Martorell, Spain: CEM, 2007.

[35] J. Salvany. *Sant Genís de Rocafort.* Biblioteca de Catalunya Ed. Barcelona, Spain: BC, 1915.

[36] J. Salvany. *Sant Genís de Rocafort.* Biblioteca de Catalunya Ed. Barcelona, Spain: BC, 1912.

[37] I. Hodder, C. Orton. *Spatial Analysis in Archaeology*, Cambridge, UK: Cambridge University Press, 1976.

[38] *Sant Genís de Rocafort. Ortophoto.* Provided by Institut Cartogràfic i Geològic de Catalunya. Barcelona. [On-line] Available: http://www.icc.cat/vissir3/

[39] C. González-Pérez. "A Conceptual Modelling Language for the Humanities and Social Sciences", in *Sixth International Conference on Research Challenges in Information Science (RCIS),* IEEE Computer Society Ed. Washington DC, United States: IEEE Digital Library, 2012, doi: 10.1109/RCIS.2012.6240430.

[40] E. Travé, P. Del Fresno, A. Mauri. "Ontology-mediated Historical Data Modelling: Theoretical and Practical Tools for an Integrated Construction of the Past", in *Digital Humanities Special Issue.* C. González-Perez Ed. *Information*, vol. 11, paper 182, doi:10.3390/info11040182.

[41] P. Del Fresno. *Sistema de Información Arqueológica: propuesta de normalización, desarrollo conceptual e informático. (Archaeological Information System: Normalization proposal, conceptual and computing development) PhD Dissertation.* Vitoria, Spain: University of the Basque Country, 2016. [On-line] Available: https://www.academia.edu/23836202/

[42] *SigArq. Sistema de Información Geoespacial Arqueológico.* [On-line] Available: http://www.sigarq.es/home/

[43] H. Wickham, 'Tidy Data', *Journal of Statistical Software*, vol. 59, no.10, pp. 1-10, 2004, doi: 10.18637/jss.v059.i10.

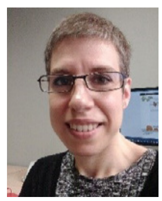
[44] A. Mauri, E. Travé, P. Del Fresno. '"An integrated implementation of written and material sources – conceptual challenge and technological resources", in *Archaeology. New Approaches in Theory and Techniques*, I. Ollich Ed. Rijeka, Croatia: InTech, 2012, pp. 41-64, doi:10.5772/38934.

[45] P. Del Fresno, A. Mauri, E. Travé. "Documentació d'intervencions de restauració en jaciments arqueològics i patrimoni edificat: cap a la construcció d'un Sistema d'informació (Recording restoration tasks in archaeological sites and building heritage: towards the construction of an information system)". *Documentar en la conservació-restauració. Mètodes i noves technologies*, Conservadors Restauradors Associats de Catalunya. Barcelona: Spain, CRAC, 2020, in press.

[46] C. M. Sperberg-McQueen, L. Burnard. *TEI P5: Guidelines for Electronic Text Encoding and Interchange TEI.* Text Encoding Initiative Consortium, 2019. [On-line] Available: https://tei-c.org/release/doc/tei-p5-doc/en/Guidelines.pdf

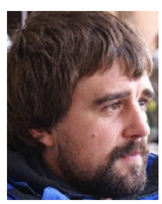
[47] E. Travé. "Los hornos medievales y la cerámica negra" in *Actas del IXI Congreso de la Asociación de Ceramología. Obra negra y alfarería de cocina*, Asociación de Ceramología Ed., Quart, Spain: Ajuntament de Quart, 2017, pp. 123-147. [On-line] Available: http://www.ceramologia.org/gestion/archivos/1084.Est.pdf

[48] R. Sprugnoli, S. Tonelli. "One, no one and one hundred thousand events: Defining and processing events in an inter-disciplinary perspective", *Natural language engineering*, vol. 23, no. 4, pp. 485-503, 2016, doi: 10.1017/S1351324916000292.



Esther Travé Allepuz

Esther Travé holds a Degree in History (Universitat de Barcelona, Spain, 2005) and a PhD in Medieval History at the same university (2009). She is a medieval archaeologist and her research field is the production and distribution of greyware pottery in medieval and postmedieval ages. She was trained in Ceramic Petrography at the University of Sheffield (2008) and in XRF and XRD techniques at the Institute of Archaeology, University College London (UCL) (2012-2014). She lectures at the University of Barcelona since 2015, and she currently is leading researcher of the archaeological project developed at the Sites of Santa Margarida and Sant Genís de Rocafort (Martorell). Her most recent research has been devoted to data analysis and the development of interdisciplinary approaches applied, in particular, to the historical and archaeological study of pottery.

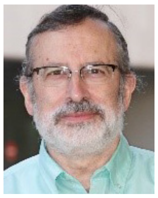


Pablo del Fresno Bernal

Pablo Del Fresno obtained his Degree in History at the University of the Basque Country (Spain, 2000). He holds an MSc in Geographic Information Technologies (Universitat Autònoma de Barcelona, Spain, 2008) and a PhD in History (University of the Basque Country, 2016). He works as a professional archaeologist at Sistemes de Gestió de Patrimoni SCCL. Based upon the theoretical reflection of his PhD dissertation, he is the development coordinator of the Archaeological Information System SigArq, within a long collaborative framework in which the Geographic Information and Remote Sensing Lab (LIGIT) at the UAB takes part. He participates in several archaeological projects in which the SigArq system is involved. Amongst these, Santa Catalina (Mansilla de la Sierra, La Rioja), the Convent de Penyafort (Santa Margarida i els Monjos, Barcelona), and the Archaeological Research Project of Santa Margarida and Sant Genís de Rocafort (Martorell, Barcelona), are the more relevant.

Sonia Medina Gordo

Sonia Medina is a PhD candidate working under Esther Travé's supervision at the University of Barcelona. Her PhD project aims at developing digital strategies for data management and exploitation and applying them to the historical analysis of the formation of the Castilian Feudal System at the Upper Arlanza Basin (Burgos, Spain). She holds a Degree in History (Universitat de Barcelona, Spain, 2018) and a MA in Advanced History Studies (Universidad de Salamanca, Spain, 2019). Approaching the medieval period according to an interdisciplinary method integrating results from historical science and archaeology led her to join this research team. She contributes to the data modelling and database development, the way of dealing with uncertainty and bias in historical discourses being one of her main interests.



Alfred Mauri Martí

Alfred Mauri has a long career in Archival Science and Archaeology since he obtained his Degree in Prehistory, Ancient History and Archaeology (Universitat de Barcelona, Spain, 1983). From 1981 to 2002 he worked in Martorell (Spain), as an archivist in charge of the Town Council Archive, while he was the leading researcher at the Archaeological Site of Santa Margarida, located in the same town. He performed his PhD Dissertation in the field of Landscape Archaeology and developed the concept of Unit of Topography. After obtaining his PhD (Universitat de Barcelona, Spain, 2006) he has been Lecturer and Degree coordinator at the High School of Archival Science and Records Management of the Universitat Autònoma de Barcelona. He has been director of this same centre for three years (2016 – 2018). His main research lines are landscape archaeology and archeaology of architecture, the integrated and georeferenced historical data management, and Big Data analysis and exploitation.

Smoke Test Planning using Answer Set Programming

Tobias Philipp[1], Valentin Roland[1], Lukas Schweizer[2] *

[1] SINA Development and Verification Team, Division Defence & Space, secunet Security Networks AG, Essen (Germany)

[2] Computational Logic Group, Technische Universität Dresden, Dresden (Germany)

Received 10 November 2020 | Accepted 23 November 2020 | Published 10 February 2021

Abstract

Smoke testing is an important method to increase stability and reliability of hardware-dependent systems. Due to concurrent access to the same physical resource and the impracticality of the use of virtualization, smoke testing requires some form of planning. In this paper, we propose to decompose test cases in terms of atomic actions consisting of preconditions and effects. We present a solution based on answer set programming with multi-shot solving that automatically generates short parallel test plans. Experiments suggest that the approach is feasible for non-inherently sequential test cases and scales up to thousands of test cases.

Keywords

Planning, Answer Set Programming, Testing.

DOI: 10.9781/ijimai.2021.02.003

I. Introduction

Quick development-test-cycles are vital for the development of functional and reliable software. Thus, it is considered best-practice to frequently run a subset of regression tests, called smoke tests, to catch as many issues as possible as early as possible [1]. In contrast to regression testing, smoke tests lean towards minimizing the time spent on a test run, sacrificing coverage if necessary. To reduce overall run time, software tests are often run concurrently on isolated instances of the *system under test* (SUT). This works well if the software system can run on virtualized infrastructure, which can be easily scaled.

However, some systems depend on real hardware and thus virtualization is impossible. Furthermore, functional and performance characteristics of a virtualized system may significantly differ from operations on bare metal, which would render performance and load tests meaningless.

Smoke testing of hardware or hardware-dependent systems requires some form of planning: A running test inevitably changes the state of the SUT and thus the hardware state. Therefore, isolation of test cases is difficult, since multiple tests may depend on the same physical resources. Consequently, tests can neither be easily parallelized nor chained and some execution strategy is required.

A straightforward solution is to bring the SUT to a clean state after each test has run individually, e.g. by a cold reboot. For a number of reasons, this is not satisfactory: Resetting the SUT to a known state is often difficult or time-consuming. Issues, such as hidden assumptions on the concurrent use of different SUT features, may not occur with a strictly sequential execution.

* Corresponding author.

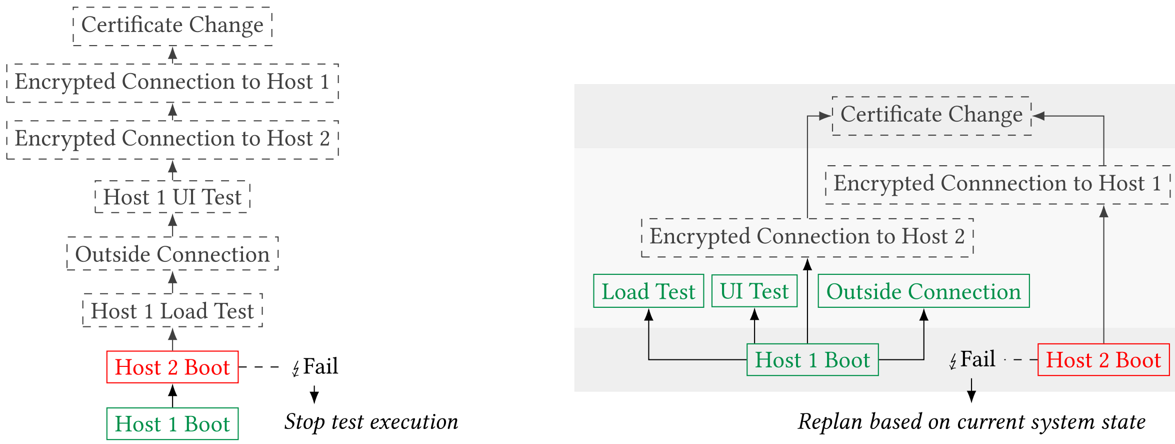
E-mail addresses: tobias.philipp@secunet.com (T. Philipp), science@vroland.de (V. Roland), lukas.schweizer@tu-dresden.de (L. Schweizer).

In this paper we propose to decompose test cases into atomic actions with preconditions and effects such that we can automatically infer a suitable test plan that is short and parallel. Each action is a small, specific instruction performed on the SUT, like “boot system X” or “connect to host Y”. Preconditions describe the state of the SUT in which an action is executable. Effects describe the changes of the state of the SUT after successful execution of the action. Furthermore, actions are associated with instructions determining their failure or success.

In our proposed method, this information is given declaratively to enable the automated generation of short parallel test plans. We demonstrate that solving the planning problem can be done using *Answer Set Programming* (ASP) and show that the system can solve planning problems with thousands of actions on non inherently-sequential problems. The main benefits of this approach include:

1. *Reduced overall execution times*: Consider Fig. 1 that compares test executions with automated planning to a predefined sequential test plan in which a secure network connection is established, user interface tests are performed and a certificate is exchanged. Note that the parallel plan consists of four time points whereas the sequential needs eight.
2. *Maximal test execution*: In case of failing tests, as many tests as possible can be executed by replanning: Assuming that the action “Host 2 Boot” fails, the predefined sequential test plan fails early, whereas one can construct a new test plan that allows the continued execution of tests that do not depend on host 2. Consequently, the state of the SUT is preserved, no reboots are necessary and we do not need to (re-)execute passed tests.
3. *Specification reuse*: Declarative definition of actions and a flexible planning framework allows reusing actions in regression tests, integration tests, fixed scenarios or component-wise testing trough different planning goals and constraints.

Due to the nature of testing we deal with uncertainty in the outcome of actions as well as in the SUT state itself, which captures the ideas of conformant planning [2]. Moreover, maximal test execution can



(a) Sequential execution.

(b) Parallel test execution with replanning.

Fig. 1. Comparison of test execution with sequential and parallel plans, in the scenario of the action "Host Boot 2" failing.

be seen as some form of oversubscription planning [3], which aims to achieve as many sub-goals as possible instead of a conjunctively defined (single) top-goal.

The paper is structured as follows: We briefly describe ASP in Section II, before a formal description of smoke test planning is then given in Section III. The ASP encoding and its correspondence to tests is then presented in Section IV. In Section V, we present an experimental evaluation on the basis of a family of benchmarks. Finally, we describe related work and conclude.

II. Preliminaries

ASP is a declarative modeling and problem solving framework that combines techniques of knowledge representation and database theory. Two of the main advantages of ASP are its expressiveness [4] and, when using non-ground programs, its advanced declarative problem modeling capability. Such programs must be transformed to a ground program prior to solving.

We review the basic notions of answer set programming [5] under the stable model semantics [6], and refer to introductory literature [4], [7], for a more comprehensive introduction.

We fix a countable set $\mathcal{U}$ of (domain) elements, also called constants; and suppose a total order $<$ over the domain elements. An atom is an expression $p(t_1, ..., t_n)$, where p is a predicate of arity $n \geq 0$ and each ti is either a variable or an element from $\mathcal{U}$. An atom is ground if it is free of variables. $B_{\mathcal{U}}$ denotes the set of all ground atoms over $\mathcal{U}$. A (normal) rule ρ is of the form

$$a \leftarrow b_1, \ldots, b_k,\ not\ b_{k+1}, \ldots,\ not\ b_m$$

with $m \geq k \geq 0$, where a, b_1, ..., b_m are atoms, and "not " denotes *default negation*. The head of ρ is the singleton set $H(\rho) = \{a\}$ and the body of ρ is $B(\rho) = \{b_1, ..., b_k, \text{not } b_{k+1}, ..., \text{not } b_m\}$. Furthermore, $B^+(\rho) = \{b_1, ..., b_k\}$ and $B^-(\rho) = \{b_{k+1}, ..., b_m\}$. A rule ρ is safe if each variable in ρ occurs in $B^+(r)$. A rule ρ is ground if no variable occurs in ρ. A fact is a ground rule with empty body. An (input) database is a set of facts. A (normal) *program* is a finite set of normal rules. For a program Π and an input database D, we often write $\Pi(D)$ instead of $D \cup \Pi$. For any program Π, let U_Π be the set of all constants appearing in Π. $Gr(\Pi)$ is the set of rules $\rho\sigma$ obtained by applying, to each rule $\rho \in \Pi$, all possible substitutions σ from the variables in ρ to elements of U_Π.

An interpretation $I \subseteq B_{\mathcal{U}}$ *satisfies* a ground rule ρ iff $H(\rho) \cap I \neq \emptyset$ whenever $B^+(\rho) \subseteq I$, $B^-(\rho) \cap I \neq \emptyset$. I satisfies a ground program Π, if each $\rho \in \Pi$ is satisfied by I. A non-ground rule ρ (resp., a program Π) is satisfied by an interpretation I if I satisfies all groundings of ρ (resp., $Gr(\Pi)$). $I \subseteq B_{\mathcal{U}}$ is an *answer set* (also called *stable model*) of Π if it is the subset-minimal set satisfying the *Gelfond-Lifschitz reduct* $\Pi^I = \{H(\rho) \leftarrow B^+(\rho) \mid I \cap B^-(\rho) = \emptyset, \rho \in Gr(\Pi)\}$ [6]. For a program Π, we denote the set of its answer sets by $\mathcal{AS}(\Pi)$.

We make use of further syntactic extensions, namely integrity constraints and choice rules, which both can be recast to ordinary normal rules as described in [8]. An *integrity constraint* is a rule ρ where $H(\rho) = \emptyset$, intuitively representing an undesirable situation; i.e. it has to be avoided that $B(\rho)$ evaluates positively. Choice rules are of the form $l\ \{a : a_1, ..., a_i\}\ u$, where a is an atom and $a_j = p_j$ or $a_j = not\ p_j$, for a_j an atom, $1 \leq j \leq i$, l and u are non-negative integers, and the expression $\{a : a_1, ..., a_n\}$ denotes the set of all ground instantiations of a, governed through $\{a_1, ..., a_n\}$. Intuitively, an interpretation satisfies a choice rule, if $l \leq N \leq u$ holds, where N is the cardinality of any subset of $\{a : a_1, ..., a_i\}$. In our encodings, we further use $-a$, to denote the classical negation of an atom a, though '−' is not an operator present in the introduced language and merely represents syntactic sugar[1].

Programs in this paper are given in the input language of Clingo [9], a state-of-the-art system combining an ASP grounder and solver. However, rule head and body are separated with ⟵ instead of :- for improved readability.

III. Smoke Tests as Planning Problems with Unknowns

In the following, we describe the smoke test planning and optimization problem in terms of a planning problem in which the state of SUT properties may be unknown and actions can be executed in parallel. As executing many actions in a short time span is crucial, we consider *parallel plans*. Here, we consider a new semantics based on the ∀-*step semantics* [10], in which actions may be executed simultaneously, as long as they can be executed in any sequential order resulting in a unique state.

A. The System State as Fluents

We describe the state of the SUT in terms of *fluents* that correspond to atomic properties of the system. The set of fluents $\mathcal{F}$ is a designated subset of the set of terms $\mathcal{T}$. For instance, the fluent system_up(host1) states that the machine host1 is running. In a *system state* S, a fluent f can hold, not hold or it may be unknown, whether f holds or not. This is indicated by unary operators +, - and unknown for *holds*, *does not hold* and *unknown*, respectively. If f is a fluent, then + f, - f and

[1] Note that $-a$ merely represents a special predicate name, and together with an integrity constraint ⟵ $-a$, a, the behavior of classical negation can be emulated.

unknown f are fluent literals. For instance, the fluent literal -system_up(host1) denotes that host1 is not running. The set of all *fluent literals* is referred to as $\mathcal{FL}$. Formally, we represent the *state* S_t of the *SUT at time t* as a finite set of fluent literals. Initially, the presence of all fluents is unknown, thus $S_0 = \{\text{unknown } f \mid f \in \mathcal{F}\}$.

The function fluents: $\mathcal{P}(\mathcal{FL}) \rightarrow \mathcal{P}(\mathcal{F})$ maps a set of fluent literals to the set of fluents occurring in that set. The function lits : $\mathcal{P}(\mathcal{F}) \rightarrow \mathcal{P}(\mathcal{FL})$ maps a set of fluents F to the corresponding set of possible fluent literals $\{+ f, - f, \text{unknown } f \mid f \in F\}$.

B. Test Cases as Actions

Smoke test cases are usually specified at varying levels of complexity and abstraction. A test describing a user interface workflow is semantically different from testing a single button press. To reason about SUT behavior for test planning, we need a unified, fine grained view on how the system state is changed. Thus, we decompose test cases into *actions*. Such actions represent small, self-contained operations. In the context of the test planning problem, we consider only the information relevant for planning.

Definition 1 (Action). An *action* a is a tuple (P, E, M), where

- $P \subseteq \mathcal{FL}$ is a finite set of *preconditions*, which do not contain fluent literals of the form unknown f,
- $E \subseteq \mathcal{FL}$ is a finite set of *effects*, and
- $M \subseteq \mathcal{F}$ is a finite set of *modifications*, which is the set of fluents an action may change if the action *fails*.

The set of all actions is denoted by $\mathcal{A}$. An action (P, E, M) is executable in state S if $P \subseteq S$, i.e. its preconditions are satisfied.

Definition 2 (Successor State). The *successor state* S_{i+1} of S_i after executing a finite set of actions A, denoted by $S_i \xrightarrow{A} S_i+1$ is defined as follows:

$$S_{i+1} = \Big(S_i \setminus \text{lits}\,\big(\bigcup_{a \in A} \text{fluents}(E_a)\big)\Big) \cup \bigcup_{a \in A} E_a$$

C. Plan Semantics

For all permutations $(a_1, \ldots, a_n)$ of a set of actions A_t, their sequential application $S_i \xrightarrow{a_1} S_{i+1} \xrightarrow{a_2} \ldots \xrightarrow{a_n} S_G$ must always result in a unique state S_G.

Definition 3 (Plan). A plan P for a set of actions $\mathcal{A}$ is a finite sequence of slots $P = \langle A_0, \ldots, A_n \rangle$.

Intuitively, all actions must be executable when applied, are mutually non-interfering, and each action a occurs at least once in a slot A_t for some $t \in \{0 \ldots n\}$.

Definition 4 (Non-interference). Let $a_1 = (P_1, E_1, M_1)$ and $a_2 = (P_2, E_2, M_2)$ be two actions. Then, a_1 and a_2 are *non-interfering*, if the following mutually holds:

1. $\text{fluents}(E_1) \cap \text{fluents}(E_2) = \emptyset$,
2. $\{+ f \mid f \in P_1\} \cap \{- f, \text{unknown } f \in E_2\} = \emptyset$,
3. $\{- f \mid f \in P_1\} \cap \{+ f, \text{unknown } f \in E_2\} = \emptyset$, and
4. $(\text{fluents}(P_1) \cup \text{fluents}(E_1) \cup M_1) \cap M_2 = \emptyset$.

Intuitively, the conditions describe the following: Condition 1. excludes concurrent modification of system properties. From the point of view of classical planning, this constraint seems too restrictive as it disallows executing actions in parallel although any order of execution results in the same final state [10]. However, in the context of smoke testing, actions represent state transitions of physical systems and fluents describe its properties. Therefore, a change of the presence of a fluent indicates concurrent access to physical resources that may be unsafe. Conditions 2. and 3 ensure that no action affects the precondition of other, concurrently executed actions. This assumes that a fluent literal which is present in the precondition and effects of an action always holds during the execution of an action. If this is not the case, the fluent must also appear in the set of *modifications* of that action. Condition 4 guarantees that two actions running in parallel do not modify fluents which are modified by another action. Thus, their parallel execution is independent of each other, in particular in case of failing actions. These conditions guarantee the prevention *false positive* or *false negative* test results.

Note that by defining non-interference and *executability* in this fashion, we model uncertainty in the SUT state as the third fluent state unknown. This is different to approaches like conformant planning [2], [11], where uncertainty in the initial state is modeled as a set of possible states. While one could construct such a *belief-state* for a state S, the initial state is not the only source of uncertainty. Even when successful, actions can introduce uncertainty through their effects: For instance, consider an action which re-starts the SUT. While some properties of the SUT will be known, some components may be in an uninitialized state, thus the related fluents are made unknown in the action's effects.

D. Planning Goal

In contrast to classical planning problems in which a desired goal state is to be reached, the goal in this paper is to find a shortest parallel plan that executes all actions. We call such a plan *complete*. In the case that there is no such plan as many actions as possible should be executed, which can be seen as some form of oversubscription planning [3]. While the system state determines which actions can be executed at any point in time, there is *no specified goal state*. The planning goal is a property of the plan itself, not of the system under test.

Example 1. Consider an arrangement of two devices, where a subset of the activities shown in Fig. 1 are tested: 1. The devices connect to each via encrypted network connections, 2. each device connects to the outside network, and 3. the devices stay connected if the encryption parameters of their connection are changed. Consider the following set of fluents that describe the relevant system properties $\mathcal{F}$ and the set of actions $\mathcal{A}$ as described in Table I:

$\mathcal{F}$ = {system_up(host1), system_up(host2)}

∪ toutside_up(host1), outside_up(host2)u

∪ tenc_conn_up(host1), enc_conn_up(host2)u

∪ tconn_to(host1, host2), conn_to(host2, host1)u

We can obtain the plan $P_e = \langle A_0, \ldots, A_3 \rangle$ with the following considerations:

- In state $S_0 = \{\text{unknown } f \mid f \in \mathcal{F}\}$, we can see that only boot(host1) and boot(host2) are *executable* in A_0. Since precondition and modifications are empty, 2, 3 and 4 are satisfied. As both actions each only affect one of the hosts, respectively, they are also *non-interfering*. Thus, we can schedule both actions in A_0 = {boot(host1), boot(host2)}
- With $S_0 \xrightarrow{A_0} S_1 \supseteq$ {+system_up(host1), unknown enc_conn_up(host1)}, the actions connect_en-crypted(host1), connect_encrypted(host2), connect_out-side(host1) and connect_outside(host2) get executable. Through checking for non-interference, we can see that they can all be scheduled for A_1.
- In S_2, change_conn(host1) as well as change_conn(host2) are executable. However, they mutually interfere due to conditions 1 and 4. Thus, we need to schedule them sequentially, for instance as A_2 = {change_conn(host1)} and A_3 = {change_conn(host2)}.
- The resulting plan $P_e = \langle A_0, \ldots, A_3 \rangle$ is a shortest plan.

TABLE I. Representation of the Example in Terms of Actions. The Variables *H1*, *H2* ∈ {host1, Host2} Are Placeholders for Any Concrete Host, Where *H1* and *H2* Are Distinct

Action	**Precondition *P***	**Effects *E***	**Modifications *M***
boot(H1) – Boot up *H1*.	∅	+ system_up(H1) unknown enc_conn_up(H1) unknown outside_up(H1) unknown conn_to(H1, H2)	∅
connect_outside(H1) – H1 initiates an (unencrypted) connection to an outside network.	+ system_up(H1)	+ outside_up(H1)	∅
connect_encrypted(H1) – *H1* initiates an encrypted connection to the other host *H2*.	+ system_up(H1)	+ conn_to(H1, H2) + enc_conn_up(H1)	∅
change_conn(H1) – *H1* initiates a change of encryption parameters, e.g. renews a certificate.	+ system_up(H1) + enc_conn_up(H1) + conn_to(H1, H2) + conn_to(H2, H1)	+ conn_to(H1, H2) + conn_to(H2, H1)	enc_conn_up(H1) enc_conn_up(H2) conn_to(H1, H2) conn_to(H2, H1)

E. Replanning

Consider a plan $\langle A_0, ..., A_{t-1}, A_t, A_{t+1}, ..., A_\ell \rangle$, the corresponding states $S_0, ..., S_\ell$ and suppose action $e = (P, E, M)$ in A_t fails. In the case that $E = \varnothing$ and $M = \varnothing$, the SUT state is consistent with the plan. Otherwise, actions in the slots $A_{t+1}, ..., A_\ell$ could be non-executable and therefore, a new plan involving as many actions as possible that have not been executed is generated.

Replanning follows the same rules outlined in Section C, but with a different initial state. Intuitively, the new initial state is constructed by applying the successful actions as scheduled in previous plan, but making any fluent the failed action may have changed unknown. Formally, let $F_{\text{affected}} = \text{fluents}(E_e) \cup M_e$. Then, the new initial state is derived from the system state S_{t-1} before the action failed as follows:

$$S = \left(S_{t-1} \setminus \text{lits}\left(\bigcup_{a \in A_t} \text{fluents}(E_a) \cup F_{\text{affected}} \right) \right)$$
$$\cup \{\text{unknown } f \mid f \in F_{\text{affected}}\} \cup \bigcup_{a \in A_t} E_a \setminus \text{lits}(F_{\text{affected}})$$

F. Finding Maximal Runnable Subsets

A complete plan may not exist, in particular in case of replanning as a failed test is excluded. For instance, an action that cannot be executed because all sequences of actions making its precondition true contain some action which has already failed. However, with the overall goal to quickly execute as many actions as possible, a *partial* plan can be constructed instead, which executes *as many outstanding actions as possible*. We can exploit the flexibility of a declarative approach to obtain such a maximal partial plan, by maximizing a score function instead of solving for satisfiability of a set goal. As a score function for a plan P of length n, we use the number of distinct outstanding actions $g(P)$ in P.

$$g(P) = \left| \bigcup_{t=1}^{n} \{a \mid a \in A_t, a \text{ not previously executed}\} \right|$$

IV. An ASP Program Using Multi-Shot Solving

For solving the smoke test planning problem, we phrase the planning and replanning problem in terms of an Answer Set Program and use Clingo [9] for solving, a state-of-the-art system combining an ASP grounder and solver supporting incremental solving [12] and assumptions. Incremental grounding allows to extend the logic program after initial grounding, by adding parameterized *subprograms*. Assumptions are realized with *external atoms* that allow to change their truth value after grounding. Together, these mechanisms allow to reuse a single ground program over the entire testing process, supplying initial conditions and goal condition via external atoms. We use the atoms as presented in Table II to encode the planning program which is divided into four *subprograms*: the instance program, the base program, the transition program, and the goal program.

TABLE II. ASP Predicates and Their Meaning: I Denotes May Atoms That May Be Added Incrementally, E Denotes May External Atoms

Atom	**I**	**E**	**Description**
apply(A,t)	✓		applies action A at time t
demands(A,F,true)			action A requires + F
demands(A,F,false)			action A requires - F
adds(A,F)			action A has positive effect + F
deletes(A,F)			action A has negative effect - F
invalidates(A,F)			action A makes F unknown F
modifies(A,F)			action A may modify fluent F
pc_changes(A, F)			action A changes fluent F
interfere(A1, A2)			actions A1 and A2 mutually interfere
available(A)			action A can be used for in a plan
has_failed(A)		✓	the execution of A has failed
need_to_plan(A)		✓	action A must occur in the plan
add(F, t)	✓		fluent F is made true at time t
del(F, t)	✓		fluent F is made false at time t
inv(F, t)	✓		fluent F is made unknown at time t
holds(F, 0)	✓	✓	fluent F holds initially
holds(F, t)	✓		$t > 0$, fluent F holds at time t
was_applied(A, t)	✓		action A has been applied at least once at t
goal_horizon(t)	✓	✓	the current horizon is t

1. Instances

As the basis of the planning program, we represent actions and the initial state as ASP facts. This part of the program is specific to the supplied set of actions and, in case of replanning, the current SUT state.

An action *a* is represented as an action(a) atom. Its precondition, effects and modifications are expressed using the atoms demands(..), adds(..), deletes(..), invalidates(..), modifies(..), respectively. If the initial state S_0 is different from {unknown $f \mid f \in \mathcal{F}$}, it is specified through holds(f, 0) and -holds(f, 0) atoms. Since variables are capitalized in the input language of Clingo, variables A and F refer to singular actions and fluents in this context. They are distinct from A and F, which refer to sets thereof.

Note that we assume that a nop action is contained in the specification.

2. Base Program

The *base program* specifies available actions, required actions and pairs of interfering actions. Mutual interference of actions is calculated by first collecting the fluents modified by its postcondition.

pc_changes(A, F) ← adds(A, F).
pc_changes(A, F) ← deletes(A, F).
pc_changes(A, F) ← invalidates(A, F).

Then, interference is calculated as described in Section C.

interfere(A1, A2) ⟵ demands(A1, F, _), pc_changes(A2, F).
interfere(A1, A2) ⟵ pc_changes(A1, F), pc_changes(A2, F).
interfere(A1, A2) ⟵ modifies(A1, F), modifies(A2, F).
interfere(A1, A2) ⟵ demands(A1, F, _), modifies(A2, F).
interfere(A1, A2) ⟵ pc_changes(A1, F), modifies(A2, F).
interfere(A1, A2) ⟵ interfere(A2, A1).

An action is available if it did not fail in a previous plan:

available(A) ⟵ action(A), not has_failed(A).

In case of *replanning*, actions which have already been executed successfully are not required in the new plan. Thus, they are excluded from the set of required actions. This is implemented through the external need_to_plan(A) atoms, whose truth values can be changed for replanning.

3. Transition Program

For a given slot $A_{t-1}1 \in P$ of a plan P, the *transition program* ensures that only a non-empty set of *executable*, *mutually non-interfering* actions is selected as A_t. It constitutes the ASP rules necessary to raise the planning horizon (and thus the maximal plan length) from $t-1$ to t and describes the new state S_t. The transition program is incrementally added.

1. At least one action is selected for execution in this slot, marked by apply(A, t). This may be the nop action, which is always executable.

   1 { apply(A, t) : available(A) }.

2. Each applied action is executable w.r.t. the previous state S_{t-1}.

   ⟵ apply(A, t), demands(A, F, true), not holds(F, t−1).
   ⟵ apply(A, t), demands(A, F, false), not −holds(F, t−1).
   ⟵ apply(A, t), not available(A).

3. Applied actions are pairwise non-interfering.

   ⟵ interfere(A1, A2), apply(A1, t), apply(A2, t), A1 < A2.

4. Describe the new state S_t based on the applied actions.

   add(F, t) ⟵ apply(A, t), adds(A, F).
   del(F, t) ⟵ apply(A, t), deletes(A, F).
   inv(F, t) ⟵ apply(A, t), invalidates(A, F).

   not holds(F, t) ⟵ inv(F, t).
   not −holds(F, t) ⟵ inv(F, t).

   holds(F, t) ⟵ add(F, t).
   holds(F, t) ⟵ holds(F, t−1), not del(F, t), not inv(F, t).
   − holds(F, t) ⟵ del(F, t).
   − holds(F, t) ⟵ −holds(F, t−1), not add(F, t), not inv(F, t).

5. Track which actions have been applied so far with was_applied(A, t)-atoms.

   was_applied(A, t) ⟵ apply(A, t).
   was_applied(A, t) ⟵ was_applied(A, t−1).

4. Goal Program

The *goal program* characterizes the desired plan for a specific horizon *h*. A straightforward goal is to have all actions (except nop) planned in some slot A_t with $t \leq h$. Through external goal_horizon(h) atoms, goals are constructed in a way that they can later be *deactivated*, i.e. have no effect on the set of models of the ground program.

← not was_applied(A, h), need_to_plan(A), goal_horizon(h).

While this is sufficient in a narrow sense, practically, additional rules such as optimization goals, optional actions, timeout bounds or other planning constraints may be added.

5. Externals

To use the same ground program for planning and replanning, some input must be supplied as external atoms:

1. *Initial Conditions*: The state of a fluent is expressed by the combination of holds/2 and -holds/2: holds(F, T) states that fluent *F* holds at slot *T*, -holds(F, T) states that fluent *F* does not holds at slot *T*, and if neither holds(F, T) not -holds(F, T) holds, the fluent *F* is unknown at slot *T*. The initial conditions describe which fluents hold *before* the first slot A_0 and are marked as external.

   #external holds(F, 0) : fluent(F).
   #external −holds(F, 0) : fluent(F).

2. *Required Actions*: When replanning, actions which have already been executed do not need to be re-run. There might be other practical reasons for not requiring some actions, such as further reducing the overall execution time for testing only selected features.

   #external need_to_plan(A) : action(A).

3. *Failed Actions*: Actions for which execution has failed have to be excluded from the set of available actions when replanning.

   #external has_failed(A) : action(A).

4. *Goals*: During solving, the search horizon is increased incrementally. As the ground program is reused for solving with different search horizons, a new goal is added. The old goal is deactivated by assigning goal_horizon(t) to false.

   #subprogram goal(t).
   #external goal_horizon(t).
   ← some_constraint(...), goal_horizon(t).

A. Solving Heuristics

As we do not know the required planning horizon, we guess a planning horizon before solving. Selecting a horizon which is too low results in the program being unsatisfiable. In contrast, an unnecessarily large horizon yields a large ground program, taking more time to preprocess and solve. Thus, the horizon should approach the plan length as quickly as possible, without overestimating it.

To balance these two aspects, we solve the planning problem in two phases: First, the search horizon is increased in exponential steps. In each step, additional *step*-subprograms are grounded. The old goal is deactivated and a new *goal*-subprogram for the new horizon h_n is added. This is repeated until the resulting ground program is *satisfiable*. Then, the minimal plan length is in the range $(h_{last}, h_n]$.

In the second phase, we proceed to find the lowest satisfiable horizon using binary search. Note that while the goal program is adjusted for every horizon, no additional *step*-subprograms are added, as they have already been ground in the first phase. This procedure is reminiscent to a parallel plan search algorithm proposed by Rintanen [10], however, we only solve one horizon at a time.

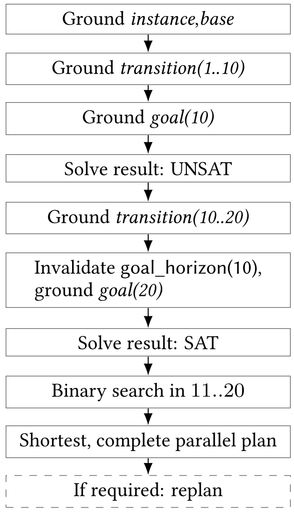


Fig. 2. Example of an incremental solving run.

An example of such an incremental solving run is presented in Fig. 2: After grounding the *instance* and *base* subprograms, the initial search horizon of 10 is increased until the instance is satisfiable at horizon 20. Then, the shortest plan is found through subsequent binary search.

A further improvement is to adapt the heuristics as follows:

#heuristic apply(A, t) : action(A). [−1, sign]

The addition of this line does not significantly reduce the runtime of Clingo, but reduces the number of executed actions in a plan.

V. Experimental Results

In the following, we report on experimental results based on a generated family of four benchmarks[2] that have been measured in a virtualized GNU/Linux environment (Debian stretch) with four 2.2GHz CPU cores (Intel Xeon Silver 4114) and 16GB of main memory.

[2] ASP instances of some generated benchmarks are provided on https://github.com/vroland/smoke-testing-benchmarks.

Version 5.4.1 of the Clingo solver was used in parallel mode 4, split with Python 3.7.3.

The benchmark families, as described in Fig. 3, have a certain structure which we believe is realistic and include corner cases:

Tree-Structured: Every action a_i depends on a fluent that is made true by $a_{\lfloor \frac{i}{2} \rfloor}$. The resulting plans are highly parallel, with plan lengths of $\lceil \log_2 N \rceil$ for N actions.

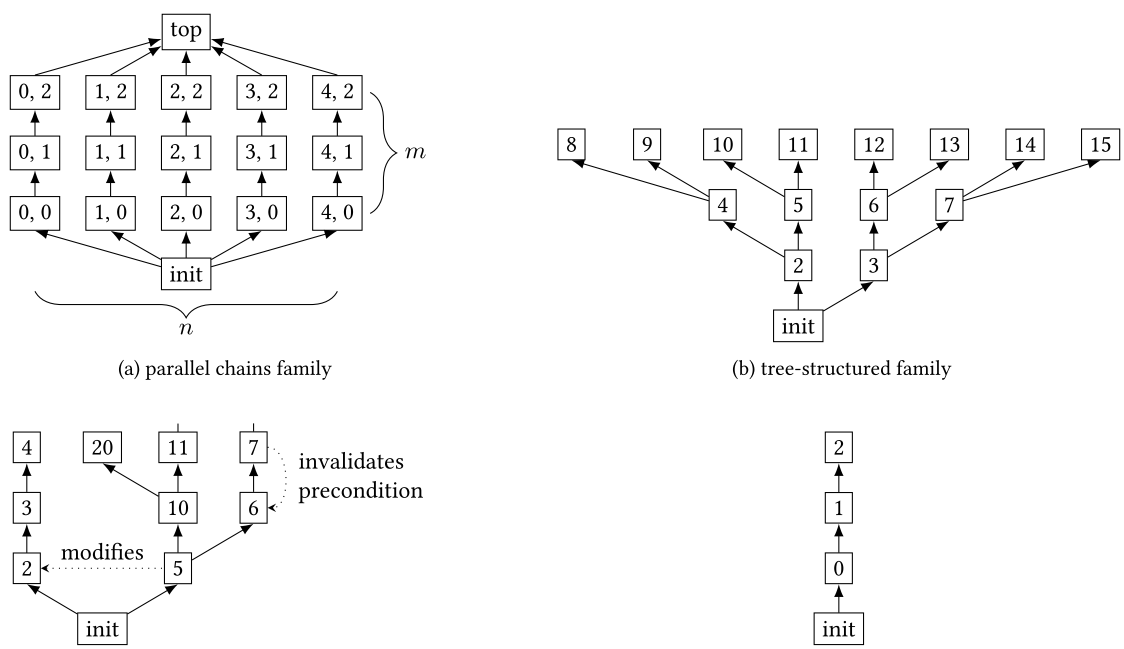
Sequential: Every action a_i depends on a fluent that is made true by a_{i-1}. The resulting plans are of length N for N actions.

Mixed: This benchmark aims to portray a more realistic test set, with tree-structured and sequential subsets of random length. Some of the branches of a tree-structured subset conflict based on their modifications. Every action a_i has a probability of $\frac{1}{5}$ producing a tree branch or a conflicting branch respectively, by depending on $a_{\lfloor \frac{i}{2} \rfloor}$ instead of a_{i-1}. With a probability of $\frac{3}{5}$, an action a_i depends on its direct predecessor a_{i-1}, producing a sequential sub-structure. In sequential subsets, actions invalidate their precondition. As the benchmark is randomized, repeated runs of this benchmark may not necessarily yield the same results.

Parallel Chains: Another realistic usage profile consists of running sequential test procedures in parallel. The benchmark is constructed by generating n *Sequential* instances of length m, which depend on a common init action and are jointly required by a final top action.

Fig. 4 shows that the size of the ground program w.r.t. number of rules and atoms increases with plan length. For the *Tree-Structured* and *Mixed* benchmarks, the search horizon stays low. Consequently, only a low number of instances of the transition subprogram have to be added to make the planning problem satisfiable. In contrast, the program size of the *Sequential* benchmark increases quickly. We observe a significant increase in the number of rules and atoms in the case that we increase the search horizon in exponential steps according to the solving heuristic (see Section A).

Fig. 5 presents the time spent in showing satisfiability and unsatisfiability and the total planning time for our benchmark families. Our upper bound on the horizon is 1000, consequently *Sequential* benchmarks are only conducted up to this size. Starting at



(a) parallel chains family

(b) tree-structured family

(c) mixed family

(d) sequential family

Fig. 3. Dependency structure of the considered benchmark families.

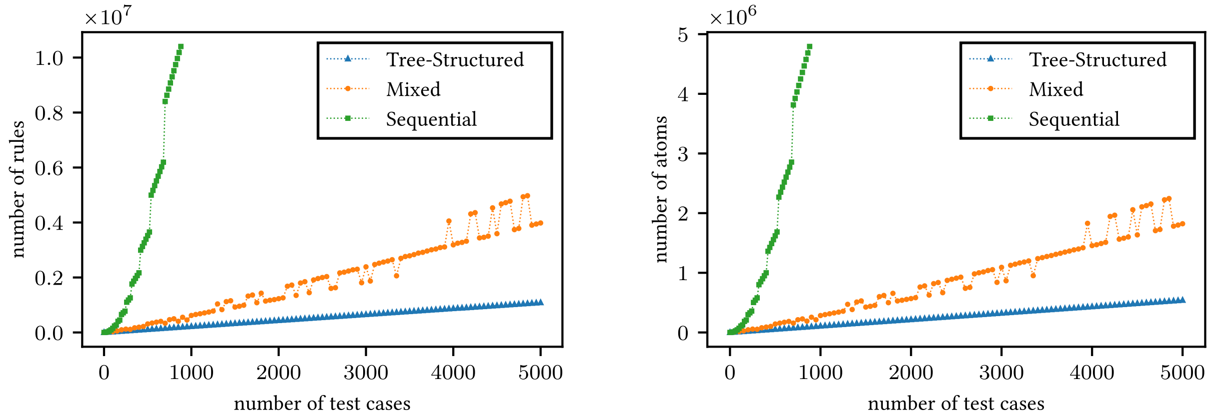


Fig. 4. Ground program size w.r.t. the number of actions.

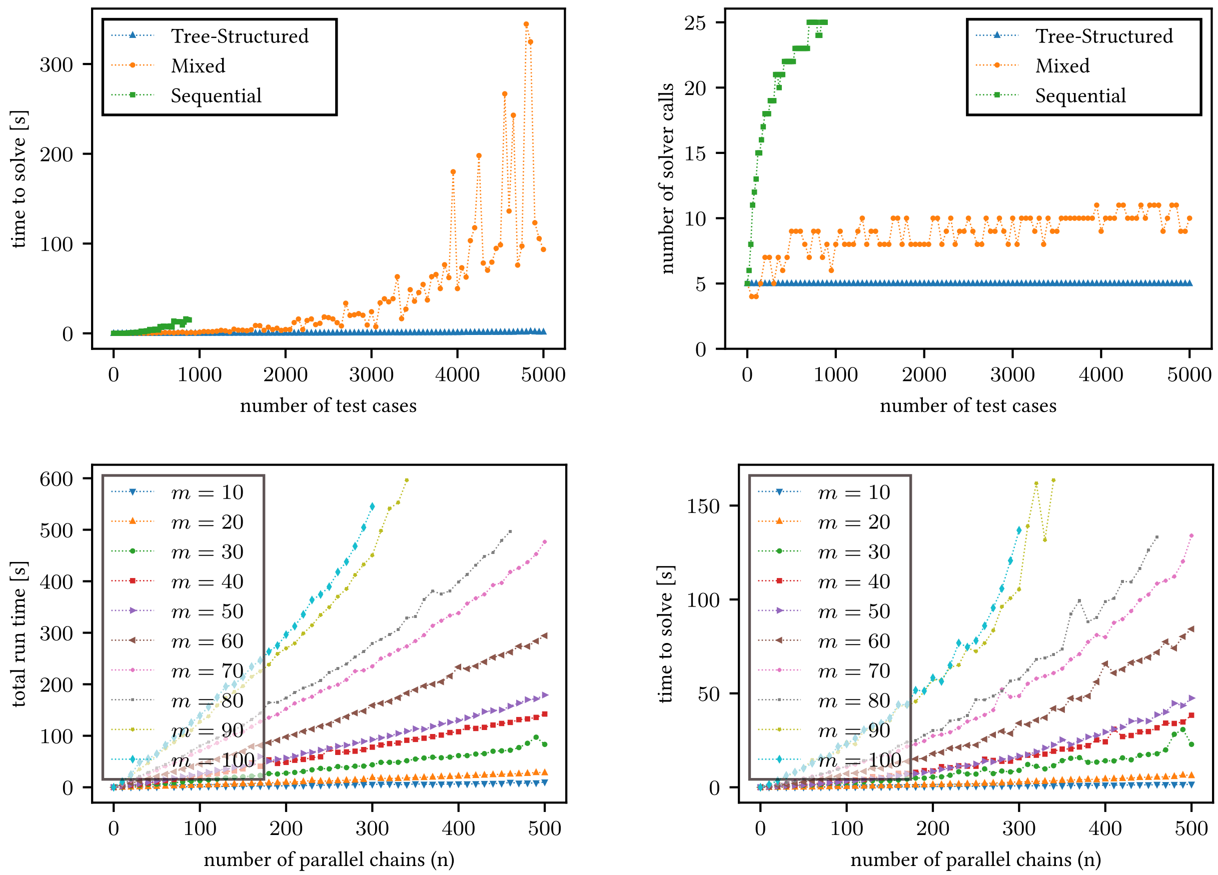


Fig. 5. Total run time and time to satisfiability for all benchmark families.

approximately 2000 actions, *Mixed* instances take significantly longer to solve than *Tree-Structured* instances. In contrast to the highly parallel *Tree-Structured* benchmark with short plan lengths, we observe an increase in the number of interactive solver calls with *Sequential* and *Mixed*, as the number of actions increases. This is a result of our search heuristic (see Section A), which incrementally approaches the larger search horizon needed for longer plans.

In case of the *Parallel Chains* benchmark, runtimes increase with both higher chain length m and more parallel chains n. However, as indicated by the previously discussed benchmark families, our solution is best suited for highly parallel instances: For the same total number of actions, instances with high n and small m require significantly less time to solve than longer, less parallel plans with high m.

Considering differences in run time for a constant n and increasing m, we notice a non-uniform distribution. This is a result of the incremental search heuristic: For instance, the difference in runtime of $m = 90$ to $m = 100$ is smaller than the difference of $m = 90$ to $m = 80$. This irregularity occurs because for increasing the horizon from 80 to 90, an additional grounding and solving step is necessary, whereas 90 and 100 fall in the same horizon step.

Presumably through memory limitations of our benchmark hardware resulting in swapping, jitter is introduced for larger problem sizes.

By comparing total runtime with time spent for solving, we can observe a difference in almost an order of magnitude. This suggests that a large portion of the total run time is spent in preprocessing and grounding.

Due to the reuse of a single ground program in the entire test execution, we avoid repeated preprocessing and keep the internal solver state. This speeds up subsequent replanning as shown in Fig. 6. The *Sequential* benchmark exhibits a steep growth of ground program size with the number of actions, as the latter is equal to the minimal search horizon. In the first phase of the horizon search, solver calls spend large parts of the execution time in preprocessing and grounding rather than solving. The peak at call 22 marks the finishing goal of the initial plan search. From call 23, the search for a new plan starts after an initial action has failed. Though solving now takes longer due to the overall larger ground program size, the total planning time is reduced, as less preprocessing and grounding time is needed.

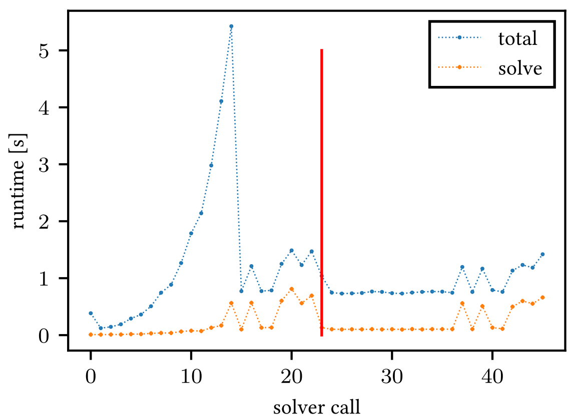


Fig. 6. Time spent per solver call for planning and replanning of a *sequential* instance with 500 actions.

VI. Conclusion

Smoke testing is an important method to increase the reliability of hardware-depending systems. Due to concurrent access to the same physical resource and the infeasibility of the use of virtualization, smoke testing requires some form of planning. In this paper, we propose to decompose test cases in terms of atomic actions consisting of preconditions and effects and developed a declarative framework that allows to automatically generate complete and incomplete plans based on ASP. By modeling tests as state transitions of a system-under-test, our method is capable of generating shortest, parallel test plans, while offering the flexibility to incorporate additional goals, constraints and knowledge [13].

Broader use cases can be modeled as variants of smoke testing, with some adjustments to the planning goals and constraints. For instance, *regression testing* can be seen as scaled-up smoke testing, which we have shown is viable for thousands of tests or more, especially if planning time is not critical.

Furthermore, functionality can easily be tested piecewise in isolation to inspect reasons for test failures. By only requiring the execution of a specific action as our goal and minimize the total number of planned actions, we obtain a minimal trace of actions to produce a particular SUT behavior.

Moreover, through the use of parallel plans, we can find issues which may not occur in sequential or isolated test runs. Such issues, typically referred to as *race conditions* or *race hazards*, arise from conflicting concurrent use of resources and lead to non-deterministic test outcomes. When such dependencies of two tests on the same resource are insufficiently specified but some property of the execution environment is implicitly assumed, we call this a *hidden assumption*. Finding *race conditions* and *hidden assumptions* can be time-consuming, but can be automated to some degree by our solution: Running a large number of structurally diverse plans [14] can provide insight on sequences of actions which cause the system to fail, by recording failing plans and looking for similarities.

Experiments show that generating *short* and *highly parallel* plans can be efficiently done using ASP. We believe that domains like networking tests or tests of distributed systems allow for such plans.

Inherently sequential test procedures prove adverse to our solution and may be better addressed by existing planning tools. These occur in domains where all actions operate on a common resource, forcing mutual exclusion, like testing workflows in a graphical user interface.

The AI planning problem is used in different contexts in the area of testing: [15] uses PDDL to find well-known security issues in web applications, [16] considers test case generation, [17] uses contingent planning in the area of penetration testing, [18] proposes to test chatbots using planning, [19] considers test case generation for systems, and [20] considers hierarchical GUI test case generation using planning methods. In the domain of software testing, to the best of our knowledge, test execution planning has not been modeled in the sense of an AI planning problem. However, work has been done on obtaining *test specification*, like Behavior-Driven Development (BDD) [21], [22] or Model-Driven Testing. At a first glance, actions may seem reminiscent of *given-when-then-style* scenarios in Behavior-Driven Development (BDD) [21], [22]. However, while BDD is concerned with how tests are specified in natural language, this work focuses on fast test execution. Moreover, while BDD scenarios may cover multiple layers of abstraction, actions should remain relatively low-level. Conceptually, a BDD test is more akin to an abstract view of a *subset of a plan* than to an action. However, synergies could emerge when using the BDD process in conjunction with planning-based test execution.

Solving planning problems based on the propositional satisfiability solvers instead of specialized planners has been explored since the 1990s [10], [23], exploiting the flexibility of general-purpose SAT solvers. Similarly, Answer Set Programming has successfully been used to implement classical [24] as well as real-world planning such as tasks robotics [25], [26]. Accord-ing to [27], ASP-based planners perform especially well for short plans with complex dependencies.

We compare our approach to modeling uncertainty as in conformant planning [2], [11] in Section C, and also point out the related idea of oversubscription planning [3] when it comes to achieving as many actions as possible. Though, in contrast to oversubscription approaches as e.g. in [3], we do not impose cost-estimates on actions such that a global constraint on the total cost is satisfied (e.g. for resources such as time, or power consumption), but merely achieve the maximum number of actions. However, such an extension could be achieved by extending actions with temporal information, such as an estimated duration, where time then can be encoded as a limited resource. For instance, temporal constraints such as a maximal run time or required parallelism could be added. Going further, the notion of a shortest plan could now be redefined as a temporal (cf. [28] for an overview). However, how this affects planning performance and which temporal extensions prove to be of use in practice, remains an open question.

In the future, we plan to investigate the test specification debugging problem, i.e. the questions, why a test specification does not admit a complete test plan and the possibilities to generate substantially different test plans.

Acknowledgment

We are grateful for the valuable feedback and additional references from the anonymous reviewers, which helped greatly to improve this work.

References

[1] G. D. Everett, R. McLeod Jr, *Software testing: testing across the entire software development life cycle*. Piscataway, NJ, Hoboken, N.J: John Wiley & Sons, 2007.

[2] H. Palacios, H. Geffner, "Compiling uncertainty away in conformant planning problems with bounded width," *Journal of Artificial Intelligence Research*, vol. 35, pp. 623–675, 2009, doi: 10.1613/jair.2708.

[3] D. E. Smith, "Choosing Objectives in Over-Subscription Planning," in *Proceedings of the Fourteenth International Conference on Automated Planning and Scheduling (ICAPS 2004)*, June 3-7 2004, Whistler, British Columbia, Canada, 2004, pp. 393–401, AAAI.

[4] G. Brewka, T. Eiter, M. Truszczynski, "Answer set program-ming at a glance," *Communications of the ACM*, vol. 54, no. 12, pp. 92–103, 2011, doi: 10.1145/2043174.2043195.

[5] I. Niemelä, "Logic programs with stable model semantics as a constraint programming paradigm," *Annals of Mathematics and Artificial Intelligence*, vol. 25, no. 3-4, pp. 241–273, 1999, doi: 10.1023/A:1018930122475.

[6] M. Gelfond, V. Lifschitz, "Classical negation in logic programs and disjunctive databases," *New Generation Computing*, vol. 9, no. 3/4, pp. 365–386, 1991, doi: 10.1007/BF03037169.

[7] T. Janhunen, I. Niemelä, "The answer set programming paradigm," *AI Magazine*, vol. 37, no. 3, pp. 13–24, 2016.

[8] M. Gebser, R. Kaminski, B. Kaufmann, T. Schaub, "Answer Set Solving in Practice," *Synthesis Lectures on Artificial Intelligence and Machine Learning*, vol. 6, pp. 1–238, 2012, doi: 10.2200/S00457ED1V01Y201211AIM019.

[9] M. Gebser, R. Kaminski, B. Kaufmann, T. Schaub, "Clingo = ASP + control: Preliminary report," *CoRR*, vol. abs/1405.3694, 2014.

[10] J. Rintanen, K. Heljanko, I. Niemelä, "Planning as satisfia-bility: parallel plans and algorithms for plan search," *Artificial Intelligence*, vol. 170, no. 12-13, pp. 1031–1080, 2006, doi: 10.1016/j.artint.2006.08.002.

[11] A. Cimatti, M. Roveri, P. Bertoli, "Conformant planning via symbolic model checking and heuristic search," *Artificial Intelligence*, vol. 159, no. 1-2, pp. 127–206, 2004, doi: 10.1016/j.artint.2004.05.003.

[12] M. Gebser, R. Kaminski, B. Kaufmann, M. Ostrowski, T. Schaub, S. Thiele, "Engineering an incremental ASP solver," in *Logic Programming, 24th International Conference, ICLP 2008*, Udine, Italy, December 9-13 2008, Proceedings, vol. 5366 of Lecture Notes in Computer Science, 2008, pp. 190–205, Springer.

[13] T. C. Son, C. Baral, T. H. Nam, S. A. McIlraith, "Domain-dependent knowledge in answer set planning," *ACM Transactions on Computational Logic*, vol. 7, no. 4, pp. 613–657, 2006, doi: 10.1145/1183278.1183279.

[14] T. Eiter, E. Erdem, H. Erdogan, M. Fink, "Finding similar/-diverse solutions in answer set programming," *Theory and Practice of Logic Programming*, vol. 13, no. 3, pp. 303–359, 2013, doi: 10.1017/S1471068411000548.

[15] F. Wotawa, J. Bozic, "Plan it! automated security testing based on planning," in *Proceedings of Testing Software and Systems - International Conference, ICTSS 2014, Madrid, Spain, September 23-25, 2014., vol. 8763 of Lecture Notes in Computer Science*, 2014, pp. 48–62, Springer.

[16] A. E. Howe, A. von Mayrhauser, R. T. Mraz, "Test case generation as an AI planning problem," *Automated Software Engineering*, vol. 4, no. 1, pp. 77–106, 1997, doi: 10.1023/A:1008607721339.

[17] D. Shmaryahu, G. Shani, J. Hoffmann, M. Steinmetz, "Simulated penetration testing as contingent planning," in *Proceedings of the Twenty-Eighth International Conference on Automated Planning and Scheduling, ICAPS 2018, Delft, The Netherlands, June 24-29, 2018*, 2018, pp. 241–249, AAAI Press.

[18] J. Bozic, O. A. Tazl, F. Wotawa, "Chatbot testing using AI planning," in *Proceedings of International Conference On Artificial Intelligence Testing IEEE, AITest 2019, Newark, CA, USA, April 4-9*, 2019, 2019, pp. 37–44, IEEE.

[19] R. T. Mraz, A. E. Howe, A. von Mayrhauser, L. Li, "System testing with an AI planner," in *Proceedings of Sixth International Symposium on Software Reliability Engineering, ISSRE 1995, Toulouse, France, October 24-27, 1995*, 1995, pp. 96–105, IEEE Computer Society.

[20] A. M. Memon, M. E. Pollack, M. L. Soffa, "Hierarchical GUI test case generation using automated planning," *IEEE Trans-actions of Software Engineering*, vol. 27, no. 2, pp. 144–155, 2001, doi: 10.1109/32.908959.

[21] M. Diepenbeck, U. Kühne, M. Soeken, R. Drechsler, "Be-haviour driven development for tests and verification," in *Proceedings of Tests and Proofs - 8th International Conference, TAP@STAF 2014, York, UK, July 24-25, 2014, vol. 8570 of Lecture Notes in Computer Science*, 2014, pp. 61–77, Springer.

[22] SmartBear Software, "Cucumber," 2019. [Online]. Available: https://cucumber.io/, Accessed: 2020-01-11.

[23] H. A. Kautz, B. Selman, "Planning as satisfiability," in *10th European Conference on Artificial Intelligence, ECAI 92, Vienna, Austria, August 3-7, 1992. Proceedings*, 1992, pp. 359–363, John Wiley and Sons.

[24] M. Gebser, R. Kaufmann, T. Schaub, "Gearing up for effective ASP planning," in *Correct Reasoning - Essays on Logic-Based AI in Honour of Vladimir Lifschitz*, vol. 7265 of Lecture Notes in Computer Science, 2012, pp. 296–310, Springer.

[25] S. Zhang, M. Sridharan, F. S. Bao, "ASP+POMDP: integrating non-monotonic logic programming and probabilistic plan-ning on robots," in *Proceedings of IEEE International Confer-ence on Development and Learning and Epigenetic Robotics, ICDL-EPIROB 2012, San Diego, CA, USA, November 7-9, 2012*, 2012, pp. 1–7, IEEE.

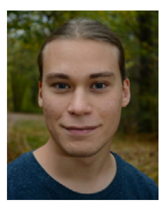
[26] J. J. Portillo, C. L. García-Mata, P. R. Márquez-Gutiérrez, R. B a r a y - Arana, "Robot platform motion planning using answer set programming," in *Proceedings of the Seventh Latin American Workshop on Non-Monotonic Reasoning, LANMR 2011, Toluca, Estado de México, México, November 7-8, 2011, vol. 804 of CEUR Workshop Proceedings*, 2011, pp. 35–44, CEUR-WS.org.

[27] Y. Jiang, S. Zhang, P. Khandelwal, P. Stone, "Task planning in robotics: an empirical comparison of PDDL- and asp-based systems," *Frontiers of Information Technology & Electronic Engineering*, vol. 20, no. 3, pp. 363–373, 2019, doi: 10.1631/FITEE.1800514.

[28] L. Vila, "A survey on temporal reasoning in artificial intelligence," *AI Communications*, vol. 7, no. 1, pp. 4–28, 1994, doi: 10.3233/AIC-1994-7102.

Tobias Philipp

Tobias Philipp is a full-time senior consul-tant and verification expert in the Development and Verification Team, Defence & Space Division, secunet Security Networks AG, Germany. His research interests include formal verification, logic-based Artificial Intelligence, planning, constraints, and knowledge representation and reason ing. Until 2017, he was a scientifc staff member of the International Center of Computational Logic (ICCL) at Technische Universität Dresden. In 2013, he graduated of the International Master's Program in Computational Logic, Technische Universität Dresden, Germany.



Valentin Roland

alentin Roland is a final year Diploma student at the Faculty of Computer Science, Technische Universität Dresden, Germany. Since 2018, he works as a student assistant in the Development and Verification Team, Defence & Space Division, secunet Security Networks AG. He is interested in formal methods and verification, knowledge representation and reasoning and declarative problem solving. This includes academic research as well as industrial applications thereof.

Lukas Schweizer

Lukas Schweizer is a research assistant in the Computational Logic Group at Technische Universität Dresden, Germany. He is also working as a research engineer at Deepreason.ai, a spin-off from the University of Oxford, developing the next generation of Datalog engines and knowledge graph systems. His interests are knowledge representation and reasoning in general, in particular rule based approaches and answer-set programming.

International Journal of Interactive Multimedia and Artificial Intelligence, Vol. 6, Nº5

An Application of Declarative Languages in Distributed Architectures: ASP and DALI Microservices

Stefania Costantini, Giovanni De Gasperis, Lorenzo De Lauretis *

Dipartimento di Ingegneria e Scienze dell'Informazione e Matematica Università degli Studi di L'Aquila, L'Aquila (Italy)

Received 15 November 2020 | Accepted 13 December 2020 | Published 10 February 2021

Abstract

In this paper we introduce an approach to the possible adoption of Answer Set Programming (ASP) for the definition of microservices, which are a successful abstraction for designing distributed applications as suites of independently deployable interacting components. Such ASP-based components might be employed in distributed architectures related to Cloud Computing or to the Internet of Things (IoT), where the ASP microservices might be usefully coordinated with intelligent logic-based agents. We develop a case study where we consider ASP microservices in synergy with agents defined in DALI, a well-known logic-based agent-oriented programming language developed by our research group.

Keywords

AI For Ubiquitous Computing, Answer Set Programming, Intelligent Software Agents, Knowledge Based Systems, Microservices.

DOI: 10.9781/ijimai.2021.02.001

I. Introduction

The remarkable success of Answer Set Programming (ASP) in a wide variety of applications calls for the definition of specific software engineering principles. ASP is a successfully logic programming paradigm (cf. [1] and the references therein) stemming from the Answer Set (or "Stable Model") semantics of Gelfond and Lifschitz [2], [3], and based on the programming methodology proposed by Marek, Truszczyński and Lifschitz [4], [5]. ASP is put into practice by means of effective inference engines, called solvers[1]. ASP has been widely applied in many fields, e.g., to information integration, constraint satisfaction, routing, planning, diagnosis, configuration, computer-aided verification, biology/biomedicine, knowledge management, and many others.

In this paper we discuss the possibility of exploiting ASP to define components for distributed systems, to be deployed over different nodes of a network. In this perspective, the connections between components and the ways of exchanging information should be clearly specified. Our approach is inspired by the microser-vices architectural abstraction, which can be described as a particular way of designing distributed software applications as suites of independently deployable in-teracting services (cf. for instance the survey [6], and https://martinfowler.com/articles/microservices.html#CharacteristicsOfAMicroserviceArchitecture).

[1] Many performant ASP solvers are available as open-source tools, a list is reported at https://en.wikipedia.org/wiki/Answer_ set_programming.

* Corresponding author.

E-mail addresses: stefania.costantini@univaq.it (Stefania Costantini), giovanni.degasperis@univaq.it (Giovanni De Gasperis), lorenzo.delauretis@graduate.univaq.it (Lorenzo De Lauretis).

A microservice is indeed a component, as it is a unit of software that is independently replaceable and modifiable: in fact, it intended as a self-contained piece of business functionality with clear interfaces that can be accessed by the "external world". This kind of architectural abstraction enables distribution, as each microservice is meant to be executed as an independent process, and heterogeneity, as it allows different services to be written in different programming languages. Microservices are a suitable architectural abstraction for the Internet of Things (IoT): a microservice may incapsulate a physical object, where service inputs and/or outputs can possibly be linked to sensors/actuators. Microservices are by their very nature heterogeneous, so open issues are: how microservices communicate with each other (synchronous, asynchronous, which is the message format, etc.); and, the protocols used for the communication.

Microservices in real distributed software architectures and in cloud computing are usually deployed via lightweight containers. In standard terminology borrowed from software engineering, a container is a standard unit of software that packages up code and all its dependencies; so, the application runs quickly and reliably and can be seamlessly transferred from one computing environment to another. A widely used tool to create containers is Docker, available in the form of an open source Docker Engine[2]. A Docker container image consists in a lightweight, standalone, executable package of software that includes all elements needed to run an application: code, run-time support, system tools, system libraries and settings.

Along this line, we propose ASP microservices that might be blended into heterogeneous systems, and even into Multi-Agent-System (MAS) since each such component may be seen as a reactive agent. They could in perspective be employed in cloud computing and in IoT, including robotic applications. In this paper we discuss how these components, that we call μASP-Services (μASPSv's), can be

[2] See https://www.docker.com

specified, how their interfaces to the "external world" can be defined, and how they should procedurally behave. In fact, a μASPSv is meant to be based upon a 'core' ASP program whose activities, however, should be triggered by external stimula/requests coming from some source, and whose results should be returned to the requesters.

In our view, the 'core' ASP program should be included into a container, that can be possibly realized via the Docker technology, which should also include: an interface, to provide the Âµ ASPSv with inputs, and to select and deliver the outputs; solving capabilities to compute the answer sets. So, a docker deployment for a Âµ ASPSv should include the so For ASP, standalone versions of the most important solvers are nowadays available. New solutions have been recently introduced [7], that allow for incremental solving of an ASP program under atoms/rules addition/deletion, and so might be used to provide a μASPSv with new inputs and cancel old ones. Thus, a docker deployment for a μASPSv should include the source program, its 'execution shell', and the solver.

A small specimen of the proposed approach is represented in the following example, which is meant to be (a fragment of) the code of a controller component/agent, acting in the IoT. This piece of code might be in fact the ASP 'core' of a μASPSv. *test_ok* is the input coming from a sensor, with value 'true' if the controlled device is working properly, (otherwise the value is set to false).

device_ok ⟵ t*est_ok*.

device_fault ⟵ not *test_ok*.

wait ⟵ not *wait*, not *sensor_input*.

In this simple example, inconsistency (due to the odd cycle over *wait*) is to be interpreted as a 'no-operation' controller state, where the component is waiting for sensor's outcome. It can be assumed that the sensor provides results at a certain frequency. The outcome, i.e., *device_ok* or *device_fault*, is to be delivered to whatever other components would ask for it.

In order to work in a standalone way within a distributed system, an interface (or 'shell') will manage the 'core' ASP program, and in particular will perform the following functions. First, manage the inputs and outputs of the μASPSv: i.e., be able to detect input arrival and to dispatch the outputs according to the request coming from the μASPSv's external environment. In the above example, inputs can be: (1) queries over the device state for which an answer has to be delivered, and (2) sensor outcomes, which are to be considered as particular inputs which activate the module. In the general case, upon the arrival of inputs, the shell will: (i) add the inputs to the ASP program as facts; (ii) evaluate the answer sets of the resulting ASP program; (iii) according to previously-received requests, extract (from the answer sets) the answers and and deliver them to the external environment. Notice that the shell, after delivering the outputs, will remove (some or all of) the last-added program facts so as to bring back the controller to the 'no-operation' state. In a 'stateless component', all inputs will be removed, while some of the inputs can be left if instead the component is meant to have a state; the shell functioning is enabled (or at least greatly simplified) by the new advanced solving capabilities provided in particular by the clingo ASP solver [7].

There are however complex devices in the Internet of Things that should be managed in a coordinated way. Take for instance a car, where modern cars include several control devices for the various parts. Each such device will be managed by a microservice, where such microservices should produce coordinated behavior. It is thus a reasonable choice to define these components as agents. In this way, the overall control over the complex device will be managed by a Multi-Agent-System (MAS), which is by definition capable of integrated behavior. Several approaches to logic-based agent-oriented languages exist (cf., e.g., [8]–[10]). We may notice that such kind of agents can be the natural complement to ASP microservices. In general terms, one might want to adopt ASP microservices whenever there is the need to cope with uncertainty, or the need to manage possible alternative scenarios. When instead immediate direct reactive/proactive behavior is required, logical agent may represent a suitable tool. Among the different existing logical-based agent frameworks, to develop our case study we choose the DALI logic-based agent-oriented language and framework (introduced in Section V), which has been developed by our research group, and that (as illustrated later) we have already used in synergy with ASP modules in past work.

In this paper we introduce a formal definition of ASP microservices and we outline a possible logic-based semantics of an overall heterogeneous distributed system encompassing such modules, and other logical components/agents. The paper is structured as follows. In Section II we introduce basic concepts about microservices. In Section III we recall (for the sake of completeness) the Answer Set Programming paradigm, and in Section IV we briefly survey and discuss existing approaches to modularity in ASP. We introduce our contribution in Sections VII and IX, i.e.: (1) how to define and implement μASPSv's so as to be able to get inputs and extract answers, and how the inner ASP program might be structured; (2) how to provide a formal semantics to a generic microservice architecture possibly encompassing μASPSv's. In Section VIII we discuss a small case study, developing a specific μASPSv which implements an intelligent agent managing a road intersection (i.e., a "virtual traffic light"), where cars are modeled as DALI logical agents. Finally, in Section X we conclude. This paper is in our view an evolution of the work in [11], in the sense that there, as illustrated in Section VI, ASP modules were invoked as auxiliary modules by agents in a DALI multi-agent system. Here, we make it possible for an ASP program to act as an independent component, that is able to interact with other components, among which agents.

II. Background: Microservices

In order to better understand Microservices, let us first introduce the concept of "Service". A Service, as a software component, is a mechanism to enable access to one or more software capabilities [12]. It provides other applications with stable, reusable software functionalities at an application-oriented, business-related level of granularity using certain standards [13]. Service-Oriented Architecture (SOA) is a software architectural style that uses services as the main building component [12]. Key features of SOA are heterogeneity, standardization and "evolvability" of services.

Microservices can be seen as a technique for developing software applications that, inheriting all the principles and concepts from the SOA style, permits to structure a service-based application as a collection of very small loosely coupled software services [14].

A MicroServices Architecture (MSA) is an evolution of the SOA architecture, making the communication lighter and the software parts (Microservices) smaller. As empasized in [15], it can be seen as a new paradigm for programming applications by means of the composition of small services, each one running its own processes and communicating via light-weight mechanisms. Key features of MSA are bounded scope, flexibility and modularity [15]. I.e., there is a clear definition of the data a microservice service is responsible for and is "bound to". So, a microservice owns this data and is responsible for its integrity and mutability.

The work in [16] shows that a distributed MSA can easily fit into an IoT system. In particular, the set of microservices can be seen as a Multi-Agent-System, cooperating to realize all system functionalities.

At the current day, microservices are still a new and emerging paradigm, having building standards not perfectly defined and communication protocols that are not well specified: in fact, following one of the definitions of microservices [14], [15], they are small loosely coupled software services that communicate, possibly exploiting service discovery to find the route of communication between any two of them. In our work, we are proposing a new approach, that is μ ASPSv's, which are based upon an inner ASP program.

III. Background: Answer Set Semantics (AS) and Answer Set Programming (ASP)

The following introduction consists of standard material taken (literally for what concerns long-established scientific terminology and definitions) from [1], [17]–[19]. "Answer Set Programming" (ASP) (cf. [1] and the references therein) is a successful programming paradigm based on the Answer Set Semantics. In ASP, one can see an answer set program (for short, just "program") as a set of statements that specify a problem, where each answer set represents a solution compatible with this specification. Whenever a program has no answer sets (no solution could be found), it is said to be inconsistent, otherwise it is said to be consistent.

Syntactically, an ASP program Π is a collection of rules of the form

$H \leftarrow A_1, ..., A_m, not\ A_{m+1}, ... , not\ A_{m+n}.$

where H is an atom, $m, n \geq 0$, and each A_i, $i \leq m+n$, is an atom. Atoms and their negations are called literals. Symbol ⟵ is often indicated as :- in practical programming. The left-hand side and the right-hand side of the clause are called head and body, respectively. A rule with empty body is called a fact. A rule with empty head is a constraint, where a constraint of the form '$\leftarrow L_1,..., L_n$.' states that literals $L_1,..., L_n$ cannot be simultaneously true in any answer set. Constraints are often rephrased as '$f \leftarrow not\ f, L_1,..., L_n$.' where f is a fresh atom. To avoid the contradiction over f, some of the Li's must be false thus forcing f to be false, and this, if achieved, fulfills the constraint.

Actually, an ASP rule can have a more general form including a disjunction of literals in the head, and "classical negation" of atoms [3]; various useful programming constructs have been introduced over time; for simplicity, we consider the basic form, i.e., "normal logic programs". The interested reader can refer, e.g., to [20] for a complete up-to-date discussion about ASP syntax and practical use.

The answer set (or "stable model") semantics (AS) [2] can be defined in several ways (cf., e.g., [21], though more recently several other definitions have appeared in the literature). However, answer sets of a program Π are found among the supported minimal classical models of the program (interpreted as a first-order theory in the obvious the model (directly or indirectly) by its own negation. This is why it can be the case that no answer set exists: take, e.g, simple ASP program $p \leftarrow not\ p$ which is equivalent to first-order theory $p \vee p$ with unique minimal model $\{p\}$ which is not an answer set as p is supported, in the model, by its own negation. As it is well-known, AS extends the three-valued Well-Founded semantics [22] for normal logic programs, where every program Π has a well-founded model $wfm(\Pi) = \langle T, F \rangle$ where T is the set of true atoms, F is the set of false atoms, and the remaining atoms (implicitly) form the set $U = Undef(\Pi)$ of the undefined atoms. For every answer set M it holds that $T \subseteq M$, so finding the answer sets accounts to suitably assigning truth values to the undefined atoms.

The ASP approach to problem-solving consists basically in the following: (i) encoding of the given problem via an ASP program; (ii) computing the "answer sets" of the ground program via an ASP solver (a list of available solvers can be found at https://en.wikipedia.org/wiki/ Answer_set_programming), where, as a preliminary step, solvers perform the "grounding" of the program, by substituting all variables with the constants occurring in the program; (iii) extracting the problem solutions by examining such answer sets; in fact, answer set contents can be in general reformulated in order to present the solution in terms of the given problem.

A top-down query answering device which is prolog-style, i.e., does not compute answer sets in advance to extract the query answers, has been defined in [23] for RAS, where RAS is a variation of AS where every program admits answer sets[3]. RAS and AS coincide however over a wide class of programs: some sufficient conditions that identify classes of programs where the two semantics coincide are reported in [26]. Queries that have been introduced are, first of all, "? A" asking whether A is true w.r.t. some answer set of given program Π. Other queries are the following: query "? not A" asks whether A is false w.r.t. some answer set of Π, and therefore it succeeds if not A is true in some of them (this implements the operator not introduced in [27]); query "? not not A" asks whether not A is false in some answer set, and therefore it succeeds if A is true in some of them, which corresponds to query "? MA", M standing for 'possibility' in the modal logic sense; query "? not not A" asks whether it is not true that A is false w.r.t. some answer set of Π, i.e., that A is true in all of them, which corresponds to "? KA", K standing for 'knowledge' in the modal logic sense; query "? not notnot A" asks whether A is false in every answer set, meaning Knot A, i.e., not MA (a new operator NOT is a shorthand for not notnot A).

IV. Background: Modularity in ASP

Existing approaches to modularization of ASP programs have been extensively reviewed in [18], to which the reader may refer for a complete account. Reporting faithfully from there, such approaches can be divided into two lines: "programming-in-the-large", where programs are understood as combinations of separate and independent components, combined by means of compositional operators; "programming-in-the-small", in which logic programming is enriched with new logical connectives for managing subprograms.

Considering the programming-in-the-small vision: in [28], program modules are viewed as generalized quantifiers; [29] proposes templates for defining subprograms; [30] developed a declarative language for modular ASP, which allows a programmer to describe a state how one ASP module can import processed answer sets from another ASP module. The work in [31] explores how to divide an ASP program into components according to its structure in terms of cycles.

Lifschitz and Turner's "splitting set theorem" (cf. [32]), or variants of it, is underlying many programming-in-the- large approaches. The basic idea is that a program can be divided into two parts: a "bottom" part and a "top" part, such that the former does not refer to predicates defined in the latter. Computation of the answer sets of a program can be simplified when the program is split into such parts.

[33] defines the notion of a "DLP-function" which is basically a module for which a well-defined input/output interface is provided; a suitable compositional semantics for modules is introduced. [34] provides a simple and intuitive notion of a logic programming module that interacts through an input/output interface. This is achieved by accommodating modules as proposed by [35] to the context of Answer Set Programming. Full compatibility of the module system with the stable model semantics is achieved by allowing positive recursion to occur inside modules only.

[36] focuses on modular non-monotonic logic programs (MLP) under the answer set semantics, where modules may provide input to other modules. Mutually recursive module calls are allowed.

[37] defines modules in terms of macros that can be called from a program. [38] provides modules specification with information hiding, where modules exchange information with a global state.

[3] To the best of our knowledge, the only alternative query-answering device for ASP that does not compute the answer sets in advance has been introduced in [24], [25], though under some syntactic-semantic limitations.

In [39] a technique is proposed to allow an answer set program to access the brave or cautious consequences of another answer set program. [40] proposes "modular logic programs" as a modular version of ASP. This work consider programs as modules and define modular programs as sets of modules. The authors introduce "input answer sets", which is the key semantic object for communication between modules.

[41] proposes to adopt ASP modules in order to simulate (within reasonable complexity) possibility and necessity operators. Such operators (given the underlying modules) are meant to be usable in ASP programs, but possibly also programs written under other programming paradigms.

It can be seen that none of the above approach tackles modularization in view of using ASP modules as standalone components in distributed systems. Therefore, our approach is a novelty in the landscape of the current literature.

V. Background: Logical Agents and DALI

The material exposed in this section, which reports about our previous work concerning logical agents so as to provide the notions needed in the subsequent sections, is largely taken (in some parts literally, to be faithful to well- established terminology) from [42]–[52] and from the DALI web site https://github.com/AAAI-DISIM-UnivAQ/DALI.

The original perspective on agents in Artificial Intelligence was focused on the agents' reasoning process, thus identifying "intelligence" as rationality, thus neglecting the interactions of the agents with the environment and with other agents. This perspective has been heavily criticized for instance in [53], [54], that adopts in an extreme way the opposite point of view, arguing that "intelligent" behavior results solely from the ability of an agent to react appropriately to changes in its environment.

Reasoning about beliefs, but also about what an agent means and chooses to do, is the basis of the seminal approach of the BDI (Belief, Desires, Intention) logic for modelling agents by [55], that resulted in the definition of the AgentSpeak agent-oriented logic programming language [56]. At the same time, in the approach of [57], agents were theories (logic programs), each one with its name, and they were able to communicate with each other via two communication primitives (tell/told). A view of logical agents, able to be both rational and reactive, i.e., capable not only to reason and to communicate, but also to provide timely response to external events, has been introduced in [58], [59].

After those seminal approaches, both the notion of agency and its interpretation in computational logic have greatly evolved. Many computational-logic-based agent- oriented languages and frameworks to specify agents and Multi-Agent Systems (MAS) have in fact been defined over time (for a survey of these languages and architectures the reader may refer, among many, to [8]–[10]). Their added value with respect to non-logical approaches is to provide clean semantics, readability and verifiability, as well as transparency and explainability 'by design' (or almost), as logical rules can easily be transposed into natural-language explanations.

DALI [42], [43], [60] is an Agent-Oriented Logic Programming language, where the autonomous behaviour of a DALI agent is triggered by several kinds of events: external events, internal, present and past events.

External events are syntactically indicated by the postfix E. Reaction to each such event is defined by a reactive rule, where the special token :>. The agent remembers to have reacted by converting an external event into a past event (postfix P). An event perceived but not yet reacted to is called "present event" and is indicated by the postfix N. It is often useful for an agent to reason about present events, that make the agent aware of what is happening in its external environment.

In DALI, actions (indicated with postfix A) may have or not preconditions: in the former case, the actions are defined by actions rules, in the latter case they are just action atoms. The new token :< characterizes an action rule that specifies an action's preconditions. Similarly to events, actions are recorded as past actions.

Internal events is the device which makes a DALI agent proactive. An internal event is syntactically indicated by the postfix I, and its description is composed of two rules. The first one contains the conditions (knowledge, past events, procedures, etc.) that must be true so that the reaction (in the second rule) may happen. Thus, a DALI agent is able to react to its own conclusions. Internal events are automatically attempted with a default frequency, customizable by means of user directives.

The DALI communication architecture [44] implements the DALI/FIPA protocol, which consists of the main FIPA primitives[4] , plus few new primitives which are peculiar to DALI. Notice that, DALI has been made compatible with the Docker technology (cf. [61] for details). So, a DALI agent can be deployed within a container.

The semantics of DALI is based upon the declarative semantic framework introduced in [45], aimed at encompassing approaches to evolving logical agents, by understanding changes determined by external events and by the agent's own activities as the result of the application of program- transformation functions.

We abstractly formalise an agent as the tuple $Ag = \langle P_{Ag}, E, I, A\rangle$ where Ag is the agent name and P_{Ag} is the "agent program" according to the specific language adopted. E is the set of the external events, i.e, events that the agent is capable to perceive and recognize: let $E = \{E_1, \dots, E_n\}$ for some n. I is the set of internal events (distinguished internal conclusions, that may include agent's desires and intentions): let $I = \{I_1, \dots, I_m\}$ for some m. A is the set of actions that the agent can possibly perform: let $A = \{A_1, \dots, A_k\}$ for some k. Let $ev = (E \cup I \cup A)$.

In the DALI syntax, used below for the examples, atoms indicated with a postfix correspond to events of various kinds. In particular, if p is an atom, pE is an external event, pA is an action and pI an internal event.

According to this semantic account, one will have an initial program Π0 obtained by the program P_{Ag} provided by a programmer. According to events that happen, agent's activities and internal reasoning, and actions which are performed, Π_0 will "evolve" through corresponding program-transformation steps (each one transforming Π_i into Π_{i+1}, cf. [45]), and thus gives rise to a Program Evolution Sequence $PE = [\Pi_0, \dots, \Pi_n, \dots]$. The program evolution sequence will imply a corresponding Semantic Evolution Sequence $ME = [M_0, \dots, M_n, \dots]$ where M_i is the semantic account of Π_i.

Different languages and different formalisms in which an agent can possibly be expressed will influence the following key points: (i) when a transition from Π_i to Π_{i+1} takes place, i.e., which are the external and internal factors that determine a change in the agent; (ii) which kind of transformations are performed; (iii) which semantic approach is adopted, i.e., how M_i is obtained from Π_i.

The semantic account includes an Initialization step, where the program P_{Ag} written by the programmer is transformed into a corresponding program Π_0 by means of some sort of knowledge compilation. In DALI for instance, the initialization step extracts the

[4] FIPA is a widely used standardized ACL (Agent Communication Language), cf. http://www.fipa.org/specs/fipa00037/SC00037J.html for language specification, syntax and semantics.

list of internal and external events, and the control directives that are associated to the program (e.g., for defining priorities among events and frequencies for checking the occurrence of events). In general in fact, Π_0 can be simply a program (logical theory) or can have additional control information associated to it.

Agents usually record events that happened and actions that they performed. Notice that an agent can describe the state of the world only in terms of its perceptions, where more recent remembrances define the agent's approximation of the current state of affairs. We thus define set $\mathscr{P}$ of current (i.e., most recent) past events, and a set PNV where we store all previous ones (under certain conditions). We define the 'history' H of an agent as the tuple $\langle \mathscr{P}, PNV \rangle$, dynamically augmented with new events that happen. In DALI, a past event in P is in the form $pP : T_i$, where p is an atom corresponding to an event, postfix P stands for 'past' and T_i is a time-stamp indicating when the event has been perceived. In [62] we have defined Past Constraints, which allow one to define when and upon which conditions (apart from arrival of more recent versions) past events should be moved into PNV, and later on possibly removed.

Definition 1 (Evolutionary semantics). Let Ag be an agent. The evolutionary semantics ε^{Ag} of Ag is a tuple $\langle H, PE, ME \rangle$, where H is the history of Ag, and PE and ME are its program and semantic evolution sequence.

DALI has been fully implemented, and a programming environment has been devised. The DALI programming environment [60] is freely available, and at the current stage of development offers a multi-platform folder environment, built upon Sicstus Prolog [63] programs, shells scripts, Python scripts to integrate external applications, a JSON/HTML5/jQuery web interface to integrate into DALI applications, with a Python/Twisted/Flask web server capable to interact with A DALI MAS at the backend. We have recently devised a cloud DALI implementation, reported in [64], [65]. As shown in [64], the preexisting DALI framework has been extended to "DALI 2.0" by using open sources packages, protocols and web-based technologies. DALI agents can thus be developed to act as high level cognitive robotic controllers, and can be automatically integrated with conventional embedded controllers. The web compatibility of the framework allows real-time monitors and graphical visualizers of the underline MAS activity to be specified, for checking the interaction between an agent and some external device, that can possibly be a robotic subsystem. The cloud package ServerDALI allows a DALI MAS to be integrated into any practical environment. In [65] illustrate the recent "Koiné DALI" framework, where a Koiné DALI MAS can cooperate without problems with other MASs, programmed in other languages (logical or non-logical), and with object-oriented applications. In summary, the enhanced DALI can be used for multi-MAS applications and hybrid multi-agents and object-oriented applications, and can be easily integrated into preexisting applications.

The DALI framework has been experimented, e.g., in applications for: unattended hardware testing of hardware-software platforms in telecommunication industry; user monitoring and training; emergencies management (such as first aid triage assignment); security or automation contexts; home automation and processes control. More generally, DALI has proved to be useful in every situation that is characterised by asynchronous events sources that require reasoning over a dynamic data collection: either simple events, and/or events that are correlated to other ones even in complex patterns. In fact, in order to be able to perform Complex Event Processing, i.e., to actively monitor event data so as to make automated decisions and take time-critical actions, DALI has been empowered with CEP capabilities [66], of which the implementation at this day is partial, but is being actively developed: since the 2018 release, DALI supports the double concurring events occurrence in a predefined time window, so that reaction rules can be defined where two events from different asynchronous sources happen to fall in the same time interval. An architecture encompassing DALI agents and called F&K (Friendly-and-Kind) system [67] has been proposed for (though not restricted to) applications the e-Health domain. We have since long equipped DALI with a plugin for invoking ASP solvers and thus executing ASP modules in the so called ASP_DALI event, postfix P stands for 'past' and T_i is a time-stamp extension available at our github organization repositories[5]. An ongoing experimentation is about emotion recognition in the context of cognitive robotics [68], were real time analysis of the non verbal communication interaction between a human and the anthropomorphic NAO robot is performed by an extended DALI, consisting in an ASP_- DALI and QuLog/Teleor [69] multi-agent system. In this experimental setup, several sub-symbolic perception systems generate real-time fluents about the emotional state of the human while interacting with the robot, and the MAS in background determines the best emotional state according to a predefined model in a timely manner, so as to suggest the most appropriate behaviour to the robot.

VI. DALI and ASP in Synergy: Past Work

The work presented in [11] studied the application of DALI and ASP to the problem of dynamic goal decomposition and planning in scenarios characterised by a strong inter-dependency between action and context, for instance those related to rescue intervention in a territory upon occurrence of some kind of catastrophic event. The paper in particular proposed an architecture that integrates DALI MASs (DALI Multi-Agent Systems) and ASP modules for reaching goals in a flexible and timely way.

The effectiveness of this solution was demonstrated by means of a case-study where DALI agents cooperate in order to explore an unknown territory. The solution is based upon a MAS instead of a monolithic software solution because it is important that each software component, implemented as an agent, can partially retain its autonomy during asynchronous event processing. In fact, in this way each agent can be enriched with high-level reasoning/control behaviours that can coexists with the planning/executing activity. The MAS solution also permits to distribute the computational effort and increases overall robustness.

The DALI MAS is intended to fulfill the so-called bounded rationality principle, by which a plan for reaching a goal has to be devised and executed in a timely manner before a ultimate T_{max} deadline. There is a second deadline $T_{PlanMax} < T_{Max}$ by which a plan has to be computed and selected, so that the remaining time is sufficient for plan execution.

In the context of microservices we might improve this solution by defining a specific agent role called "micro- meta-planner" that shall supervise the task allocation over ASP and DALI agents, and which is responsible of the real-time compliance of the overall system. For example, in those situation were the ASP module could not deliver answer sets in polynomial time, the micro-planner shall take over either by providing a fail-safe plan, or by providing a set of short plans' definitions aimed to obtain better working conditions for the ASP solver and its grounding subsystem, such as the GRINGO grounder [70].

Thus, given the input set $T_{PlanMax}$, T_{Max}, G, N, where G is the goal, and N is the instance size of the problem to be solved (if applicable), the MAS operates via the following steps.

i) Decompose the overall goal into suitable sub-goals;

[5] https://github.com/AAAI-DISIM-UNIVAQ

ii) For each sub-goal, generate (via an ASP module) a plan within the $T_{PlanMax}$ deadline;

iii) Execute the plan within the T_{Max} deadline; in case of failure (insufficient time), maximize the length of the partially executed plan;

iv) In case of a change of conditions in the environment, re- plan, possibly limiting this activity to specific sub-goals resulting from the partitioning.

Sub-goals can be determined by any kind of goal partitioning algorithm. In the disaster management case study, it was obtained simply by sub-dividing the main geographical area into slightly overlapping sub-territories.

The planner agent equipped with an ASP module may find more than one plan for each (sub-)goal; so, metrics can be applied by which a plan could be preferred to another one.

VII. μ ASPSv's: Specification and Implementation Guidelines

The present work can be seen as an evolution of the work in [11], in the sense that we make it possible for an ASP program to act as an independent component, instead of being invoked as an auxiliary module by an agent.

In this section we provide in fact an abstract definition of a μASPSv, and some more specific indication of how such a component might be enacted and inserted into a distributed system, and how the inner ASP program might be structured.

Definition 2. Let Π be an ASP program, and let $U = Undef(\Pi)$. A μASPSv based upon Π, denoted as μASPSv(Π), has the following specification:

- Inner ASP program Π;
- Activation signal A (optional), with $A \in Undef(\Pi)$;
- Stop signal S (optional), with $S \in Undef(\Pi)$;
- Input set $\{I_1, ..., I_k\} \subseteq Undef(\Pi)$;
- Output set $\{O_1, ..., O_h\} \subseteq Heads(\Pi)$.
- Query result set $\{Q_1 = v_1, ..., Q_r = v_r\}$ where $\{Q_1, ..., Q_r\}$ are queries[6], formulated over atoms occurring in $Heads(\Pi)$ and the v_is can have value "true" or "false".

The elements listed above have the following meaning.

Whenever the activation signal is expected, if A is not true in Π, then μASPSv(Π) is in a state of no-operation.

Whenever the stop signal is expected, if S becomes true in Π, then μASPSv(Π) will go back into a state of no-operation.

The input set is a set of atoms that, when some of them are added to Π, contribute to answer sets computation. Each of such atom corresponds to an input/request received from the μASPSv's surrounding environment.

The output set is a set of atoms extracted from the answer sets of Π plus the current input set. Each of these atoms corresponds to an output/answer to be delivered into the μASPSv's surrounding environment.

The query result set is a set of truth values elicited from the answer sets of Π. Each of these values corresponds to result of a query, to be delivered into the μASPSv's surrounding environment.

Notice that, we admit as inputs atoms included in $Undef(\Pi)$, i.e., atoms that have truth value "undefined" in the well-founded model. This means that external inputs are intended to activate behaviors in program Π, without however threatening is basic functioning, represented by the atoms which are true or false in the well-founded model.

[6] c.f. previous section for possible queries.

In order to make it possible for μASPSv(Π) to operate dynamically, thus receiving inputs and delivering outputs and answers, a suitable shell program must be defined, in any programming language able to be interfaced with an answer set solver. Below we provide a schematic essential definition of such a shell program, to be used as a guideline for actual definition and implementation. The shell program will rely upon an input-output table, where each potential and actual input and potential and actual output will be annotated, together with the list of external components sending inputs, and the list of external components to which outputs are to be delivered.

Definition 3. The shell responsible to manage an ASP microservice μASPSv(Π) can be specified by the following pseudo-code.

begin

1. while not activation then no-operation endwhile;
2. if activation then add atom A to Π as a fact to bring μASPSv(Π) into operation;
3. while not stop do at frequency f
   (a) detect and annotate actual inputs $\{I_{j1}, ..., I_{jr}\} \subseteq \{I_1, ..., I_k\}$;
   (b) add $\{I_{j1}, ..., I_{jr}\}$ to Π as facts;
   (c) obtain the answer sets $\{S_1, ..., S_n\}$ of (the augmented) Π;
   (d) elicit outputs $\{O_1, ..., O_v\} \subseteq \{O_1, ..., O_h\}$;
   (e) extract query results $\{Q_1, ..., Q_t\} \subseteq \{Q_1, ..., Q_r\}$;
   (f) deliver outputs and query results according to requests;
   (g) remove $\{I_{v1}, ..., I_{vs}\} \subseteq \{I_{j1}, ..., I_{jr}\}$ from Π
   (h) and remove relative annotations;
   endwhile;
4. add atom S to Π as a fact and remove atom A, to bring Π into no-operation.

end.

This shell program is able to activate and stop a μASPSv, and to execute, until possibly a stop signal arrives, a loop where: the inputs are received from the external environment and delivered to Π; and, outputs and query results are extracted from the answer sets of Π (given the inputs) and delivered to the external environment. Precisely, each input will arrive from some external component, and each output will have to be delivered to some other (or to the same) component. At the end of each cycle some or all of the inputs will be removed from Π and the relative annotations will be eliminated; removing all the inputs determines a stateless component, while omitting to remove some of the inputs, forever of for some time interval, accounts to defining a stateful component. Input detection will occur at a certain frequency, suitable for each particular kind of component, environment, and application domain. Some of the inputs may come from sensors (and therefore they do not require any answer) and some of the outputs may go to actuators. This is also annotated in the input-output table. The parts concerning the activation and stopping of the μASPSv (first and second line after the begin, and last line before the end) will be omitted if the component is running forever rather being first activated and then stopped.

Notice that the above definitions can find easier practical application thanks to the advanced features of modern solvers such as clingo [7], that provides "multi-shot"solving features, coping with grounding and solving in continuously changing logic programs. In particular, "multi-shot"solving allows a given ASP program to evolve during the reasoning process, because data or constraints are added, deleted, or replaced. This is exactly what is needed in order to send to a μASPSv the activation and stop signals, and to cancel old inputs and add new ones.

Many practical aspects remain however to be defined in order to obtain an implementation. For instance, if a μASPSv is to be situated within a multi-agent system, input-output-query exchange might happen by means of the above-mentioned FIPA ACL. The shell program can be made FIPA-compliant (i.e., able to exchange and understand FIPA messages) either by developing suitable code, or, better, by importing a suitable library such as, e.g., the freely available JADE library[7]. The JADE library is an advance middleware that offers many functionalities to "agentify" imperative or object-oriented or other kinds of programs. In fact it provides: the agent abstraction (i.e., a given program, when running, is seen by the external environment as an agent); the ability of peer to peer inter-agent FIPA asynchronous message-passing; a yellow pages service supporting subscription of agents and a discovery mechanism, and many other facilities to support the development of distributed systems.

So for instance, an input can be sent to a μASPSv via a FIPA "request" message with the input as argument, to be interpreted on the μASPSv's side as a request to reply with a "confirm" message, containing the corresponding output. A query can be sent to the μASPSv via a FIPA "query-if" message whose answer will be again a "confirm", conveying the truth value of the query. Notice that, to avoid ambiguities, the FIPA syntax provides the facility to identify each message via a certain arbitrary identifier, so that the answer message can indicate that it is 'in-reply-to' to that identifier.

The JADE yellow pages services might be exploited by μASPSv's which would want to register as agents with a name and a role, and then communicate with each other in an asynchronous way. Or, since most MASs offer such a mediator service, μASPSv's might enroll in any agent community. Finally, they might communicate peer-to-peer with other agents that they are aware of, or that they locate via the mediator.

Let us now consider how to structure the 'core' program Π, on which a microservice μASPSv(Π) is based. First, activation and stopping of a module can be simply obtained by a couple of constraints, that make the program inconsistent (in no-operation state) if either activation A has not arrived, or stopping signal S has been issued:

$:- \; not \; A.$ % *module activation*

$:- \; S.$ % *module stop*

Then, when the module has been activated, upon arrival of new inputs, the inner program Π will in general 'produce' (admit) answer sets. If the answer set is unique then the outputs can be univocally identified. Otherwise the shell, in the 'elicit outputs' part, will have to adopt some kind of policy (e.g., preferences, utilities, costs or other) to select which answer set to consider. The queries, being by definition specified upon the whole set of answer sets, will always return an univocal result. In case, given the present input, Π should be inconsistent, then the output will consist in a failure signal (e.g., in the FIFA ACL, there is the "failure" primitive to be used in such cases).

VIII. Case Study

The case study that we propose here is inspired to issues raised by applications related to autonomous vehicles. Presently, machine learning mechanism have been defined to allow autonomous cars to comply with traffic lights by detecting their color, so as to pass with green and stop with red similarly to traditional cars. Such mechanisms must be trained, are prone to errors, and are potentially subject to adversarial machine learning.

In our view, physical traffic lights might in perspective disappear, to be substituted by monitoring agents that would receive requests to pass from cars and consequently issue authorisations. This either in routes dedicated to autonomous vehicles, or in the (very reasonable) hypothesis to equip also 'traditional' cars with a device to interact with the monitoring agents.

Below we propose the sample design of the inner program concerning a μASPSv which implements the monitoring agent of a road intersection, taking the place of a physical traffic light. In the example, the traffic light agent is called *tl* and, for the sake of simplicity, behaves like a 'real' traffic light but just takes the colors green (*g* for short) and red (*r* for short). In fact, the yellow is no longer necessary as we assume that the involved cars (each one equipped with its own driver agent) will obey the directives. We have two lanes, one going north-south (*ns* for short) and the other one east-west (*ew* for short), crossing at the traffic light. If the traffic light is green in one direction it must be red in the other one, and vice versa. The traffic light is activated by a signal $active(t1)$, and never stopped unless there is a fault, detected by the module itself by means of a sensor. A fault is supposed to have occurred whenever $fault\_tl$ is true, i.e., it has been returned by the sensor.

$tln(t1).$ % *Traffic – Light Identifier*

$active(t1).$

$:- not \; active(t1).$ % *Sensor Check activation*

$:- lane(L), fault\_tl(t1, L, T).$ % *Sensor Check Possible Fault*

Each car, say here $c1$, $c2$, $c3$, $c4$ and $c5$[8], wants to go, but it is allowed to proceed only if it gets the green traffic light. Otherwise, it remains dummy. We assume that all cars behave in the same way. Each one issues a request of format $car(C)$, $want\_go(C, t1, L, T)$ where L is the lane, with possible values *ns* for north-south and *ew* for east-west; T is the time of the request. Requests by various cars may for example give rise to the addition of the following facts to the μASPSv's program.

$\%INPUT : CARS$

$car(c1).$

$car(c2).$

$car(c3).$

$car(c4).$

$car(c5).$

$\%INPUT : REQUESTS$

$want\_go(c1, t1, ns, 2).$

$want\_go(c2, t1, ns, 2).$

$want\_go(c3, t1, ew, 2).$

$want\_go(c4, t1, ns, 4).$

$want\_go(c5, t1, ew, 4).$

The following facts and rules define the lanes, and specify that this monitoring agent has a lookahead of five time instants: after that, it will have to be re-run.

$lane(ns).$

$lane(ew).$

$time(1..5).$

$next(Y, X) :- time(X), time(Y), Y = X + 1.$

[7] https://jade.tilab.com where references to several related publications can also be found

[8] The specification of which cars come and go in the traffic light surroundings can be within the module's inputs, and so the car list will be updated by the shell.

The rules below define the color that the traffic light takes (in a very standard way) as transitions from green to red and vice versa, where the initial color is green. In reality, such a monitoring agent can employ a much more sophisticated protocol such as for instance the Contract Net Protocol (CNP). If adopting CNP, the agent might grant priority to particular kinds of vehicles, e.g., police cars, ambulances, cars transporting a disabled person, etc. More generally, any policy to grant passage according to criteria could be implemented.

tl (*r, TL, L*1*, T* 1) :–
    time(*T*), *lane*(*L*1), *lane*(*L*2), *tln*(*TL*), *L*1! = *L*2,
    next(*T*1*, T*), *tl*(*g, TL, L*1*, T*), *tl*(*r, TL, L*2*, T*).
tl (*g, TL, L*1*, T*1) :–
    time(*T*), *lane*(*L*1), *lane*(*L*2), *tln*(*TL*), *L*1! = *L*2,
    next(*T* 1*, T*), *tl*(*r, TL, L*1*, T*), *tl*(*g, TL, L*2*, T*).
tl (*g, TL, ns,* 1) :– *tln*(*TL*).
tl (*r, TL, ew, T*) :–*tln*(*TL*), *time*(*T*), *tl*(*g, TL, ns, T*).

In our case the implemented protocol is fair, as cars that cannot go now because it is red on their lane will be deferred to the next time instant (by delaying their request), when the color will be green (output in format *go* (*Car, t*1*, Lane, Time*)).

go (*C, TL, L, T*) :–
    time(*T*), *car*(*C*), *tln*(*TL*), *lane*(*L*),
want_go(*C, TL, L, T*), *tl*p*g, TL, L, T*).
wait(*C, TL, L, T*) :–
    time(*T*), *car*(*C*), *tln*(*TL*), *lane*(*L*),
want_go(*C, TL, L, T*), *tl*(*r, TL, L, T*).
want_go(*C, TL, L, T* 1):– *car*(*C*), *tln*(*TL*), *lane*(*L*),
wait(*C, TL, L, T*), *next*(*T* 1*, T*).
:– *time*(*T*), *car*(*C*), *tln*(*TL*), *lane*(*L*),
  go(*C, TL, L, T*), *tl*(*r, TL, L, T*).

Clearly, this program can be 'cloned' (mutatis mutandis) to manage any number of traffic lights. For the reader's convenience, this program is standalone and can be run exactly as it is to check its results.

We now provide a definition of a car in DALI. Or rather, we define an agent capable to manage the situation where the car has to pass an intersection controlled by a μASPSv such as the one defined above. This agent will presumably be a component of an overall multi-agent system managing the many appliances included in most recent cars.

The agent will receive data about its present position from an infrastructure (which the road system may be equipped with at low cost), that will periodically broadcast the information, that will be received by cars. Then, the car will sense the presence of a crossing (with its associated traffic-light component) from a signal broadcasted up to a certain distance, that will communicate the identifier *tl* of that traffic light. The car will annotate the present position's external events as past events (a reaction that does nothing has exactly the purpose of annotating), where the most recent past event will be taken by default in consideration during subsequent operation, to extract position parameters. An external event signalling the presence of a crossing will determine a reaction where the agent issues a request to pass to *tl*. The request will be issued by sending a message whose performative will be the FIPA primitive *request*. The message will include the agent's name (available in the predefined special variable *Me*) and the present time, obtained by the system's primitive *time* (*T*). The predefined predicate *messageA* (...) is processed by the DALI communication architecture, which will fill the remaining unspecified parameters expected by the FIPA syntax with default values, and will actually send out a correct FIPA message. The agent becomes aware of being enabled to pass when, via the *enabled_passI* first rule (where posfix *I* indicates an internal event), that will be attempted automatically at a certain frequency, it will detect the arrival of a message containing the FIPA primitive *accept_proposal*. This primitive signals that the traffic light accepts the request, and thus grants the permission, in this case unconditionally: the list which occurs as second parameter (here empty) might in general indicate conditions to be fulfilled. So, success of the internal event via the first rule determines a reaction (second rule), which consists in the action *passA* that will be physically enacted by the car.

present_positionE(*Road, Direction*) :> *true.*
crossingE(*TL*) :> *request_to_pass*(*TL*)*.*
request_to_pass(*TL*) :–*time*(*T*)*,*
    present_positionP(*_, Direction*)*,*
    messageA(*TL,*
        request(*want_go*(*Me, TL, Direction, Time*)*, Me*))*.*
enabled_passI :–*messageA*(*TL,*
        accept_proposal(*want_go*(*_, _, _*)*,* []*, Me*))*.*
enable_passI :> *passA.*

To make the two components interact it is not needed to import the whole FIPA protocol. For this simple case, the traffic light μASPSv's shell may extract the request from the input message, and "package" the permission to pass (when granted) into the required syntax before sending it back to the agent. An underlying (though minimal) middleware must be implemented, so that each component (many cars and traffic lights might in fact be present) can send/receive input/outputs to the others. Notice that, as said before, DALI has been integrated with the Docker technology, that may help to get this part "for free" or almost.

IX. Overall System's Semantics

The semantics of a single μASPSv is fully specified by: (i) the answer sets of the inner ASP program; (ii) the policy employed in its shell to select one single answer set; (iii) the set of queries that the shell possibly performs over the entire set of answer sets, whose meaning is formally specified in [26], [71]. We aim however to provide a semantics for the overall distributed system composed of heterogeneous microservices (where some of them can be agents), in order to provide a firm ground and a guideline for implementation.

To do so, we resort to Multi-Context Systems (MCSs), that are a well-established paradigm in Artificial Intelligence and Knowledge Representation, aimed to model information exchange among heterogeneous sources [72]–[74]. However, with some abuse of notation (and some slight loss of generality) we adapt and readjust the definitions to fit into our framework. To represent the heterogeneity of sources, each component in a Multi-context system, called 'context', is supposed to be based on its own logic, defined in a very general way [73]. In particular, a logic is defined by the following features.

- A set *F* of possible formulas (or *KB*-elements) under some signature.
- A set *KB* of knowledge bases built out of elements of *F*. in our framework, *KB* can also be a program in some programming language.
- A function *ACC*, where *ACC* (*kb, s*) means that *s* is an acceptable set of consequences of knowledge base $kb \in KB$, i.e., $s \subseteq Cn$, where *Cn* is the set of all possible consequences that can be drawn from *kb*. We assume here that *ACC* produces a unique set of consequences. In case of a program written in a non-logical programming language, such set can be the set of legal outputs given some input, that will be a subset of all possible outputs *Cn*; for logical

components, it will be (one of) the *kb* model(s). For instance, as we have seen the shell of a µASPSv will produce as consequences the elements occurring in the answer set selected according to some policy, along with query results.

A (Managed) multi-context system (MCS)

$$M = \{C_1, ..., C_r\}$$

is a set of $r = |M|$ contexts, each of them of the form $C_i = \langle c_i, L_i, kb_i, br_i \rangle$, where:

- c_i is the context name (unique for each context; if a specific name is omitted, index i can act as a name). In [75] a context's name can be a term called "context designator", denoting the kind of context (for instance, mycardiologist(c), customercare(c), helpdesk(h), etc.).
- L_i is a logic.
- $kb_i \in KB$ is a knowledge base.
- br_i is the set of bridge rules this context is equipped with.

Contexts in an MCS are meant to be heterogeneous distributed components, that exchange data. In fact, bridge rules are the key construct of MCSs, as it describes in a uniform way the communication/ data exchange patterns between contexts. Each bridge rule $\rho \in br_i$ has the form

$$op_i(s) \longleftarrow (c_1 : p_1),..., (c_j : p_j) \quad (1)$$

where the left-hand side s is called the head, and the right-hand side is called the body, and the comma stands for conjunction. The meaning is that, each data item p_i is supposed to come from context c_i. Whenever all the c_1,..., c_j have delivered their data item to the destination context *ci*, the rule becomes applicable[9]. In case context designators are employed, prior to checking a bridge rule for applicability, such terms must have been substituted by actual context names from which to acquire the data. For µASPSv's, this task will be performed by the shell, that must then be endowed with a list of contexts of eachtype. When the rule is actually applied (where, in our approach, application is optional and must be explicitly triggered in the destination context's code), its conclusion *s*, once elaborated by operator op_i, will be added to c_i's knowledge base. Operator op_i can perform any elaboration on the "raw" input *s*, such as format conversion, filtering, elaboration via ontologies, etc. Its operation is specified via a management function mng_i, which is thus crucial for knowledge incorporation from external sources. For simplicity, here we assume mng_i to be monotonic (i.e., to produce from *s* one or more data items) . Therefore, we can extend the previous definition of a context as

$$Ci = \langle c_i, L_i, kb_i, br_i, mng_i \rangle.$$

Notice that, in [66], [76], [77], the MCS approach has been extended so that a context can possibly be a logic- based agent, and extensions to bridge-rules format have been introduced for data and ontologies exchange in this new setting.

A data state (or belief state) $\vec{S}$ of an MCS M is a tuple $\vec{S} = (S_1, ..., S_r)$ such that for $1 \leq i \leq r$, $S_i \subseteq Cn_i$. A data state can be seen as a view of the distributed system by an external "observer". $app(\vec{S})$ is the set composed of the head of those bridge rules which are applicable in $\vec{S}$. This means, in logical terms, that their body is true w.r.t. $\vec{S}$. In practical terms, we may say that a bridge rule ρ associated to context *ci* is applicable in $\vec{S}$ if all the data mentioned in the body of the bridge rule can be delivered to the destination context. This is the case whenever they are available in the contexts of origin, i.e., they occur in the present respective data state items in $\vec{S}$. In the original formulation of MCS, all applicable bridge rules are automatically applied, and their results, after the elaboration by the management function, are added to the destination context's knowledge base, that therefore grows via bridge-rule application. Starting from a certain specific data state, some bridge rules will be applicable and therefore they will be applied. This will enhance the knowledge base in some of the contexts, thus determining (in these contexts) a new set of acceptable consequence, and therefore a new overall data state. In the new state other bridge rules will be applicable, and so on, until a "stable" state, called Equilibrium, will be reached. Technically, $\vec{S}$ is an equilibrium for an mMCS M iff, for $1 \leq i \leq |M|$,

$$S_i = ACC_i(mng_i\,(app(\vec{S}), kb_i))$$

which states that each element of the equilibrium is an acceptable set of consequences after the application of every applicable bridge rule, whose result has been incorporated into the context's knowledge base via the management function.

In [75] it is proved that, in the kind of MCS that we have just described, an equilibrium will be reached in a finite number of steps. Notice however that this definition assumes the system to be isolated from any outside influence, and that an equilibrium, one reached, will last forever. Instead, in real systems there will be interactions with an external environment, and so equilibria may change over time. Moreover, each context is not necessarily a passive receiver of data sent by others.

To take these aspects into account, [75] proposes some extensions to the original formulation, among which the following, that are relevant in the present setting.

- It is noticed that contexts' knowledge bases can evolve in time, not only due to bridge-rule application. In fact, contexts receive sensor inputs (passively or in consequence to active observation), or can be affected by user's modification (e.g., a context may encompass a relational database that can be modified by users). So, each context c_i will have an associated Update Operator $\mathcal{U}_i$ (that can actually consist in a tuple of operators, each one performing a different kind of update). Updates and bridge rules both affect contexts' knowledge base over time. So (assuming an underlying discrete model of time) we will be able to consider, when necessary, $c_i\,[T]$ meaning context c_i at time T, with its knowledge base $kb_i\,[T]$; consequently we will have an evolution over time of contexts. Therefore, we will have a definition (not reported here) of Timed Equilibria. Notice only that a timed equilibrium can be reached at time $T+1$ only if the actual elapsed time between T and $T+1$ is sufficient for the system to "stabilize" by means of bridge-rules application on the updated knowledge bases.
- Mandatory bridge-rule application (as it is in the original MCS definition) constitutes a limitation: in fact, contexts would be forced to accept inputs unconditionally, and this may be often inappropriate. Consider for instance a context representing a family doctor: the context may accept non-urgent patient's requests for appointments or consultation only within a certain time windows. So, [75] introduces conditional bridge-rule application, formalized via a timed triggering function, tr_i, which specifies which applicable bridge rules are triggered (i.e., they are practically applicable) at time *T*. It does so either based on certain pre-defined conditions, or by performing some reasoning over the present knowledge base contents. Therefore, the implementation of $tr_i\,[T]$ my require an auxiliary piece of program, that in a µASPSv's shell will presumably be a logic program.

So, considering contexts which are µASPSv's, in order to fit in the vision of the overall system as an MCS, their shell must be empowered as follows.

[9] In the original formulation of bridge-rule syntax, there can be additional literals not $(c_1 : p_{j+1})$, ..., not $(c_j : p_n)$ in the body, meaning that in order for the bridge rule to be applicable, the p_{j+1} ... p_n must be false in the relative contexts. We disregard this part, as non-logical components cannot use logical negation. There is no loss of generality however, as each of the p_1, ..., p_j can state a negative fact.

- Include the bridge rules associated to a μASPSv, and the definition/implementation of the triggering function.
- Include a facility to resolve the context designators, so as to check for applicability a triggered bridge rule after substituting context designators occurring therein with actual contexts' names.
- Include the definition of the specific management function, so as to be able to apply it on bridge-rules' results.

In the case study of previous section, each traffic light should be equipped with a bridge rule that, by means of the instantiation of a suitable context designator (say, anycar(c)) collects the cars' requests. Symmetrically, cars should be equipped with a bridge rule to collect the permission to go by the traffic light (the nearest one, whose identifier should replace a context designator of the form, e.g., a_traffic_light(t)). The triggering function may allow cars to enable reception of traffic-light communications only when needed.

Context designators are therefore useful to write general bridge rules to be then customized to the particular situation at hand. They also allow to devise a system where components do not know or are aware of each other in advance, and where components can possibly join/leave the system at any time. A suitable middleware should be realized to allow component's shells to instantiate bridge rules. In our case study, that concerns an infrastructure for car traffic, both cars and traffic lights might for instance broadcast their name and geo-localization. In this way, cars might locate the traffic light of interest, and traffic lights might become aware of nearby cars that might send them a request.

X. Concluding Remarks

We have proposed a methodology for developing microservices in Answer Set Programming, by means of the creation of a particular kind of components, that can be activated/stopped, can receive external requests and can deliver answers. We have provided a definition of μASPSv's and explained how they might be implemented, and we have outlined a programming methodology. We have shown by means of a case study how such components can be defined, and how they might interact with other heterogeneous components, e.g., DALI logical agents.

We have also outlined a possible uniform semantics to specify an heterogeneous system in which μASPSv's could be situated, also in synergy with logic-based autonomous agents. This is an absolute novelty for microservices in general, as no attempt has ever been made to provide such a uniform model for an overall system. The proposed semantics can constitute the ground for principled implementations. Overall, this work can be considered as a creative combination of existing technologies, in view of entirely new application domains of answer set programming and logic programming in general.

Important application fields for μASPSv's are Cloud computing and IoT. We consider particularly important the various kinds of robotic applications and the underlying infrastructural aspects (as shown in the case study related to autonomous vehicles), and human-robot interaction. Promising future applications might concern personalised assistance in healthcare, where heterogeneous components might include: μASPSv's that manage sensors such as wearable devices to monitor the patient's conditions; personal assistant (possibly robotic) agents; and components representing the available appliances for patient's management and vital support, and knowledge sources that provide criteria for, e.g., evaluation of medical checks, dosage of drugs, and medical diagnosis.

Future work includes: develop a real implementation; refine the programming methodology; provide a user friendly graphical interface, and perform experiments in realistic environments. We plan to carry out an effective integration of μASPSv's and DALI multi-agent systems, and extend it to heterogeneous systems, possibly including also QuLog/Teleor and AgentSpeak agents. We will then perform experiments in the various domains where DALI is being applied, including robotics. We have in mind applications concerning cognitive robotic architectures, comprising hybrid multi-agent systems with object detectors as perception layer, and DALI-ASP as reasoning layer.

Acknowledgements

We thank the anonymous reviewers for their interest in the DALI language, and for insightful suggestions about modern ASP grounders/solvers. This work has been realized and funded over the years only by internal basic research activities of the authors, and by consulting work concerning the exploitation of DALI tools in real world applications, covered by no-disclosure agreements.

References

[1] G. Brewka, T. Eiter, M. T. (eds.), "Answer set programming: Special issue," AI Magazine, vol. 37, no. 3, 2016.

[2] M. Gelfond, V. Lifschitz, "The stable model semantics for logic programming.," in Logic Programming, Proceedings of the Fifth International Conference and Symposium, Seattle, Washington, August 15-19, 1988 (2 Volumes), vol. 88, 1988, pp. 1070–1080, MIT Press.

[3] M. Gelfond, V. Lifschitz, "Classical negation in logic programs and disjunctive databases," New generation computing, Springer, vol. 9, no. 3-4, pp. 365–385, 1991, doi: 10.1007/BF03037169.

[4] V. W. Marek, M. Truszczyński, "Stable models and an alternative logic programming paradigm," in The Logic Programming Paradigm, Springer, 1999, pp. 375–398, doi: 10.1007/978-3-642-60085-2_17.

[5] V. Lifschitz, "Answer set planning," in Logic Programming: The 1999 International Conference, Las Cruces, New Mexico, USA, November 29 - December 4, 1999, 1999, pp. 23–37, MIT Press.

[6] P. Jamshidi, C. Pahl, N. C. Mendonça, J. Lewis, S. Tilkov, "Microservices: The journey so far and challenges ahead," IEEE Software, vol. 35, no. 3, pp. 24–35, 2018, doi: 10.1109/MS.2018.2141039.

[7] M. Gebser, R. Kaminski, B. Kaufmann, T. Schaub, "Multi-shot ASP solving with clingo," Theory Pract. Log. Program., vol. 19, no. 1, pp. 27–82, 2019, doi: 10.1017/S1471068418000054.

[8] R. H. Bordini, L. Braubach, M. Dastani, A. E. F. Seghrouchni, J. J. Gómez-Sanz, J. Leite, G. M. P. O'Hare, A. Pokahr, A. Ricci, "A survey of programming languages and platforms for multi-agent systems," Informatica (Slovenia), vol. 30, no. 1, pp. 33–44, 2006.

[9] A. Garro, M. Mühlhäuser, A. Tundis, M. Baldoni, C. Baroglio, F. Bergenti, P. Torroni, "Intelligent agents: Multi-agent systems," in Encyclopedia of Bioinformatics and Computational Biology - Volume 1, S. Ranganathan, M. Gribskov, K. Nakai, C. Schönbach Eds., Elsevier, 2019, pp. 315-320, doi: 10.1016/b978-0-12-809633-8.20328-2.

[10] R. Calegari, G. Ciatto, V. Mascardi, A. Omicini, "Logic-based technologies for multi-agent systems: a systematic literature review," Auton. Agents Multi Agent Syst., vol. 35, no. 1, p. 1, 2021, doi: 10.1007/s10458-020-09478-3.

[11] S. Costantini, G. De Gasperis, "Dynamic goal decomposition and planning in MAS for highly changing environments," in Proceedings of the 33rd Italian Conference on Computational Logic, Bolzano, Italy, September 20-22, 2018, vol. 2214 of CEUR Workshop Proceedings, 2018, pp. 40–54, CEUR-WS.org.

[12] D. Ameller, X. Burgués, O. Collell, D. Costal, X. Franch, M. P. Papazoglou, "Development of service-oriented architectures using model-driven development: A mapping study," Information and Software Technology, vol. 62, pp. 42 – 66, 2015, doi: https://doi.org/10.1016/j.infsof.2015.02.006.

[13] C. Legner, R. Heutschi, "Soa adoption in practice-findings from early soa implementations," 2007, Association for Information Systems.

[14] N. Dragoni, S. Giallorenzo, A. L. Lafuente, M. Mazzara, F. Montesi, R. Mustafin, L. Safina, "Microservices: yesterday, today, and tomorrow," in Present and ulterior software engineering, Springer, 2017, pp. 195–216, doi: 10.1007/978-3-319-67425-4_12.

[15] L. De Lauretis, "From monolithic architecture to microservices

architecture," in 2019 IEEE International Symposium on Software Reliability Engineering Workshops (ISSREW), 2019, pp. 93–96, IEEE.

[16] P. Krivic, P. Skocir, M. Kusek, G. Jezic, "Microservices as agents in iot systems," in KES International Symposium on Agent and Multi-Agent Systems: Technologies and Applications, 2017, pp. 22–31, Springer.

[17] S. Costantini, A. Formisano, "Answer set programming with resources," Journal of Logic and Computation, vol. 20, no. 2, pp. 533–571, 2010, doi: 10.1093/logcom/exp071.

[18] A. Dyoub, S. Costantini, G. De Gasperis, "Answer set programming and agents," Knowledge Eng. Review, vol. 33, p. e19, 2018, doi: 10.1017/S0269888918000164.

[19] S. Costantini, "About epistemic negation and world views in epistemic logic programs," Theory Pract. Log. Program., vol. 19, no. 5-6, pp. 790–807, 2019.

[20] W. Faber, "An introduction to answer set programming and some of its extensions," in Reasoning Web. Declarative Artificial Intelligence - 16th International Summer School 2020, Oslo, Norway, June 24-26, 2020, Tutorial Lectures, vol. 12258 of Lecture Notes in Computer Science, 2020, pp. 149–185, Springer.

[21] V. Lifschitz, "Twelve definitions of a stable model," in Proc. of the 24th Intl. Conf. on Logic Programming, vol. 5366 of LNCS, 2008, pp. 37–51, Springer.

[22] A. Van Gelder, K. A. Ross, J. S. Schlipf, "The well-founded semantics for general logic programs," Journal of the ACM, vol. 38, no. 3, pp. 620–650, 1991, doi: 10.1145/116825.116838.

[23] S. Costantini, A. Formisano, "Negation as a resource: A novel view on answer set semantics," in Logic Programming and Nonmonotonic Reasoning, 12th Intl. Conf., LPNMR 2013, vol. 8148 of Lecture Notes in Computer Science, 2013, pp. 257–263, Springer.

[24] K. Marple, G. Gupta, "Dynamic consistency checking in goal-directed answer set programming," TPLP, vol. 14, no. 4-5, pp. 415–427, 2014, doi: 10.1017/S1471068414000118.

[25] G. Gupta, E. Salazar, K. Marple, Z. Chen, F. Shakerin, "A case for query-driven predicate answer set programming," in ARCADE 2017, 1st International Workshop on Automated Reasoning: Challenges, Applications, Directions, Exemplary Achievements, Gothenburg, Sweden, 6th August 2017, vol. 51 of EPiC Series in Computing, 2017, pp. 64–68, EasyChair.

[26] S. Costantini, F. A. Lisi, R. Olivieri, "Digforasp: A european cooperation network for logic-based AI in digital forensics," in Proceedings of the 34th Italian Conference on Computational Logic, Trieste, Italy, June 19-21, 2019, vol. 2396 of CEUR Workshop Proceedings, 2019, pp. 138–146, CEUR-WS.org.

[27] Y. Shen, T. Eiter, "Evaluating epistemic negation in answer set programming," Artificial Intelligence, vol. 237, pp. 115–135, 2016, doi: 10.1016/j.artint.2016.04.004.

[28] T. Eiter, G. Gottlob, H. Veith, "Modular logic programming and generalized quantifiers," in Logic Programming and Nonmonotonic Reasoning, Springer, 1997, pp. 289–308, doi: 10.1007/3-540-63255-7_22.

[29] G. Ianni, G. Ielpa, A. Pietramala, M. C. Santoro, F. Calimeri, "Enhancing answer set programming with templates.," in 10th International Workshop on Non-Monotonic Reasoning (NMR 2004), Whistler, Canada, June 6-8, 2004, Proceedings, 2004, pp. 233–239.

[30] L. Tari, C. Baral, S. Anwar, "A language for modular answer set programming: Application to ACC tournament scheduling.," in Answer Set Programming, Advances in Theory and Implementation, Proceedings of the 3rd Intl. ASP'05 Workshop, Bath, UK, September 27-29, 2005, vol. 142 of CEUR Workshop Proceedings, 2005, CEUR-WS.org.

[31] S. Costantini, "On the existence of stable models of non-stratified logic programs," Theory and Practice of Logic Programming, vol. 6, no. 1-2, 2006, doi: 10.1017/S1471068405002589.

[32] V. Lifschitz, H. Turner, "Splitting a logic program.," in Logic Programming, Proceedings of the Eleventh International Conference on Logic Programming, Santa Marherita Ligure, Italy, June 13-18, 1994, vol. 94, 1994, pp. 23–37, MIT Press.

[33] T. Janhunen, E. Oikarinen, H. Tompits, S. Woltran, "Modularity aspects of disjunctive stable models," Journal of Artificial Intelligence Research, pp. 813–857, 2009, doi: 10.1613/jair.2810.

[34] E. Oikarinen, Modularity in answer set programs. PhD dissertation, Helsinki University of Technology, Finland, 2008.

[35] H. Gaifman, E. Shapiro, "Fully abstract compositional semantics for logic programs," in Proceedings of the 16th ACM SIGPLAN-SIGACT symposium on Principles of programming languages, 1989, pp. 134–142, ACM.

[36] M. Dao-Tran, T. Eiter, M. Fink, T. Krennwallner, "Modular nonmonotonic logic programming revisited," in Logic Programming, Springer, 2009, pp. 145–159, doi: 10.1007/978-3-642-02846-5_16.

[37] C. Baral, J. Dzifcak, H. Takahashi, "Macros, macro calls and use of ensembles in modular answer set programming," in Logic Programming, Springer, 2006, pp. 376–390, doi: 10.1007/11799573_28.

[38] M. Balduccini, "Modules and signature declarations for a-prolog: Progress report," Workshop on Software Engineering for Answer Set Programming (SEA'07), pp. 41–55, 2007.

[39] W. Faber, S. Woltran, "Manifold answer-set programs for meta-reasoning," in Logic Programming and Non-monotonic Reasoning, Springer, 2009, pp. 115–128, doi: 10.1007/978-3-642-04238-6_12.

[40] Y. Lierler, M. Truszczyński, "Modular answer set solving," Late-Breaking Developments in the Field of Artificial Intelligence, Bellevue, Washington, USA, July 14-18, AAAI, vol. WS-13-17, 2013.

[41] S. Costantini, "Answer set modules for logical agents," in Datalog Reloaded - First Intl. Workshop, Datalog 2010, Revised Selected Papers, O. de Moor, G. Gottlob, T. Furche, A. J. Sellers Eds., no. 6702 in Lecture Notes in Computer Science, Springer, 2011, pp. 37–58, doi: 10.1007/978-3-642-24206-9_3.

[42] S. Costantini, A. Tocchio, "A logic programming language for multi-agent systems," in Logics in Artificial Intelligence, Proc. of the 8th Europ. Conf.,JELIA 2002, LNAI 2424, 2002, pp. 1–13, Springer-Verlag, Berlin.

[43] S. Costantini, A. Tocchio, "The DALI logic programming agent-oriented language," in Logics in Artificial Intelligence, 9th European Conference, JELIA 2004, Proceedings, vol. 3229 of Lecture Notes in Computer Science, 2004, pp. 685–688, Springer.

[44] S. Costantini, A. Tocchio, A. Verticchio, "Communication and trust in the DALI logic programming agentoriented language," Intelligenza Artificiale, vol. 2, no. 1, pp. 39–46, 2005. Journal of the Italian Association AI*IA.

[45] S. Costantini, A. Tocchio, "About declarative semantics of logic-based agent languages," in Declarative Agent Languages and Technologies III, Third International Workshop, DALT 2005, Selected and Revised Papers, vol. 3904 of Lecture Notes in Computer Science, M. Baldoni, U. Endriss, A. Omicini, P. Torroni Eds., Springer, 2005, pp. 106–123, doi: 10.1007/11691792_7.

[46] S. Costantini, P. Dell'Acqua, L. M. Pereira, "A multilayer framework for evolving and learning agents," in Proceedings of Metareasoning: Thinking about thinking workshop at AAAI 2008, Chicago, USA, 2008.

[47] S. Costantini, A. Tocchio, "DALI: An architecture for intelligent logical agents," in Proceedings of the Int. Workshop on Architectures for Intelligent Theory-Based Agents (AITA08), AAAI Spring Symposium Series, 2008.

[48] S. Costantini, "Self-checking logical agents," in Proceedings of the Eighth Latin American Workshop on Logic, Languages, Algorithms and New Methods of Reasoning LA-NMR 2012, vol. 911 of CEUR Workshop Proceedings, 2012, pp. 3–30, CEUR-WS.org. Invited Paper, Extended Abstract in Proceedings of AAMAS 2013, 12th Intl. Conf. on Autonomous Agents and Multi-Agent Systems.

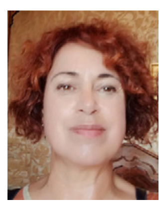
[49] S. Costantini, A. D'Andrea, G. De Gasperis, N. Florio, Tocchio, "DALI logical agents into play," in Proceedings of the AI*IA Workshop "Popularize Artificial Intelligence" (PAI-2012), 2012.

[50] S. Costantini, G. D. Gasperis, "Complex reactivity with preferences in rule-based agents," in Rules on the Web: Research and Applications - 6th International Symposium, RuleML 2012, Montpellier, France, August 27-29, 2012. Proceedings, vol. 7438 of Lecture Notes in Computer Science, 2012, pp. 167–181, Springer.

[51] S. Costantini, G. De Gasperis, G. Nazzicone, "Exploration of unknown territory via DALI agents and ASP modules," in Distributed Computing and Artificial Intelligence, 12th International Conference, DCAI 2015. Proceedings, vol. 373 of Advances in Intelligent Systems and Computing, 2015, pp. 285–292, Springer.

[52] S. Costantini, G. De Gasperis, G. Nazzicone, "DALI for cognitive robotics: Principles and prototype implementation," in Practical Aspects of Declarative Languages - 19th International Symposium, PADL 2017,

Proceedings, vol. 10137 of Lecture Notes in Computer Science, 2017, pp. 152–162, Springer.
[53] R. A. Brooks, "Intelligence without reason," in Proceedings of the 12th International Joint Conference on Artificial Intelligence. Sydney, Australia, August 24-30, 1991, 1991, pp. 569–595, Morgan Kaufmann.
[54] R. A. Brooks, "Intelligence without representation," Artif. Intell., vol. 47, no. 1-3, pp. 139–159, 1991.
[55] A. S. Rao, M. Georgeff, "Modeling rational agents within a BDI-architecture," in Proc. of the Second Int. Conf. on Principles of Knowledge Representation and Reasoning (KR'91), 1991, pp. 473–484, Morgan Kaufmann.
[56] A. S. Rao, "AgentSpeak (L): BDI agents speak out in a logical computable language," in Agents Breaking Away, 7th European Workshop on Modelling Autonomous Agents in a Multi-Agent World, Eindhoven, The Netherlands, January 22-25, 1996, Proceedings, Springer, 1996, pp. 42–55, doi: 10.1007/BFb0031845.
[57] S. Costantini, P. Dell'Acqua, G. A. Lanzarone, "Reflective agents in metalogic programming," in Meta-Programming in Logic, 3rd International Workshop, META-92, Proceedings, vol. 649 of Lecture Notes in Computer Science, 1992, pp. 135–147, Springer.
[58] R. A. Kowalski, F. Sadri, "Towards a unified agent architecture that combines rationality with reactivity," in Logic in Databases, International Workshop LID'96, Proceedings, vol. 1154 of Lecture Notes in Computer Science, 1996, pp. 137–149, Springer.
[59] P. Dell'Acqua, F. Sadri, F. Toni, "Combining introspection and communication with rationality and reactivity in agents," in Logics in Artificial Intelligence, European Workshop, JELIA '98, Dagstuhl, Germany, October 12-15, 1998, Proceedings, vol. 1489 of Lecture Notes in Computer Science, 1998, pp. 17–32, Springer.
[60] G. De Gasperis, S. Costantini, G. Nazzicone, "Dali multi agent systems framework, doi 10.5281/zenodo.11042." DALI GitHub Software Repository, July 2014. Accessed: December 2020, DALI: http:// github.com/AAAI-DISIM-UnivAQ/DALI.
[61] S. Costantini, G. De Gasperis, V. Pitoni, A. Salutari, "DALI: A multi agent system framework for the web, cognitive robotic and complex event processing," in Joint Proceedings of the 18th Italian Conference on Theoretical Computer Science and the 32nd Italian Conference on Computational Logic co-located with the 2017 IEEE International Workshop on Measurements and Networking (2017 IEEE M&N), vol. 1949 of CEUR Workshop Proceedings, 2017, pp. 286–300, CEUR-WS.org.
[62] S. Costantini, "Defining and maintaining agent's experience in logical agents," in Proc. of the Seventh Latin American Workshop on Non-Monotonic Reasoning LANMR 2011, vol. 804, 2011, pp. 151–165. (also in the Informal Proc. of the LPMAS "Logic Programming for Multi-Agent Systems" Workshop at ICLP 2011).
[63] M. Carlsson, P. Mildner, "Sicstus prolog–the first 25 years," arXiv preprint arXiv:1011.5640, 2010.
[64] S. Costantini, G. De Gasperis, G. Nazzicone, "DALI for cognitive robotics: Principles and prototype implementation," in Practical Aspects of Declarative Languages - 19th International Symposium, Proceedings, vol. 10137 of Lecture Notes in Computer Science, 2017, pp. 152–162, Springer.
[65] S. Costantini, G. De Gasperis, V. Pitoni, A. Salutari, "Dali: A multi agent system framework for the web, cognitive robotic and complex event processing," in Proceedings of the 32nd Italian Conference on Computational Logic, vol. 1949 of CEUR Workshop Proceedings, 2017, pp. 286–300, CEUR-WS.org. http://ceurws.org/Vol-1949/CILCpaper05.pdf.
[66] S. Costantini, "ACE: a flexible environment for complex event processing in logical agents," in Engineering Multi-Agent Systems, Third International Workshop, EMAS 2015, Revised Selected Papers, vol. 9318 of Lecture Notes in Computer Science, 2015, pp. 70–91, Springer.
[67] F. Aielli, D. Ancona, P. Caianiello, S. Costantini, G. De Gasperis, A. D. Marco, A. Ferrando, V. Mascardi, "FRIENDLY & KIND with your health: Human-friendly knowledge-intensive dynamic systems for the e-health domain," in Highlights of Practical Applications of Scalable Multi-Agent Systems. The PAAMS Collection International Workshops of PAAMS 2016, Proceedings, vol. 616 of Communications in Computer and Information Science, 2016, pp. 15–26, Springer.
[68] S. Costantini, G. De Gasperis, P. Migliarini, "Multiagent system engineering for emphatic human-robot interaction," in 2019 IEEE Second International Conference on Artificial Intelligence and Knowledge Engineering (AIKE), 2019, pp. 36–42, IEEE.
[69] K. L. Clark, P. J. Robinson, "Concurrent task programming of robotic agents in teleor.," in RuleML+ RR (Supplement), 2017.
[70] M. Gebser, T. Schaub, S. Thiele, "Gringo: A new grounder for answer set programming," in International Conference on Logic Programming and Non- monotonic Reasoning, 2007, pp. 266–271, Springer.
[71] S. Costantini, A. Formisano, "Query answering in resource-based answer set semantics," Theory and Practice of Logic Programming, vol. 16, no. 5-6, pp. 619–635, 2016, doi: 10.1017/S1471068416000478.
[72] G. Brewka, T. Eiter, "Equilibria in heterogeneous nonmonotonic multi-context systems," in Proc. of the 22nd AAAI Conf. on Artificial Intelligence, 2007, pp. 385–390, AAAI Press.
[73] G. Brewka, T. Eiter, M. Fink, "Nonmonotonic multicontext systems: A flexible approach for integrating heterogeneous knowledge sources," in Logic Programming, Knowledge Representation, and Nonmonotonic Reasoning - Essays Dedicated to Michael Gelfond on the Occasion of His 65th Birthday, vol. 6565 of Lecture Notes in Computer Science, 2011, pp. 233–258, Springer.
[74] G. Brewka, S. Ellmauthaler, J. Pührer, "Multi-context systems for reactive reasoning in dynamic environments," in ECAI 2014, Proc. of the 21st European Conf. on Artificial Intelligence, 2014, pp. 159–164, IJ- CAI/AAAI.
[75] P. Cabalar, S. Costantini, G. De Gasperis, A. Formisano, "Multi-context systems in dynamic environments," Ann. Math. Artif. Intell., vol. 86, no. 1-3, pp. 87–120, 2019, doi: 10.1007/s10472-019-09622-0.
[76] S. Costantini, G. De Gasperis, "Exchanging data and ontological definitions in multi-agent-contexts systems," in Proceedings of the RuleML 2015 Challenge, the Special Track on Rule-based Recommender Systems for the Web of Data, the Special Industry Track and the RuleML 2015 Doctoral Consortium hosted by the 9th International Web Rule Symposium (RuleML 2015), Berlin, Germany, August 2-5, 2015, vol. 1417 of CEUR Workshop Proceedings, 2015, CEUR-WS.org.
[77] S. Costantini, "Knowledge acquisition via non-monotonic reasoning in distributed heterogeneous environments," in 13th Int. Conf. on Logic Programming and Nonmonotonic Reasoning LPNMR 2013. Proc., vol. 9345 of Lecture Notes in Computer Science, 2015, pp. 228–241, Springer.

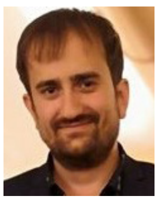


Stefania Costantini

Stefania Costantini (Female, married, two children, Daniele and Alice, born in 1993 and 1996) graduated with honors at the University of Pisa, Italy, in 1983. She worked at Italtel SIT (a telecommunications company) in Milan, Italy. In 1986 she became research associate and then (in 1990) Assistant Professor of Computer Science at the University of Milan, Italy. In 2001 she became Associate Professor at the University of L'Aquila (Italy) where, since 2005, she is Full Professor in Computer Science at the Department of Computer Science and Engineering and Mathematics (DISIM). She is the Head of the research group AAAI@AQ (Autonomous Agents and Artificial Intelligence at the University of L'Aquila). She has more than 150 publications. Her research interests are in (theory and practice of) Artificial Intelligence and Computational Logic, including Intelligent Software Agents and Multi-Agent Systems, Answer Set Programming, Non-Monotonic Reasoning, Knowledge Representation, Cognitive Robotics. She invented, defined and coordinated the first implementation of the DALI agent- oriented logic programming language. She served in the Program Committee of the main Conferences of her fields of interest, and she is a member of the Editorial Board of the journal Theory and Practice of Logic Programming (Cambridge). She is currently the President of the Italian Association of Computational Logic (GULP), and Member of the Board of the Italian Association for Artificial Intelligence (AIxIA).

Giovanni De Gasepris

Master Degree in Electronics Enginering cum Laude (1991) and Ph.D. in Electronics Engineering (1995) from the University of L'Aquila, Italy. He has been a Post Doctoral Fellow at University of Texas, M.D. Anderson Cancer Center in Houston, TX, USA (1995-1998). Technologist and freelance computer consultant in Italy (1998-2006). Contract lecturer of Computer Engineering at University of L'Aquila, Italy (2000-2006). Since 2007 he is an Assistant Professor at Department of Information Engineering and Computer Science and Mathematics, at the University of L'Aquila, Italy. He is teaching since 2015 the course of ”Intelligent Systems and Robotics Laboratory” at the Master Degree in Computer Science, and the course ”Virtual Reality and Archeomatics” at the Master Degree in Cultural Heritage at the Human Science Department. His current research interests are: Cognitive Robotics, Internet of Things, Natural Language Processing, Emotion Recognition, Virtual Reality. He is Core Developer and Coordinator for the development of many open source software packages through the research group github organization (https://github.com/AAAI-DISIM-UNIVAQ), among which the DALI logic multi agent system framework and its ASP_DALI extension. He is member of the Italian Association for Artificial Intelligence (AIxIA) since 2009.



Lorenzo De Lauretis

born in L'Aquila in 1991, graduated in Computer Science at Università degli studi dell'Aquila in 2016. He got his Master Degree cum Laude in Computer Science in 2018 at Università degli studi dell'Aquila. He became a PhD student in October 2018.

Improving Asynchronous Interview Interaction with Follow-up Question Generation

Pooja Rao S B, Manish Agnihotri, Dinesh Babu Jayagopi *

International Institute of Information Technology, Bangalore (India)

Received 28 October 2020 | Accepted 23 December 2020 | Published 24 February 2021

Abstract

The user experience of an asynchronous video interview system, conventionally is not reciprocal or conversational. Interview applicants expect that, like a typical face-to-face interview, they are innate and coherent. We posit that the planned adoption of limited probing through follow-up questions is an important step towards improving the interaction. We propose a follow-up question generation model (*followQG*) capable of generating relevant and diverse follow-up questions based on the previously asked questions, and their answers. We implement a 3D virtual interviewing system, *Maya*, with capability of follow-up question generation. Existing asynchronous interviewing systems are not dynamic with scripted and repetitive questions. In comparison, *Maya* responds with relevant follow-up questions, a largely unexplored feature of virtual interview systems. We take advantage of the implicit knowledge from deep pre-trained language models to generate rich and varied natural language follow-up questions. Empirical results suggest that *followQG* generates questions that humans rate as high quality, achieving 77% relevance. A comparison with strong baselines of neural network and rule-based systems show that it produces better quality questions. The corpus used for fine-tuning is made publicly available.

Keywords

Asynchronous Video Interview, Follow-up Question Generation, Language Model, Question Generation, Virtual Conversational Agent.

DOI: 10.9781/ijimai.2021.02.010

I. Introduction

Conventional hiring process is laden with challenges. The amount of time required to hire candidates, lack of inter viewers, expensive labour costs, scheduling conflicts are a few ex-amples. Traditionally, at the employer's location, candidates take tests in a calm, distraction-free environment chosen by the employers where their presence is required. It includes various costs like scheduling, infrastructure, workspace and many more. To reduce these costs and challenges, recruiters are heeding to futuris-tic choices like social recruitment, online assessments, and video interviews [1]. Organisations are adopting innovative methods like social media, proctored assessments, asynchronous or one-way interviews.

Online interviews for hiring are conducted using computer-mediated communication like instant messaging, email or video. Online interviews can be of the types synchronous, near-synchronous and asynchronous. [2] Synchronous interviews happen in real-time with simultaneous communication exchange. Near-synchronous interviews are near-immediate, on-going post and response. In the case of asynchronous interviews, there is a time-lapse between the communicating parties. These also called one-way interviews, are usually conducted via online video interviews using internet-enabled igital devices. The candidates can take the interview whenever and wherever it is convenient for them.

Asynchronous video interviews (AVI) have evolved as a tool to conduct first round of screening as well as interview coaching. These are gaining increased attention due to its scalability and ease of use. Many automatic talent assessment solutions like Talview[1], Hirevue[2], Sonru[3] offer asynchronous, ubiquitous interviewing and screening. Automatic interview and coaching systems simulate the behaviour of an interviewer helping interviewees with mock interviews. The feasibility and ease of automatic assessment of the AVIs when compared to in-person interviews [3] is persuading the wide spread use of the system. They provide the advantage of taking the test at candidate's convenience and facilitate efficient screening with minimal human intervention.

Conventional AVIs adopt structured interviews which are the standardized way of interviewing job candidates. Candidates are asked same questions in the same order with limited or no prompting and follow-up, and no elaboration on questions [4]. However, with large scale implementation of these systems, it will inevitably become predictable and uninteresting for recruiters and candidates alike. The dialogue will be monotonous and far from human-human interviews. Therefore, seeking the right balance between structure and probing is the imperative next step.

* Corresponding author.

E-mail addresses: pooja.rao@iiitb.org (P. Rao S B), manish.agnihotri@ii-itb.ac.in (M. Agnihotri), jdinesh@iiitb.ac.in (D. B. Jayagopi).

[1] www.talview.com

[2] www.hirevue.com

[3] www.sonru.com

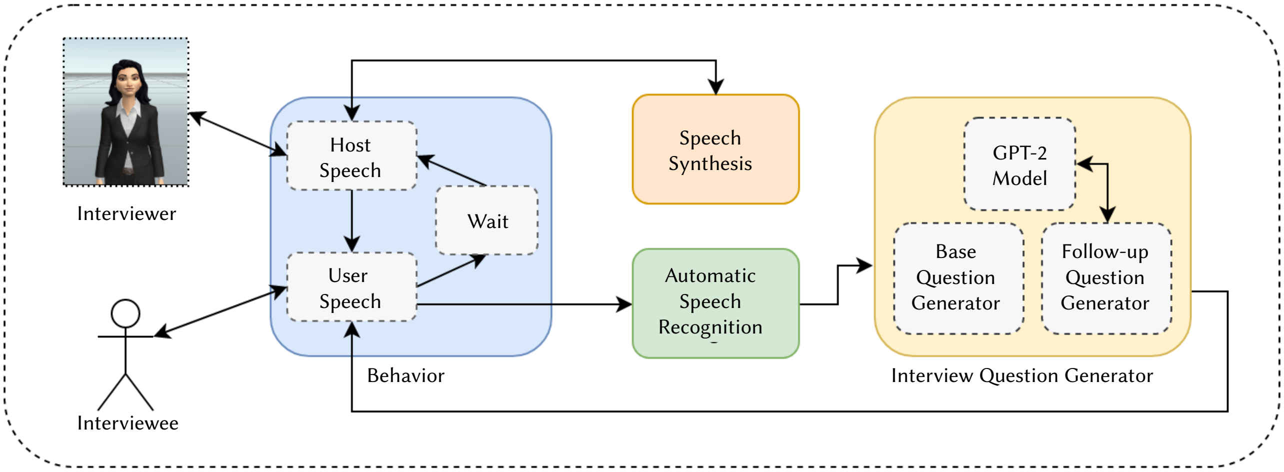


Fig. 1. Framework of Interviewing System.

An effective interviewing agent is one that has qualities similar to that of an effective human interviewer [5]. The ability to understand and respond to a conversation partner properly is one such inherent ability of human interviewer. This ability, also known as active listening, is shown to assist the progress of interviews [6]. Building a fully conversational interviewing agent with these abilities would be very challenging. A follow-up question in an interview can act as a proxy to such a human-like understanding of the answer in a crude way. It is crucial that an asynchronous system is conversational and acknowledges the candidates' response. We have enough evidence to suggest that the asynchronous interviewing systems are the need of the hour and the usage of such intelligent interviewing systems can only be scaled, if it's experience is interactive and personalized.

Levashina et al. [4] define follow-up question as the one that is intended to augment an inadequate or incomplete response provided by the applicant, or to seek additional or clarifying information. A relevant follow-up question not only improves the interaction between the interviewer and the interviewee but also makes it less predictable as the follow-up question is dynamic based on the interviewee's answer.

The methods for building a question generation model can be adapted for follow-up question generation, with one major difference. In the task of question generation, the question generated will already be answered in the input to the model. A follow-up question should not be already answered and should seek additional information not present in the input. Traditionally, the task of question generation has been tackled using hand-engineered features and crafting pattern-based templates and linguistic rules. More recently, with the advent of Deep Learning, question generation is treated as a sequence-to-sequence [7] problem with the reader sequentially parsing the input answer word by word and the generator producing the probabilities of the next word until the entire question is generated. This is a data intensive process and the availability of interview question answers with follow-ups is highly scarce. We address this challenge of data scarcity by leveraging the implicit knowledge from a pre-trained language model and adapting it with a small in-domain interview corpus.

In our work, we describe *Maya*[4], an interviewing virtual agent that is capable of generating follow up questions. This paper is an extended version of [8] published in IntelLang workshop of ECAI 2020. Our main contributions are as follows. First, we describe our follow-up question generation model - *followQG*, and each of the individual modules. Our contributing module is the use of a large-scale transformer language model to generate relevant and diverse follow up questions. Second, we benchmark the follow up generation output against other strong question generation/selection models, using human evaluation. Third, in order to show how well the system scales with regard to speech input vs text input, we perform experiments with manually transcribed text vs automatically transcribed spoken text, where an Automatic Speech Recognition (ASR) engine does the speech transcription. The results show the effectiveness of *Maya* even with speech input. Overall *Maya* works real-time and is able to hold an interactive interview with a candidate. Finally, we also investigate how robust is *Maya* with multiple follow up turns, and present qualitative results.

[4] The demo of the system can be found at – https://www.youtube.com/watch?v=gdPxdi82nV0

II. Related Work

We organise the related work in three categories: Research addressing language model pretraining (as our solution is based on neural network based text generation models), question generation task (of which follow up question generation is a special task), and agent based interviewing systems (which is exactly the task we are addressing).

A. Language Model Pretraining

Training the Deep Learning models from scratch, starting with random layer initializations, with large datasets taking a long time to converge is the de facto standard for tackling various NLP tasks. The effective breakthrough strategy to this is the greedy layer-wise training using an unsupervised learning criterion (pre-training) followed by tuning all parameters of the network on a global supervised cost function (fine-tuning) [9], [10]. Pre-training in an unsupervised fashion on vast quantities of text has resulted in state-of-the-art development on various tasks of natural language processing [11] [12]. These pre-training objectives are mostly variants of language modelling.

ULMFiT [13] is a transfer learning method for text classification tasks. A language model is pretrained on Wikipedia data and fine-tuned for a target task with a smaller amount of labelled in-domain data. Several works follow this fine-tuning approach and produce remarkable outcomes. ELMo [14] is a bidirectional language model that uses bi-LSTM networks to predict the next and previous tokens. OpenAI's GPT [12] trains huge text data in a unidirectional language model. BERT [11] is a masked language model trained with the next sentence prediction as an additional objective. On several downstream NLP functions, like the GLUE [16] benchmark, these models have achieved state-of-the-art results. Generative tasks such as end-to-end dialogue systems [17] and automated knowledge base completion [18], use pre-training with the GPT model, obtaining significant improvements over the models trained only with in-domain data. Both the works use the

TABLE I. A Comparison of Asynchronous Interview Systems. The Verbal Interaction in Maya Differs from Other Works With a Follow-up Question Mechanism as It Uses a Question Generation Model Rather Than Using Template-based Question Selection Method

System	Agent	Nonverbal Interaction	Verbal Interaction	Follow-up Q
Rao S B et al. [19]	Text Medium	No interaction	Fixed Script of Questions	No
SPECIES [20]	Embodied Agent	Head Movement and Facial Expressions	Template based	Yes
MACH [21]	Embodied Agent	Head Nodding and Smile Sharing	Fixed Script of Questions	No
TARDIS [22]	Embodied Agent	Body Motions, Gestures and Facial Expressions	Fixed Script of Questions	No
ERICA [23]	Robotic Agent	Head Movement, Gestures and Eye Gaze	Template based	Yes
Maya (Ours)	Embodied Agent	Gestures, Facial Expressions and Follow-up Question	Dynamic Question Generation	Yes

transformer language model GPT for initialization. Our work builds on this to develop a Follow-up Question Generation model.

B. Natural Language Question Generation

The goal of the Question Generation (QG) task is to automatically generate questions based on some form of text input [24]. This task became popular ever since the First Question Generation Shared Task Evaluation Challenge [25]. Recently, neural networks have enabled end-to-end training of question generation models influenced by the sequence-to-sequence (Seq2Seq) data-driven learning methods [7]. Serban et al. generate simple natural questions from structured triples - subject, relation, object using a neural system [26]. This has been successfully extended to unstructured data. Du et al. [27] generate question to test comprehension, using the encoder-decoder model with attention on the machine comprehension dataset SQuAD [28]. Wang et al. generate questions from educational content using an RNN-based encoder-decoder model, trained on SQuAD [29].

Follow-up question generation in interviews is a relatively new task, addressed first by Su et al. [30]. Instead of using a text generation model, they generate question patterns filling it up with words from a list. They adopt a pattern-based Seq2Seq model on a small interview corpus in Chinese. To create a word class table and turn all sentences in the corpus into patterns, they use a word clustering based process. In order to select a question-worthy sentence from the answer, they use a convolutional neural tensor network [31] and generate follow-up question patterns. These patterns are filled with words from the word class table to obtain potential follow-up questions. In a subsequent work, Su et al. [32] utilize the domain knowledge from ConceptNet to fill up relevant words in the follow-up template.

In contrast, we develop a follow-up question generation model utilizing knowledge from large-scale language model and a small corpus which does not involve pattern matching and template filling.

C. Agent-based Interviewing Systems

A recent trend in Interviewing Systems is the use of Intelligent Virtual Agents. Asynchronous Video Interviews (AVIs) are more common, where questions are posed by an interface. The use of intelligent virtual agents in AVIs allows for a more interactive and immersive experience than traditional voice and text-based systems [33], [34]. A job interview is aimed to analyze the hiring feasibility of an interviewee, while a training interview gives accurate feedback about their performance. While the initial works in AVIs were restricted to the skill assessment [35], [19], improving the interview experience has gained momentum. One standard approach is the usage of virtual agents as interviewers instead of textual prompts to conduct interviews [20]. This approach makes the interview experience more interactive.

In an early work, Nunamaker et al. introduced the usage of Embodied Conversational Agents (called SPECIES) in automated interviews [20]. One of their goals was to study the difference in perceptions with varying attributes of agent, and hence their work concerned agent design aspects.

Later, two coaching-based conversational agents were proposed - MACH [21] and TARDIS [22]. Both of them focus on skill assessment and non-verbal behavior analysis to improve the feedback to interviewees significantly, but the questions are taken from a small fixed pool of questions and do not take into account the interviewee's response. In these works, the virtual agents acknowledge the interviewee's answers to questions by smiling and nodding at a randomly generated frequency. Though this might make the agent credible, it can soon become superficial. Acknowledging the interviewee's answer by understanding the context and reacting appropriately through verbal means can be a better alternative to increase the interactivity and is the aim of our work.

Apart from the use of Virtual agent, a robotic agent (called ERICA) [23] was also built for spoken dialog. ERICA had the capabilities of human-like eye gaze, head movement and gestures, and a statement-response system which is response retrieval method based on pattern and focus token matching. Although the behavior synthesis is a notable improvement, it still lacks robustness in dialogue generation.

Automated Social Skills Trainer (ASST) [36] focused on Social Skills Training as their interview objective. The embodied agent in ASST is capable of head nodding, and eye blinking, and the dialogue is based on MMDAgent, a Japanese spoken dialogue system which selects an appropriate response using pattern matching.

While a lot has been done in automatic analysis of interviewee's response [19], [37] to improve the quality of the interview, not much has been done to make the interview more verbally interactive. All the previous works have either used a fixed script of questions or used a pattern matching based question selection (see Table I). We aim to improve the question generation system to make it more personal and response-based by generating relevant and grammatically correct follow-up questions.

III. Follow-up Question Generation - FollowQG

FollowQG is an adaptation framework for generating follow-up questions using language models by training it on an in-domain corpus of question, response and follow-up triplets. These data triplets help followQG to understand the structure of the question and the relationship between the triplets, and novel questions arise from the knowledge of the language model pre-training.

A. Task

The training samples consist of $\{q, r, f\}$ in natural language, where q is the interviewer question, r is the candidate response and f is the follow-up question. The task is to generate f given q and r as inputs.

B. Transformer Language Model

In this work, we use the transformer language model architecture, Generative Pre-trained Transformer (GPT-2) introduced in Radford et al. [38]. This uses an architecture similar to the decoder part of the original transformer encoder-decoder model of Vaswani et al. [39].

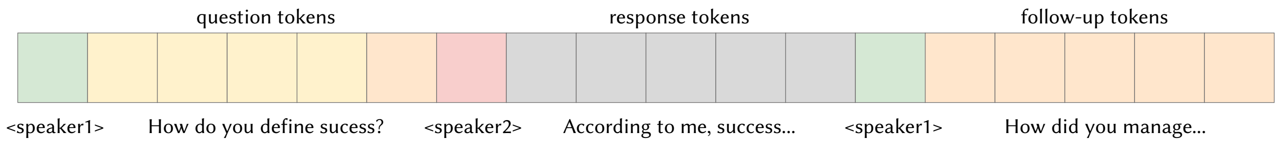


Fig. 2. Input representation for training Follow-up Question Generation model.

It uses several layers, each comprising two sub-layers. The multi-headed self-attention process is employed in the first followed by position-wise feed-forward layers to generate an output distribution over target vocabulary tokens. Our model is based on the recently published adaptation of GPT-2 from HuggingFace[5].

We initialise followQG with 12-layer decoder-only transformers with 12 heads of self-attention comprising of 768 dimensional states. Parameters are initialised to the smallest version open-sourced by Radford et al. 2019, GPT-2 small [38]. A WebText dataset containing the text of 45 million internet links is used to pre-train the GPT-2 model.

C. Dataset

We need the training samples of {*q, r, f*} triplets to train followQG. We employ the Rao S. B et al. [19] asynchronous interview dataset. This dataset consists of behavioural interviews of university students, referred to as the Asynchronous Video Interview dataset-AVI dataset and Asynchronous Written Interview dataset-AWI dataset, respectively, through asynchronous video and written media. Using AWI dataset interview excerpts, we perform a restricted crowd-sourcing to obtain follow-up questions. We advise the volunteers to compose a follow-up question based on the posed interviewer question snippet and the answer of the applicant. An instruction video of the same can be found here[6]. Therefore, with more than 1000 samples, each sample containing the triplet of a question, response and a follow-up, we obtain a follow-up question dataset. Some samples from the dataset can be seen in Table II. You can find the complete dataset here[7].

TABLE II. Examples of Human Written Follow-up Questions from the Dataset

Q: Are you a self-motivator? **A**: Absolutely. For me, internal motivation works far more than external motivation ever could. Yes, at first, it may seem like I want some sort of external motivation, but the very end, my heart goes into the work assigned only when my own self pushes me to do it.
FQ: Awesome. How would you spread motivation to others?
Q: What matters to you more - job satisfaction or salary? **A**: According to me, job satisfaction covers all - the quality and quantity of work, salary, company environment and others. Yes, cooperation and adjustment is needed in a company.
FQ: Are you ready to work in a company who offers you 5 times more than what you get now but the area may not be very interesting to you?
Q: Have you worked with someone unprofessional, how did you handle it? **A**: During my B-Tech final semester internship, I have experienced unprofessional behavior. I did not do anything to show my displeasure, instead, I kept behaving professionally. It didn't solve the issue, but did reduce the magnitude to some extent.
FQ: Do you think not showing the displeasure is the only way to tackle that situation?

[5] https://github.com/huggingface/transformers

[6] https://youtu.be/KbHF7_kMaA8

[7] https://ms-by-research-thesis.s3.amazonaws.com/followMLdata.xlsx

D. Fine-tuning

Using the dataset mentioned above, we fine-tune the GPT-2 language model. For training, 80% of the data is used and the rest is used for validation. Model input consists of tokens each from sequence of {*q, r, f*} concatenated and embedded in order. For this sequence, a set of input embeddings is constructed. The word and position embeddings are derived from the GPT-2 model learnt during the pre-training phase. To indicate whether the token belongs to the question, answer or the follow-up, we use an additional set of embeddings, speaker embeddings. Fig. 2 illustrates how the tokens in {*q, r, f*} are organised to form the speaker embeddings. These embeddings are learnt during the fine-tuning phase. The input to the model is the sum of all three forms for each token— word, position and speaker embedding.

1. Multi-task Objective

Following [17], [11], the fine-tuning is done by optimizing two loss functions – a language modelling loss, and a next-question classification loss. We use a multi-task objective where the total loss is the weighted sum of two losses. The language modelling loss is the commonly used cross-entropy loss. The last hidden state of the self-attention model is fed into a softmax layer over all the tokens in the vocabulary to obtain next token probabilities. These probabilities are then scored using the cross-entropy loss where the human written follow-up question tokens are used as labels.

With randomly sampled questions from a pool of 200 (same as those used in Section V), serving as distractors, we train a next-question classifier to recognize the correct next question. This trains the model to acquire a sense of sentence ordering. The classifier is a linear layer that applies a linear transformation to the last hidden state of self-attention model. A softmax layer obtains the classification probabilities using the computed values. We then apply a cross-entropy loss to correctly classify the follow-up question. We use $n = 2$ as the number of choices for classification making it a binary classification task. The parameters of the transformer language model and the next-question classifier layer are fine-tuned jointly to maximize the log-probability of the correct label.

2. Decoding Details

We use the top-k random sampling strategy for decoding [40]. The top-k probabilities of most likely next word is given at each point. The decoder randomly samples a word from these k candidates. Here, k is a hyperparameter determined experimentally to be k=10.

E. Results

We report the results of the follow-up question generation model in terms of perplexity [41]. The classification accuracy of the next-question task is also recorded. Perplexity is typically used to measure the quality of language models. It indicates how well the next word is correctly predicted by the model. Our model obtains an average validation perplexity of 20.6 and average validation accuracy of 63.1%. Considering the small size of the in-domain dataset used for fine-tuning, these values can be considered reasonable. The questions generated are novel and relevant and are not present in human written follow-up questions as we'll see in Section D.

IV. Experiments

In this section, we demonstrate the effectiveness of followQG with quantitative and qualitative experiments. First, we quantitatively do a relative comparison of followQG with strong baselines through human evaluation. These baselines loosely mimic the different interviewing agents discussed in Section C. We then individually evaluate followQG with human annotations on relevance and grammar. We also investigate the robustness of followQG model to errors in speech recognition. Finally, we qualitatively validate the single and multiple follow-up questions on same interview question-answer pairs.

A. Baselines

We compare followQG with two strong baselines. One is a rule-based system based on similarity measure and other is the reader-generator based QG-Net model [29]. We choose these baselines as representatives of the existing asynchronous video interview systems (Table I) which either pose questions from a fixed set or template based generation.

1. Similarity-based Question Selector

This is a rule-based question selector with a rule on cosine similarity to select questions from a fixed pool of 200 behavioural questions (same as the ones used in Section V).The original interview questions and the pool of questions are all represented using GloVe vectors [42]. We calculate the cosine similarity metric between the original interview question and each of the questions from the pool. We take into account the top-10 questions with highest similarity values and randomly select one to be the follow-up question. The selected follow-ups are based only on the question and not the candidate response. This question selector loosely mimics the different rule-based selectors in the existing systems which pose a fixed question and do not take the response into account.

2. QG-Net

With a context reader and question generator, QG-net is a Seq2Seq model. The context reader is a bi-LSTM network that processes and transforms each word into a fix-sized representation in the input context. The question generator is a uni-directional LSTM which generates the question word-by-word incorporating pointer network [43] into the vocabulary of the generator. This model design enables the generator to output questions that focus on specific parts of input text. The *focus tokens* are encoded as an additional feature with each input word, using one-hot encoding to indicate if the word is a focus token. QG-Net is trained on SQuAD dataset consisting of context, query and span of answer tokens within the context. These span of answer tokens are used by QG-Net as focus tokens. For a detailed description, we refer the readers to the original paper . QG-Net effectively adapts a general purpose question generation model trained on SQuAD to generate questions from educational content, addressing the problem of insufficient training data. We therefore select this as our baseline model of the neural network. The candidate's response, in our case, is the context and the question to be generated is the follow-up question.

The interview question-answer pairs have to undergo preparatory techniques like finding focus of the answer and extractive summarization to make the input format compatible with QG-Net model. QG-Net model trained on SQuAD dataset released by Wang et al. [29] is our second baseline.

Finding Focus of the Answer QG-net uses a binary valued indicator to indicate whether a word in context is important to generate a question, regarded as *focus tokens*. To automatically find these tokens in candidate responses, we employ a simple technique similar to Hu et al., [44]. In interview question (Q) and response (A), there are overlapping tokens seen as topics exchanged between the interviewer and candidate, that can be considered as focus tokens.

After removal of the stop words, A and Q are represented as a sequence of tokens $[a_1, .., a_n]$ and $[q_1, .., q_m]$ respectively. We consider all the tokens in A as candidates for focus tokens and all the tokens in Q as voters polling for the candidates. GloVe [42] vectors are used to represent tokens from Q and A. The ith answer token a_i gets a cumulative score S_i from all the tokens in the question calculated as

$$S_i = \sum_{j=1}^{m} p_{ij}.sim(a_i, q_j)$$

$$p_{ij} = \begin{cases} 1, & sim(a_i, q_j) > \lambda \\ 0, & \text{otherwise} \end{cases}$$

where $sim(a_i, q_j)$ is the cosine similarity between a_i and q_i. If the averaged S_i is above a certain threshold, a_i is included in the *focus*. This process is repeated for every answer token. We allow non-contiguous and multiple focus tokens which aid in the generation of distinct follow-ups.

Extractive Summarisation The input to the QG-Net model should be a sentence worthy of a follow-up representing the answer. To find this representative sentence, we employ a simple extractive summarization technique on the answer. The aim is to iteratively identify similar sentences in the answer using the focus tokens of those sentences and consider the most similar sentence as the summary sentence. We use the method described above to find the focus of each sentence. We then compare the focus of each sentence with the focus of other sentences using the cosine similarity measure.

R and S are two sentences from the candidate response with their focus tokens represented as $[fr_1, ..., fr_p]$ and $[fs_1, ..., fs_q]$ respectively. The cumulative score for each focus token of R is calculated as

$$W_i = \sum_{j=1}^{q} p_{ij}.sim(fr_i, fs_j)$$

$$N = \sum_{i=1}^{p} W_i$$

where p_{ij} is the indicative variable same as described above. If N crosses a certain percentage of the mean length of two sentences R and S, they are considered to be similar.

Once we have the pair(s) of similar sentences, we choose the one with more information content (more number of focus tokens) as the summary sentence. If more than one pair of sentences are similar to each other, *S* (pre-determined) number of sentences with the highest frequency of similar sentences is considered. The summary sentence along with the focus words is fed to the trained QG-Net model to generate questions.

B. Quantitative Human Evaluation

We obtain human annotations to evaluate the quality of the generated follow-up questions and compare them with the baselines[8]. Graduate students (non-native English speakers) with a background in Computer Science and Digital Society are the human annotators involved in this research. We sample 100 unseen question-answer pairs randomly from the AWI dataset and generate one follow-up question from all three models per QA pair – Similarity-based Question Selector, QG-Net question generation and followQG. We present the QA pair to three human annotators along with the follow-up questions produced by each model. Based on their choice, they are asked to rank the questions in the order of two metrics — the relevance of follow-up question to the given QA pair of interviews

[8] The customized web interface used for human annotation can be found here https://poorao.github.io/followML/

and their grammar. We define relevance as to how closely connected or appropriate the follow-up question is to the question-answer pair. We refer to this definition in the whole paper. The annotators are instructed to rank the ones with high relevance and high grammar as Rank 1, low relevance and low grammar as Rank 3 among the three choices. It is to be noted that these rankings are relative in measure.

For each follow-up question, we consider the statistical mode of the three annotators ranking. When the mode is not unique, i.e. when a different rank is selected by all three annotators (10% of the annotations), we resolve the disagreement by getting an additional set of rankings from an experienced interviewer. The statistical mode is then calculated uisng the extra annotations.

The findings can be seen in the Fig. 3. For each of the baseline and followQG, we calculate the frequency of the mode ranking for all three ranks. FQG model significantly outperforms (well beyond p=0.01 level) the other two models. With 54% of questions generated securing Rank 1, followQG is capable of high quality follow-up question generation. 34% of the questions generated by QG-Net obtain Rank 1. 50% of the questions from SQS secure Rank 2. It can be observed that the grammatically correct SQS selected questions are preferred second to the grammatically incorrect and somewhat relevant QG-Net model questions. We conclude that the FQG model, more frequently than the current baselines, produces valid and grammatically correct follow-up questions.

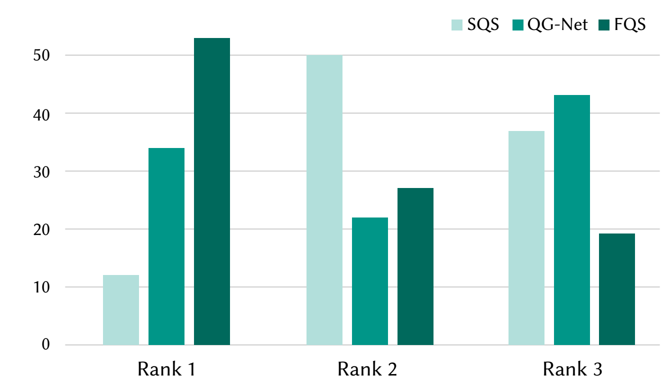


Fig. 3. Human ranking of preferred follow-up questions from followQG comparing with two other baseline models based on relevance and grammar. The bar indicates the frequency of rankings, indicating that the followQG model is the most preferred for highest ranking.

To improve our evaluation further, we obtain human ratings for followQG individually. Three human annotators measure the quality of the questions on a scale of 1-3, 1 being the lowest. The annotators are directed to annotate the follow-ups based on the grammar and relevance to the original interview QA pair. A score of 3 meant the follow-up question is grammatically acceptable and contextually relevant. Either a grammatically unacceptable question or not a follow-up question due to its low relevance meant a score of 1. And anything in between was considered a score 2.

The average scores from three annotators are considered for evaluation. Fig. 4 shows the statistics of the average ratings for the follow-up questions generated. 77% of the questions are scored ≥ 2. And 27% are rated ≥ 2.5. This shows that the followQG generates superior quality follow-up questions and are scored well by humans.

C. Robustness to Errors in Speech

Investigating the robustness of Follow-up Question Generator has two important motivations. First, the model is trained on human-written triplets of {*q, r, f*} whereas it will be inferred on the candidates's response obtained from ASR transcript in the virtual interviewing system. Secondly, ASR engines are generally prone to errors in recognition. Generation on misrecognized words may lead to an irrelevant question. Hence, analyzing how follow-up question generation varies for ASR transcripts when compared with human transcripts helps to investigate the robustness of followQG.

Average rating	Avg Rating on written QA pair	Avg Rating on manual transcripts	Avg Rating on automatic transcripts
1	2	0	4
1.3	9	11	15
1.67	12	21	18
2	23	22	22
2.3	27	21	21
2.6	20	17	20
3	7	11	3

Fig. 4. Frequency distribution of average human ratings on the quality of generated follow-up questions from followQG on a scale of 1-3 on the different types of question-answer pair inputs (hand-typed text, manually and automatically transcribed spoken text).

To this end, as they have manual transcripts of the interviews, we use the asynchronous interface-based video interview dataset from Rasipuram et al [3]. We randomly choose 103 interview QA pairs. Automatic transcripts for the same pair of 103 question answers are also obtained using the Web Speech API [45]. For each of these pairs, we generate a follow-up question. This brings us 206 triplets of questions, responses and follow-up questions, 103 for manual and automated transcripts each.

Three human annotators evaluate the quality of the question on a scale of 1-3, 1 being the lowest. Same instructions given in Section B are used. The annotators are displayed with the questions and answers from the manual transcripts and the follow-up questions generated on both manual and automatic transcripts to rate. We consider the average rating of the three annotators for evaluation.

The Fig. 4 indicates the number of average ratings for manual and automated transcript follow-up questions. 69% of the questions generated on manual transcripts and 64% of the questions generated on ASR transcripts get a score of ≥ 2. This means that, on both manual and automated transcripts, followQG produces reasonably high quality follow-up questions.

D. Qualitative Analysis

We present some examples generated by the followQG model on randomly sampled QA pairs in Table III.

Validation Examples Results show that the questions generated on the validation samples are relevant and grammatically correct but not same as the ground truth questions. We can see few examples in the first part of Table III where the generated follow-up questions are different from the ground truth follow-up questions. This may be due to the fact that followQG model leverages the implicit knowledge from the pre-trained language model to generate previously unseen questions different from ground-truth.

Unseen Examples The second part of the table shows the follow-up questions generated on completely unseen data. The model generalizes well on the unseen data and generates grammatically correct and suitable questions.

Unseen ASR Examples The third part of the table shows the questions generated on the text from the ASR. Despite the errors in transcription, the questions are relevant and appropriate. It is to be noted that the original base question is not exposed to the errors of the ASR and can be a reliable input to followQG.

TABLE III. Examples of the Follow-up Questions Generated on the Interview Snippets From the Validation Data, Unseen Data and Unseen Data From ASR

Validation Examples			
Question	Candidate Response	Generated Follow-up	Human Generated Follow-up
How would you be an asset to this organisation?	I can be an asset as I am patient while handling people and situations. I can participate and coordinate well with teams I get to work with. I can take in criticisms and work on to improve myself. I like to learn new things from anybody and will take necessary help whenever needed. I quite friendly with everyone.	How would you handle criticism?	Can criticism always be seen as a means for self-improvement?
Have you ever had to sell an idea to your co-workers? How did you do it?	I had to, when we were brainstorming ideas for a hacking session. I first tried to understand their concerns and requirements. Later I pitched my idea in accordance to that. I ensured that I understood and answered all their queries.	How did you handle the stress?	What were their first reactions?
What obstacles or difficulties have you ever faced in communicating your ideas?	I am not very expressive and can't put thoughts into words, so i have failed in communicating my ideas and explaining it or even getting work done.	How do you cope with a difficult situation?	And what steps have you taken to overcome this hardship?
What do people most often criticize about you?	I am not very social, an introvert and can't get along with people easily. I hvae been criticized about this.	What do you think about people who criticize you?	Have you ever tried to work on these drawbacks?

Unseen Examples		
Question	Candidate Response	Generated Follow-up
What are your salary expectations?	As of now, I haven't thought much about it. I'm more focused on learning the requirements for this position that I am applying for.	How much do you think you will be able to meet these requirements?
Are you a team player?	Definitely. While I will deny the fact that I can work independently with minimal supervision, I'm also one companion every leader would ever want to be in his team. Whatever the task is, I make sure it meets and exceeds what is expected of me. I also make it a point to reach out to teammates whenever needed.	Do you have any tips for team members?
Why do you want to work for this company?	I would be proud to work for a company like yours with such a long history of leadership in the industry. I have carried out web research and believe that the company's products and its future projections are very impressive.	Why do you think this company is successful?

Unseen Examples with ASR Transcriptions		
Question	Candidate Response	Generated Follow-up
Do you consider yourself successful?	I do consider myself successful because I have been able to achieve the things that I wanted for myself so ja in the respect yes I too feel them successful	Do you think you can achieve the things you wanted for yourself?
Tell me something about yourself.	I am a research assistant at the <name>... institute I work in the area of human interaction and machine learning	How did you learn to work with machine learning?
How do you think this job will help you fulfill your career goals?	when I apply for this job applied for this job and applied for this job many things which I like to do and I will do in the stop my career goals are just to have some work satisfaction and study whatever I want to so I think this job will help me do that and	How would you handle a situation where you have no work satisfaction?

TABLE IV. Comparison of Examples Generated from Three Different Models for Various Question-answer Pairs. SQS: Similarity-based Question Selector, QG-Net: QG-Net, FQG: Follow-up Question Generation

Q: What is the biggest challenge you have faced so far? **A**: The biggest challenge that i have faced so far is IIT- JEE examination after completing my class 12. I prepared for it for two years. Unfortunately I didn't cracked it. **SQS**: Where do you see yourself 3 years from now? **QG-Net**: did you think was cracked ?
FQG: What are the best lessons you have learned from this experience?
Q: Short self-presentation: Introduce yourself. **A**: Hello, I am X, a passionate software engineer. Till my 12th standard I studied at my native place itself. My strongest plus point according to me is that I dont trust anyone easily and negative point is that I cant resist anything wrong happening to anyone around me.... **SQS**: What has been the greatest disappointment in your life? **QG-Net**: what is you that you can't resist happening?
FQG: How do you manage your negative point?
Q: What has been the greatest disappointment in your life? **A**: I didn't get any high disappoint from my life. I was disappoint from myself when I didn't performed that well in GATE exam and not able to get admission in X, but I got admission in Y and now I feel that it was okay and not feel disappoint. **SQS**: What changes would you make if you came on board? **QG-Net**: what did you do in gate exam ?
FQG: How would you explain your failure in GATE exam?
Q: How would you be an asset to this organisation? **A**: I think I have a hunger to learn and solve a problem. For me, each problem is a puzzle which I need to solve and I will try till I finish it. If I don't know something, I will learn it and get the job done. My hardworking and passionate qualities can be asset to your company. **SQS**: How quickly do you adapt to new technology? **QG-Net**: what do you have a hunger to do ?
FQG: How do you think you can improve your skills?

Comparison Table IV compares the follow-up questions generated by the followQG with the two strong baselines, similarity-based question selector and QG-Net. It is seen that question generated by followQG is a grammatically correct and relevant follow-up in all cases. We can see that the QG-Net model fails the most in terms of grammar. FollowQG, despite the grammatical errors in the reply, always adheres to grammar. Grammatical adherence is irrelevant with SQS model as it selects from a pool of grammatically correct questions.

In terms of relevance, we find that the follow-up related to the question or answer or both is often generated by the followQG model. SQS generates follow-ups that are remotely connected or unrelated to the responses. Follow-ups by QG-Net model do not always adhere to relevance. It produces questions that are irrelevant or incomplete or grammatically incorrect. We should also notice that the QG-Net model produces already answered questions (e.g. Questions 2, 4).

Variety In addition to being high quality, the generated follow-up questions from followQG are diverse and novel. Majority of the questions are not present in the training set, showing that the model is capable of generating new questions. The questions generated can be modified versions of training samples or completely different from the training samples addressing unique details in the answer. For example in Table V, the last two questions address very specific details in the responses and give an impression to the candidate that the agent has understood and responded to the answer advancing the listening skills of the agent. Table V illustrates sample follow-up questions generated on the same question but different candidate responses.

TABLE V. Questions Generated by FollowQG for the Interview Question "Describe a decision You Made that Was a failure. What Happened and Why?" but with Different Candidate Responses. We Can See that the Follow-ups Are Novel and Diverse

Describe a decision you made that was a failure. What happened and why?
A: My initial plan for post-graduation was to join MS by research in IIT Madras. But I have failed in my first attempt due to some medical issues. So I was very determined to give it a second try. Then I took admission in IIIT-B for M.tech. I planned to prepare for the interview along with my course work during first semester. But unknowingly, I got into some bad company and ignored the preparation. I realised my mistake when I finally got the interview call again. It was too late by then. I got depressed because of this and I made a lot of effort to motivate myself again and do my best in my M.tech and work on my area of interest. If I achieve the goals I've set for the next year,hopefully I'll be pursuing my Ph.D too.
FQ: How did you cope with this situation?
A: During my final year B.E. project, we had decided to do a project on detection sign language using image processing and Machine Learning. I as a team lead, decided to use Neural Networks to be used as a classification mechanism for classifying images based on some features. After proceeding with the NN we found out that it was not feasible to implement it on a wearable device. Later we changed the decision from NN to SVM and achieved our goal.
FQ: What was your experience when you realized that you did not have the skills to implement a machine learning algorithm
A: When i passed my 12 standard i chose to take admission in one private engineering college rather to follow my family decision to take coaching for iits .I thought this was my failure because i thought all engineering colleges are not much different but later i realized that iits are way better than these colleges so on that very day i decided to qualify GATE exam to do my masters from one of the best college of India.
FQ: Why did you decide to take admission in one private engineering college rather than another?

To explore further, we investigate followQG by generating many follow-ups on the same interview question-answer pair. As we see from Table VI, followQG generates a variety of questions addressing different parts of the question and answer. Though example questions 1 and 3 focus on the "challenges in the organization", they are rephrased differently. The other example questions attend to other concepts related to "organization".

TABLE VI. Questions Generated by FollowQG for a Single Question-Answer Pair. We Can See the Variety in Follow-ups though the Question and Answer Input Is the Same

Explain how would you be an asset to this organisation?
A: My skills are a perfect match for the company's requirements. The company would definitely benefit from me as I have domain experience. I have a good idea about the company's goals. For me it's a place I can learn and be productive. I am young and skilled.
FQ: Do you think you will be able to overcome the challenges of the organisation? **FQ**: What would you learn from the company? **FQ**: Do you think that your skills are enough to be able to handle the challenges in the organization? **FQ**: How would you handle the pressure?

Multiple Follow-ups We further test followQG by examining its capability to generate multiple follow-up questions and hold a two or three turn interview. Since the model is trained only on a single pair of QA, the input to the model is only the previous question answer pair. Table VII shows a series of follow-up questions generated by followQG. The first question is a pre-determined interview question. This is followed by follow-up generated with the original question and candidate response. The next follow-up is generated based on the previous follow-up and answer and so on. We can note that questions generated still hold relevant to the overall theme of the interview i.e., about the teamwork.

TABLE VII. Multiple Follow-up Questions Generated by FollowQG as a Conversation Agent. The Input to the Model Is Only the Previous Question-answer Pair

Give an example of how you worked in team.
A: I have handled three teams together during my second semester of M.Tech for my course projects. I used to have complete discussion with the team before landing to a conclusion. When even I or someone got stuck at something we used to help each other to solve the issue. I managed the time to work flawlessly with all the teams.
FQ1: How do you manage team problems?
A: I motivate my team mates to complete the work and also in a perfect manner. It's important to be friendly with our team mates and I do the same. I divide the work into different parts and all the team members will share the equal work and I will be open to solve the doubts and get the work done.
FQ2: How do you motivate your teammates?
A: You can only motivate and inspire your team if they know what they are working towards and see the bigger picture. I will share the larger goal with the team and fire up the energy to keep motivated.
FQ3: Do you think you have enough energy?

V. Maya - Interactive Interviewing System

Our interactive interviewing system, *Maya*, consists of two main components – 3D Virtual Interviewer and Interview Question Generator. The overall framework of *Maya* can be seen in Fig. 1. The first component is an Amazon Sumerian [46] based 3D virtual interviewing agent which asks questions and collects the interviewee's responses. We have defined behavior for the interviewer as a state machine component which makes the scene dynamic and interactive.

A. Interviewer's Behaviour

The state machine component defined for the host (interviewer) consists of 4 states: Initialisation, Maya Response, User Response and Wait.

Maya Response, which executes the HostSpeech script defines the host response and initiates the Speech Component. After configuring the speech body and voice, it plays the audio. We use Amazon Polly text-to-speech toolkit [82] service to synthesize speech at runtime. The host response is either a greeting followed with the first question or is the follow-up question based on the interviewee's response of the previous question.

Upon successful execution of the host speech, the state changes from Maya Response to User Response. The User Response state executes the UserSpeech script, which takes the interviewee's response to the question asked and returns an appropriate follow-up question. It uses the Web Speech API [80] to get the transcript of the interviewee's response. The word error rate of this ASR engine is 45.7, calculated on 5 randomly chosen videos from the Asynchronous Interview dataset at the utterance level.

Once the transcript is collected, it is fed to our Interview Question Generator hosted on a server using an API call. The response of the API call is a follow-up question which is set as the host's next response.

B. Interview Question Generator

The Interview Question Generator component contains two modules which communicates with the 3D virtual interviewer namely, Base question selector and followQG. Base question selector selects a question randomly from 200 questions commonly asked in an HR interview. Next question is a follow-up question generated by followQG. This repeats for a fixed number of times. In our experiments, we limit the number of follow-up question to one. The next base question is selected after one follow-up question. Hence the follow-up question is based on single previous response from the candidate and not the history. We consider one follow-up question as a proxy to planned or controlled probing and try to improve the interactivity and listening skills of the asynchronous interviewing system.

VI. Discussion and Limitations

After analyzing the quantitative and qualitative results of our study, we now discuss some reflections and limitations. Traditionally, asynchronous interview media do not enable interaction. To address this interactivity attribute of the medium, we propose follow-up question generation enabling one level of probing. Since structured interviews are known to reduce different biasing factors [4], we limit the level of probing to one. This balances the structure of the interview as well as conversational flow between the system and candidate.

We use a relatively small in-domain corpus of interview question, answer and follow-up question to train our model. In spite of being small, the dataset helps the model understand the nuances of the concept of follow-up question. With the power of knowledge from the pre-trained language model, the questions are generated to be appropriate follow-ups.

The use of knowledge from the large-scale transformer language model induces external knowledge generating diverse questions adhering to grammar. The data samples help FQG to learn the question structure and the relation between the triplets, and the knowledge from the language model pre-training produces novel questions. This model can be seen as the one with improved performance when compared with two strong baseline models. The model always generates grammatically correct questions and the quality of questions is enhanced.

The task of follow-up question generation differs from question generation majorly in one dimension. In question generation, the input to the model already has the answer to the question to be generated. Whereas in follow-up question generation, the question to be generated must seek more information or related information that is not present in the input to the model. The QG-Net baseline model is originally trained to address question generation and not follow-up generation. This drawback is evident in the questions generated from QG-Net which are already answered. The similarity-based question selector relies solely on the similarity techniques and is constrained to the pre-defined number of questions in the pool. We try to address these gaps in the existing asynchronous agent-based interviewing systems with followQG.

While our evaluations are encouraging, there are certain limitations to our system. Currently, we use a dataset that is small in size to fine-tune the model. Even though the model finetuned on a small dataset is producing appropriate follow-up questions, increasing the size of this dataset might improve the quality of the questions further. All the annotators and volunteers for crowdsourcing in this study are non-native English speakers. Obtaining the data or annotations from native English speakers can add variety to the mix.

A. Future Work

Below we discuss some of the future directions to our system.

1. Considering the History of the Interview Conversation

Currently, *Maya* considers only the previous question and answer to generate a follow-up. Taking into account the whole context of the interview and remembering the information from the previous answers will be helpful in improving the overall experience with the system. An immediate next step would be to train followQG with the history of the interview conversation to generate the next follow-up question.

2. Interpreting and Understanding Deeper Meaning

Though our current system asks follow-up questions addressing specific details, it can be seen as understanding the gist of the candidate response. To generate more meaningful questions, the system must extract deeper concepts and relationships among them. Conditioning the questions on the background of the candidate (like information from resume), an external knowledge source to improve commonsense reasoning can be potential directions.

3. Interrelating Question Topics

In the current system, *Maya* asks each question independently while in a real world scenario the questions are interrelated. It is necessary to explicitly bring out these relations.

4. Building Empathetic Interviewing Agents

An important aspect of an engaging human conversation that receives relatively less focus is emotional understanding and empathy. There is a need to include emotion into the interviewing agents and respond in a way that acknowledges the feelings of the candidates. The next step towards that would be to train followQG explicitly with emotion labels and emotionally coloured words.

VII. Conclusion

We introduce *Maya*, a virtual agent-based interviewing system equipped with verbal interactivity from follow-up question generation. We leverage the implicit knowledge of a large scale transformer language model fine-tuned on follow-up questions dataset to generate relevant, novel and diverse questions based on the candidates' response in an interview. With availability of limited data, this approach scales as it uses external knowledge from a language model trained on a huge corpus. With human evaluation, we show that the questions generated are of good quality. We can also see that the FQG model is often robust to the errors of speech recognition. We restrict the generation of follow-up questions to one as existing research suggests the advantage of limited probing and follow-up. But the model is capable of generating multiple follow-up questions based on the previous response. These positive results point to future work in extending the approach to a variety of other types of interviewing agents not limited to behavioural domain, as well as investigating whether followQG can be trained on any other domain descriptive questions to generate follow-up questions.

Acknowledgment

This work was partially funded by SERB Young Scientist grant (Grant no: YSS2015001074) of Dr. Jayagopi, Karnataka government's MINRO grant and a grant from Accenture Technology Labs. We would like to thank all the participants who contributed for data collection.

References

[1] talview.com, "Understanding recruitment troubles and trends," 2016. [Online]. Available: https://info.talview.com/understanding-recruitment-troubles-trends-research-2016.

[2] J. E. Salmons, *Qualitative Online Interviews: Strategies, Design, and Skills. Thousand Oaks*, CA, USA: Sage Publications, Inc., 2nd ed., 2014.

[3] S. Rasipuram, S. B. P. Rao, D. B. Jayagopi, "Asynchronous video interviews vs. face-to-face interviews for communication skill measurement: A systematic study," in *Proceedings of the 18th ACM International Conference on Multimodal Interaction*, ICMI '16, New York, NY, USA, 2016, pp. 370–377, ACM.

[4] J. Levashina, C. J. Hartwell, F. P. Morgeson, M. A. Campion, "The structured employment interview: Narrative and quantitative review of the research literature," *Personnel Psychology*, vol. 67, no. 1, pp. 241–293, 2014.

[5] J. Li, M. X. Zhou, H. Yang, G. Mark, "Confiding in and listening to virtual agents: The effect of personality," in *Proceedings of the 22nd International Conference on Intelligent User Interfaces*, IUI '17, New York, NY, USA, 2017, p. 275–286, Association for Computing Machinery.

[6] C. R. Rogers, R. E. Farson, University of Chicago., Industrial Relations Center., *Active listening*. Chicago: Industrial Relations Center, the University of Chicago, 1957.

[7] I. Sutskever, O. Vinyals, Q. V. Le, "Sequence to sequence learning with neural networks," in *Advances in neural information processing systems*, 2014, pp. 3104–3112.

[8] S. B. P. Rao, M. Agnihotri, D. B. Jayagopi, "Automatic follow-up question generation for asynchronous interviews," in *Proceedings of the 1st Workshop on Intelligent Information Processing and Natural Language Generation, ECAI*, 2020.

[9] D. Erhan, Y. Bengio, A. Courville, P.-A. Manzagol, P. Vincent, S. Bengio, "Why does unsupervised pre-training help deep learning?," *J. Mach. Learn. Res.*, vol. 11, p. 625–660, Mar. 2010.

[10] H. Larochelle, Y. Bengio, J. Louradour, P. Lamblin, "Exploring strategies for training deep neural networks," *Journal of Machine Learning Research*, vol. 10, no. 1, pp. 1–40, 2009.

[11] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *CoRR*, 2018.

[12] A. Radford, K. Narasimhan, T. Salimans, I. Sutskever, "Improving language understanding by generative pre-training," 2018.

[13] J. Howard, S. Ruder, "Universal language model fine-tuning for text classification," in *ACL*, 2018.

[14] M. E. Peters, M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee, L. S. Zettlemoyer, "Deep contextualized word representations," *ArXiv*, 2018.

[15] Z. Huang, W. L. Xu, K. Yu, "Bidirectional lstm-crf models for sequence tagging," *ArXiv*, 2015.

[16] A. Wang, A. Singh, J. Michael, F. Hill, O. Levy, S. R. Bowman, "Glue: A multi-task benchmark and analysis platform for natural language understanding," in *BlackboxNLP@EMNLP*, 2018.

[17] T. Wolf, V. Sanh, J. Chaumond, C. Delangue, "Transfer-transfo: A transfer learning approach for neural network based conversational agents," *ArXiv*, 2019.

[18] A. Bosselut, H. Rashkin, M. Sap, C. Malaviya, A. Çelikyil-maz, Y. Choi, "Comet: Commonsense transformers for au-tomatic knowledge graph construction," *ArXiv*, 2019.

[19] S. B. P. Rao, S. Rasipuram, R. Das, D. B. Jayagopi, "Automatic assessment of communication skill in non-conventional interview settings: a comparative study," in *Proceedings of the 19th ACM International Conference on Multimodal Interaction*, 2017, pp. 221–229, ACM.

[20] J. F. Nunamaker, D. C. Derrick, A. C. Elkins, J. K. Burgoon, M. W. Patton, "Embodied conversational agent-based kiosk for automated interviewing," *Journal of Management Infor-mation Systems*, vol. 28, no. 1, pp. 17–48, 2011.

[21] M. E. Hoque, M. Courgeon, J.-C. Martin, B. Mutlu, R. W. Picard, "Mach: My automated conversation coach," in *Proceedings of the 2013 ACM international joint conference on Pervasive and ubiquitous computing*, 2013, pp. 697–706, ACM.

[22] K. Anderson, E. André, T. Baur, S. Bernardini, M. Chollet, E. Chryssafidou, I. Damian, C. Ennis, A. Egges, P. Gebhard, et al., "The tardis framework: intelligent virtual agents for social coaching in job interviews," in

International Conference on Advances in Computer Entertainment Technology, 2013, pp. 476–491, Springer.
[23] T. Kawahara, "Spoken dialogue system for a human-like conversational robot erica," in *International Workshop Spoken Dialogue Systems*, 2018.
[24] V. Rus, A. C. Graesser, "The question generation shared task and evaluation challenge," in The *University of Memphis. National Science Foundation*, 2009.
[25] V. Rus, W. Brendan, P. Piwek, M. Lintean, S. Stoyanchev, C. Moldovan, "The question generation shared task and evaluation challenge," in *The University of Memphis. Na-tional Science Foundation*, 2009.
[26] I. Serban, A. García-Durán, Çaglar Gülçehre, S. Ahn, A. P. S. Chandar, A. C. Courville, Y. Bengio, "Generating factoid questions with recurrent neural networks: The 30m factoid question-answer corpus," *CoRR*, 2016.
[27] X. Du, J. Shao, C. Cardie, "Learning to ask: Neural question generation for reading comprehension," in *ACL*, 2017.
[28] P. Rajpurkar, J. Zhang, K. Lopyrev, P. S. Liang, "Squad: 100, 000+ questions for machine comprehension of text," in *EMNLP*, 2016.
[29] Z. Wang, A. S. Lan, W. Nie, A. E. Waters, P. J. Grimaldi, R. G. Baraniuk, "Qg-net: a data-driven question generation model for educational content," in *L@S*, 2018.
[30] M.-H. Su, C.-H. Wu, K.-Y. Huang, Q.-B. Hong, H.-H. Huang, "Follow-up question generation using pattern-based seq2seq with a small corpus for interview coaching," in *Proc. Interspeech 2018*, 2018, pp. 1006–1010.
[31] X. Qiu, X. Huang, "Convolutional neural tensor network architecture for community-based question answering," in *IJCAI*, 2015.
[32] M.-H. Su, C.-H. Wu, Y. Chang, "Follow-up question gen-eration using neural tensor network-based domain ontol-ogy population in an interview coaching system," in *INTER-SPEECH*, 2019.
[33] W. Swartout, R. Artstein, E. Forbell, S. Foutz, H. Lane, B. Lange, J. Morie, D. Noren, S. Rizzo, D. Traum, "Virtual humans for learning," *AI Magazine*, vol. 34, pp. 13–30, 1 2013, doi: 10.1609/aimag.v34i4.2487.
[34] R. López-Cózar, Z. Callejas, D. Griol, J. F. Quesada, "Review of spoken dialogue systems," *Loquens*, vol. 1, no. 2, p. 012, 2014.
[35] L. S. Nguyen, D. Frauendorfer, M. S. Mast, D. Gatica-Perez, "Hire me: Computational inference of hirability in employment interviews based on nonverbal behavior," *IEEE transactions on multimedia*.
[36] H. Tanaka, S. Sakti, G. Neubig, T. Toda, H. Negoro, H. Iwasaka, S. Nakamura, "Automated social skills trainer," in *Proceedings of the 20th International Conference on Intelli-gent User Interfaces*, 2015, pp. 17–27, ACM.
[37] L. Hemamou, G. Felhi, V. Vandenbussche, J.-C. Martin, C. Clavel, "Hirenet: a hierarchical attention model for the automatic analysis of asynchronous video job interviews," 2019.
[38] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever, "Language models are unsupervised multitask learners," 2019.
[39] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, I. Polosukhin, "Attention is all you need," in *NIPS*, 2017.
[40] A. Fan, M. Lewis, Y. Dauphin, "Hierarchical neural story generation," in *ACL*, 2018.
[41] Y. Bengio, R. Ducharme, P. Vincent, C. Jauvin, "A neural probabilistic language model," *Journal of machine learning research*, vol. 3, no. Feb, pp. 1137–1155, 2003.
[42] J. Pennington, R. Socher, C. D. Manning, "Glove: Global vectors for word representation," in *EMNLP*, 2014.
[43] A. See, P. J. Liu, C. D. Manning, "Get to the point: Summarization with pointer-generator networks," in *ACL*, 2017.
[44] W. Hu, B. Liu, J. Ma, D. Zhao, R. Yan., "Aspect-based question generation," in *ICLR Workshop*, 2018.
[45] G. Shires, "Web speech api: Draft community group report," July 2019. [Online]. Available: https://w3c.github.io/speech-api/, [Online; posted 17-July-2019].
[46] T. Walker, "Presenting amazon sumerian: An easy way to create vr, ar, and 3d experiences," November 2017. [Online]. Available: https://aws.amazon.com/blogs/aws/launch-presenting-amazon-sumerian/, [Online; posted 27-November-2017].

Pooja Rao S B

Pooja Rao S B obtained her Master's by Research degree from IIIT Bangalore in 2019. She is currently a doctoral student at University of Lausanne. Her current research interests lie in the intersection of Human-Computer Interaction and Machine Learning with a focus on designing and building intelligent systems capable of multimodal processing, natural language processing, understanding and generation.

Manish Agnihotri

Manish Agnihotri obtained his Bachelor of Technology degree from MIT Manipal in 2019. He is currently a Machine Learning Engineer at Merlin AI, Zycus. His research interests lie in the areas of machine learning and affective computing with a focus on natural language processing and multimedia analysis.

Dinesh Babu Jayagopi

Dr. Dinesh Babu Jayagopi is an Associate Professor at IIIT Bangalore since Dec 2013, where he heads the Multimodal Perception Lab. His research interests are in Audio-Visual Signal Processing, Applied Machine Learning, and Social Computing. He obtained his doctorate from Ecole Polytechnic Federale Lausanne (EPFL), Switzerland, beginning of 2011. He received the Outstanding paper award in the International Conference on Multimodal Interaction (ICMI), 2012, Idiap PhD student research award for the year 2009. Subsequently, his research papers with students has received Best paper awards and nominations. He has also received funding from several agencies including DST, CAIR(DRDO) and Accenture.

Neural Scoring of Logical Inferences from Data using Feedback

Allmin Susaiyah[1]*, Aki Härmä[2]*, Ehud Reiter[3], Milan Petković[1]

[1] Eindhoven University of Technology, Eindhoven (Netherlands)
[2] Philips Research, Eindhoven (Netherlands)
[3] University of Aberdeen, Aberdeen (Scotland)

Received 16 October 2020 | Accepted 31 January 2021 | Published 15 February 2021

LA UNIVERSIDAD EN INTERNET

Abstract

Insights derived from wearable sensors in smartwatches or sleep trackers can help users in approaching their healthy lifestyle goals. These insights should indicate significant inferences from user behaviour and their generation should adapt automatically to the preferences and goals of the user. In this paper, we propose a neural network model that generates personalised lifestyle insights based on a model of their significance, and feedback from the user. Simulated analysis of our model shows its ability to assign high scores to a) insights with statistically significant behaviour patterns and b) topics related to simple or complex user preferences at any given time. We believe that the proposed neural networks model could be adapted for any application that needs user feedback to score logical inferences from data.

Keywords

Artificial Intelligence, Feedback Learning, Neural Network, Self-supervised Learning, Transfer Learning, Logical Inference, Natural Language Generation, Statistical Learning.

DOI: 10.9781/ijimai.2021.02.004

I. Introduction

Technological advancements in this century have led to a rise in the number of applications that claim to improve human lifestyle. A few important examples of these are popular health, diet and fitness mobile applications that have stormed the market. These applications obtain data from activity trackers or loggers and play the role of an artificial health or fitness agent by generating actionable insights of users' behaviour [1], [2]. For this purpose, it is desired that insights should be valid, represent a significant pattern of the user behaviour and should align with their interests. Insights can be of different levels of complexity. An example of a simple count-based absolute insight is: "You need to take 100 more steps to reach your daily goal!". Whereas, a more complex comparison based insight is "You sleep less when the room temperature is above 20 degrees than when it is lower.". Henceforth, in this paper, we consider only the comparative insights that are more complex and challenging. In general, these insights talk about a measure in two contexts, for example, by stating that a measure X is larger in context A than in context B, see [3]. This statement requires the test for statistical significance on two distributions, one from each of the contexts that are being compared.

For the purpose of testing the statistical significance of comparative insights, parametric and nonparametric significance tests have been widely used. However, there has been no specific technique to understand user interests in an intuitive manner. The biggest challenge in performing this is the very nature of user interests: they keep changing. Hence, a highly flexible model is required for this purpose. The artificial neural network (ANN) model, commonly described as a universal function approximator has shown great ability to learn, unlearn and transfer knowledge from one domain to another. Additionally, the ability of ANNs to learn multiple input characteristics encourages us to model multiple domains at once, such as the statistical significance domain and *interestingness* domain. This makes ANNs to be a favourable choice to model user preference.

* Corresponding author.

E-mail addresses: a.p.s.susaiyah@tue.nl (A. Susaiyah), aki.harma@philips.com (A. Härmä).

Insight generating systems shouldn't produce inferences that may harm the goal of the application. We can best guarantee this, by a system where all texts are selected from a pre-generated and manually curated collection of *validated insight candidates*. This is similar to the PSVI method introduced in [3]. These validated insights form the base for the rest of this paper. The statistical significance domain considered in this paper corresponds to a well-known nonparametric significance test, namely, the Kolmogorov-Smirnov (KS) test. The interestingness domain incorporates how much a user is interested in knowing about a particular comparative insight.

In this work, firstly, we develop an insight generation system that generates validated insights using a behaviour insight mining pipeline. Secondly, we train a self-supervised neural network that can upscale and replace traditional nonparametric tests (with 92% accuracy at 5% alpha). Lastly, we show how this ANN can also be used to learn user preference using an interactive learning strategy. For this, we use a evaluate using a single user-preference scenario and multi user-preference scenario.

The characteristics of insights that are considered by our model are essential for highly scalable behaviour insight mining (BIM) systems.

Applications of this can be in fitness coaching, office behaviour [4], behaviour change support systems [5], [6], business insight mining systems [3], and other relevant systems.

The structure of this document is as follows: section II gives a brief background about insight with examples, section III provides a in-depth explanation of how we developed the neural network architecture and how the online learning system is implemented. The results of our ANN and the simulated user scenarios are covered in section IV, discussion is included in section V while section VI presents the conclusions.

II. Background

In this section, we provide provide more context to the concepts discussed in this paper.

A. Desirable Characteristics of Insights

Based on recent literature, an insight should have the following characteristics, namely, statistical significance [3], [7], interestingness or personal preferences [3], [8]–[11], Causal confidence [10], surprisingness [8], actionability or usefulness [8], [9], syntactic constrains [7], presentability [11] timely delivery [11], and understandability [9]. Among these, the most essential characteristics are statistical validity and interestingness.

B. Types of Insights

1. Generic insight: These are insights that talk about a rather common or scientific phenomenon. These are not grounded on the user's behaviour. For example: Excessive caffeine consumption can lead to interrupted sleep as can ingesting caffeine too late in the day.
2. Personalised (Manual/Automated) insight [12]: These are insights that are tailored to the user either by a human-in-loop or by an algorithm.
    - Absolute insights or simple insights: These insights talk about user behaviour in one context. We do not focus on such insights in this paper as they are less actionable.
    - Comparative insights: These insights compare the user behaviour between two contexts [3] as shown in Table I.

TABLE I. Examples of Comparative Insights in BIM

Comparison	Example
time-specific	On **weekdays** you walk less than on **weekends**
parameter-specific	Your **heart rate** is higher on Mondays than on other days
event-specific	**When you bike**, you spend less calories per minute than **when you run**

C. Generation of Validated Insights From Data

Thousands of insights can be generated from even a simple database by slicing and dicing the data into different views. To streamline this process, we formulated a behaviour insight mining pipeline [13]. It consists of specialised blocks to look at data (what-to-look, where-to-look, how-to-look) and to generate text (what-to-say, when-to-say and how-to-say). For example, to generate the insight "On Weekdays you sleep less than on Weekends", the database should have logs of user's sleep duration and corresponding dates (what-to-look). The rows of the database corresponding to weekdays are considered as bin A and those corresponding to weekends are considered as bin B (how-to-look). Relevant filters are used to extract these rows (where-to-look). On comparing the average user's sleep duration in each bin, we find that bin A has a significantly lower value than bin B (what-to-say). Subsequently, a statistical significance test is performed to prove its statistical validity. A text realisation block structures and generates the appropriate textual output (when-to-say and how-to-say). Similarly, many comparisons (how-to-look) could be made between two periods such as:

- Mondays and other days
- Workdays and holidays
- February and March

Generally, thousands of such insights can be generated from even a moderately sized data. We validate these insights with the help of domain experts and proceed further.

A detailed description of how insights are generated and validated is explained in [3], [13].

D. Nonparametric Statistical Significance Tests

The data extracted from the two periods mentioned above come from two nonparametric sample distributions. The two most commonly adapted techniques to determine the statistical significance of such distributions are KS test and Mann-Whitney U (MW) test. The former is based on the shape of the distributions and the latter is based on the ranks of the samples. In this paper, we use data from a sleep monitoring device that measure the duration of sleep, sleep latency, etc [14]. Although these measures follow normal distribution, when looked at different slices and dices of the data such as Mondays vs Other days, they become nonparametric. Hence, we choose the KS test in this study. However, the MW test can also be used instead.

E. Neural Statistics

Neural networks have been used for wide range applications in Machine Learning such as signal denoising [15], image classification [16], stock prediction [17], and optical character recognition [18]. The ability of the neural network to learn basically any complex function makes it a universal function approximator. The simplicity in the way by which a neural network generates an inference makes it a suitable choice for many applications. Additionally, the transfer learning capability of the network [19]–[21] allows us to transfer the pre-learned knowledge of the network to solve different and more complex problems. This inspired us to use the neural network to approximate the statistical significance test.

F. Online Learning of User Preference

By permuting different contexts one may often find a large number of statistically significant insights but not all of these insights are useful to the user. Hence, the user's preference must be considered before presenting the insights to them. The personal preferences of end-users change with time. Filtering the insights based on statistical validity alone is not sufficient to satisfy their interests. A method to learn a user's preference in a convenient and flexible manner will solve this problem. Online learning technology can train models in a flexible manner while still being deployed in a consumer product or health coaching service [22], [23]. There is no existing literature on online learning of user preference nor the neural learning of statistical validity. Such learning will be of great use in BIM applications.

In this work, we present an online learning strategy that learns user preference while simultaneously maintaining the ability to realise the statistical significance. We make use of self supervision and transfer learning techniques to achieve this. In our technique, we assume that the user is interested in $N \in \mathbb{N}$ types of insights simultaneously at any point in time. The results indicate consistent performance for various values of N.

III. Methodology

The entire methodology was carried out in two stages, namely, the self-supervised learning stage and the online learning stage. Although each stage has a different data source, model architecture, training, and validation strategy, they share an important connection. The second stage model was transfer-learned from the first stage model. In this section, we describe the above-mentioned stages in detail.

A. Stage I: Self-Supervised Learning Stage

Self-supervision approach has been widely used to enrich a neural model using the input data and transformations of the input without the need for manual labelling. it has been widely used in fields like computer vision [24] language modelling [25] and speech modelling [26]. As a first stage, we conceptualised and developed a neural network model that learned rich feature representations to determine the statistical validity of comparative insights. We achieved this by training the model with highly diverse synthetic data. The data generation and model training are described below.

1. Problem Formulation

Let us consider an insight *i* that compares two distributions d_1 and d_2. The KS significance test can be represented as a function $f(d_1, d_2)$ that determines the p-value of d_1 and d_2. If the p-value is less than the significance level α, then, d_1 and d_2 are considered significantly different. We formulated a neural network N that approximates f as shown in Equation 1.

$$f \sim N \tag{1}$$

The neural network learns the function f by minimising the mean squared error loss function J_1 as shown in Eq 2.

$$J_1(\theta) = \frac{1}{n}\sum_{i=1}^{n}(f(d_{1i}, d_{2i}) - N_\theta(d_{1i}, d_{2i}))^2 \tag{2}$$

2. Data Generation for Base Model Selection

A dataset containing 300000 pairs of histograms of uniform distributions was generated using the NumPy-python package. The number of samples, mean and range of each distribution was chosen randomly. The ground truth labels for each pair of distribution were generated using the p-values of the two-sample KS test. The SciPy-python package was used for this. We compared it with our less optimised implementation of KS test and found it to give the same p-values. The dataset was subdivided into three equal parts, each for training, validation, and testing. We also made sure that each portion had balanced cases of significant and insignificant pairs.

3. Finalisation of Base Model Architecture

A domain-induced restriction of comparative insights is that the number of inputs is two and the number of outputs is one. Here, each input is the histogram of one of the distribution and the output is the statistical significance. Based on previous works on similar input/output constraints [27], [28], we came up with three neural network architectures, namely, a recurrent neural network (RNNA), a modified RNN (RNNB) and a siamese network (SIAM). The schematics of the RNNA architecture are shown in Fig. 1. The layers Ip1 and Ip2 are input layers, each having a fixed size of 100 elements. The layers F1 and F2, are fully connected layers, each with 50 neurons activated by a Leaky Rectified Linear Unit (ReLU) function. In fact, all layers in the network except the Final layer are activated by the Leaky ReLu function. Another level of fully connected layers, namely, F3 and F4 follow F1 and F2 respectively. We chose the number of neurons in each of these layers to be 20, which is less than the preceding layer, to have a compressed representation of the input signal. This type of step down architecture is commonly seen in the encoder part of auto-encoder neural networks [29]. This type of compression is helpful in transforming the input from spacial domain to meaningful feature domain. The layers F3 and F4 are concatenated and fed to a Simple Bidirectional Recurrent Neural Network (RNN) with 100 units. The rationale behind using an RNN is that the input needs to be considered a sequence rather than a vector as the inputs belong to two different contexts. We added another fully connected layer (F5) having 100 neurons to the output of the RNN. We believe that this layer generates rich features learned from the input data. The final layer is also a fully connected layer with one neuron activated by a thresholded ReLU activation function.

The RNNB model has every layer similar to the RNNA layer, except that it has 100 neurons in the F1 and F2 layers instead of 50. This is to see if increasing neurons would increase performance for a fixed purpose and input size. The SIAM network is also similar to the RNNA architecture, except that the F3 and F4 layers are subtracted rather than being concatenated and the RNN layer is replaced by a fully connected layer with 100 neurons.

4. Base Model Training and Testing

We trained and validated the three models in a self-supervised manner using the pairs of uniform distributions (histogram). The histogram was squeezed to 100 bins and the minimum and maximum range of histograms are fixed to be the minimum and maximum range of the dataset. This allows all the histograms to be comparable. Uniform distributions were chosen due to their close resemblance to real data that is commonly encountered in insight-mining tasks. In total, each of the training, validation and testing phases consisted of 100000 data samples. We did not go for an unequal split as we did not have that necessity due to the possibility to synthesise infinite data. The training was governed by Adam optimiser with a mean-squared-error loss function. The model that gave the best performance on the test set was considered as the base model. However, in real life, the data could also arise from complex or mixed distributions. Hence, we proceeded further with another level of fine-training.

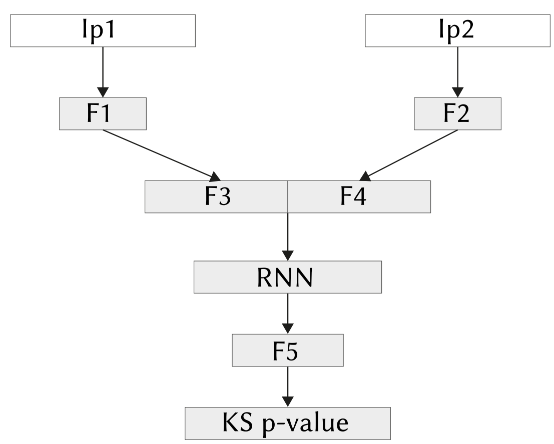


Fig. 1. Self-supervised neural network architecture for significance testing.

5. Improving the Base Model

In reality, the models encounter complex and nonparametric data distributions. For example, The distribution of hours of sleep on Mondays may not be normally distributed, but might follow a nonparametric distribution. Such scenarios are not considered during the base model training. Hence, we further train it with more diverse pairs of distributions (histogram) such as Gamma, Gumbel, Laplace, Normal, Uniform and Wald. On the whole, a total of 360000 pairs of distributions were generated and were equally split into training, validation and testing sets. Each of these sets consists of 120000 pairs of distributions (20000 pairs of each distribution). Both inputs of the network are always fed the same type of distribution, but with different parameters. For example, if one input of the network is a normal distribution, the other input is also a normal distribution but with different mean, range, and cardinality. The training labels are generated earlier. The training was governed by Adam optimiser with a mean squared error loss function. Once trained, the model can be used as a smart alternative to statistical significance testing to filter significant insights among all insights.

B. Stage II: Online Learning Stage

In the second stage, we transformed the base model to detect interesting insights while preserving its ability to detect significant insights.

1. Problem Formulation

In this stage, apart from the two distributions d_1 and d_2, we are also interested in the user model ∅. The user's preference can be represented by a function $p_u(k)$ that generates an interestingness value for a given insight k. This function can also be considered as a user interestingness/preference model. We formulated a transfer learning approach that uses a portion of network N i.e, N' and augments it with features representations generated from another neural network Δ that uses the state vector s of the insight k. Finally, the augmented network drives the overall network O that approximates $p_u(k)$ shown in Equation 3.

$$p_u(k) \sim O(N^{'}(d_1, d_2), \Delta(s)) \tag{3}$$

The neural network learns the function p_u by minimising the mean squared error loss function J_2 as shown in Eq 4.

$$J_2(\theta) = \frac{1}{n}\sum_{i=1}^{n}(p_u(k) - O_\phi(N^{'}(d_{1i}, d_{2i}), \Delta(s_i))^2 \tag{4}$$

In this work, we show that any improvement in approximating p_u does not have an impact on the approximation of f in Equation 1.

2. User Model Acquisition

The online learning strategy detects interesting insights with-out being instructed by the user explicitly. It uses a feedback form in a mobile application that displays a few insights that were scored high by the base model. We simulated a user who may choose the insights that they are interested in and the neural model learns from it. A sample feedback form is shown in Table II. In this work, the preference of the simulated user changes every month.

This feedback is equivalent to "labelling" in traditional online learning theory. To generate the insights to validate our online learning system, we obtained sleep and environmental sensor-data collected from a bedroom of a volunteer over a period of 4 months from May 2019 to August 2019. We logged various parameters such as the timestamp of the start of sleep, sleep duration, sleep latency, ambient light, ambient temperature, ambient sound and timestamp of waking-up. We generated insights for each day of the user using the procedure explained in [3]. The insight texts talk about the two contexts that it compares and an expression of the comparison such as "less than", "longer than", etc. The number of insights per day varied between a few hundred to few thousand. We simulated the user preference given below by automatically filling the feedback form for each day.

1. May: The user is interested in Insights related to Weekdays.
2. June: Weekend insights are interesting to the user.
3. July: The user prefers to know more about his sleep duration.
4. August: The user is again interested to know if he/she is doing well on weekends.

TABLE II. A Sample Insight Feedback Form

Insight	**Are you interested to see more of these type of insights?**
On Weekdays you sleep less than on Weekends	◯
...	...
Your take longer to fall asleep on Mondays than other days	◯

Collecting daily-feedback from a real user is expensive and time-consuming. Hence, we simulated the above monthly user-preference pattern. With this, we forced the model to adapt to abrupt changes in preferences; posing significant challenges to the network. Initially, all insights were initialised with an interestingness score of 0. The simulator re-assigns all statistically significant insights per day on a given month that satisfy the corresponding preference criteria with an interestingness score of 1. Although we labelled all the insights as interesting or not interesting, we observed later (section IV.D) that in actual practice, only a fraction of these labelled data were be used for training. Additionally, to simulate conflicting feedback, we randomly toggled 10% of the interestingness scores from 1 to 0 and viceversa. Since neural networks understand only numbers, we encoded each comparison insights into a single dimension binary vector s containing 220 elements where each element corresponds to one parameter of comparison. For example, one element corresponds to each day of the week. Hence, if the comparison is related to Mondays and weekends, the elements corresponding to Mondays, Saturdays, and Sundays are assigned a binary one and the rest are assigned zero. We injected this vector into the model while transfer learning for interestingness recognition. In the following subsection, we explain how the model was transfer-learned and how the online learning pipeline was implemented and evaluated.

3. Transfer Learning

Transfer learning was performed to enable the model to learn insight interestingness in addition to significance. The self-learned model was frozen from the input layers up to and including the F5 layer. The vector s was passed as input to another fully connected layer F6 with 100 neurons. This layer was concatenated with the F5 layer as shown in Fig. 2. The concatenated layers are fed to another fully connected layer F7 having 100 neurons. While the layer F6 is linearly activated, the F7 layer is activated by the ReLu function. Finally, the

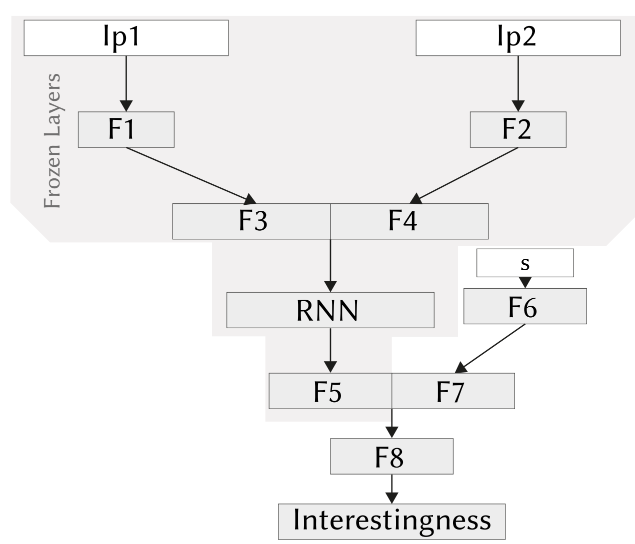


Fig. 2. Augmenting the base network for online learning.

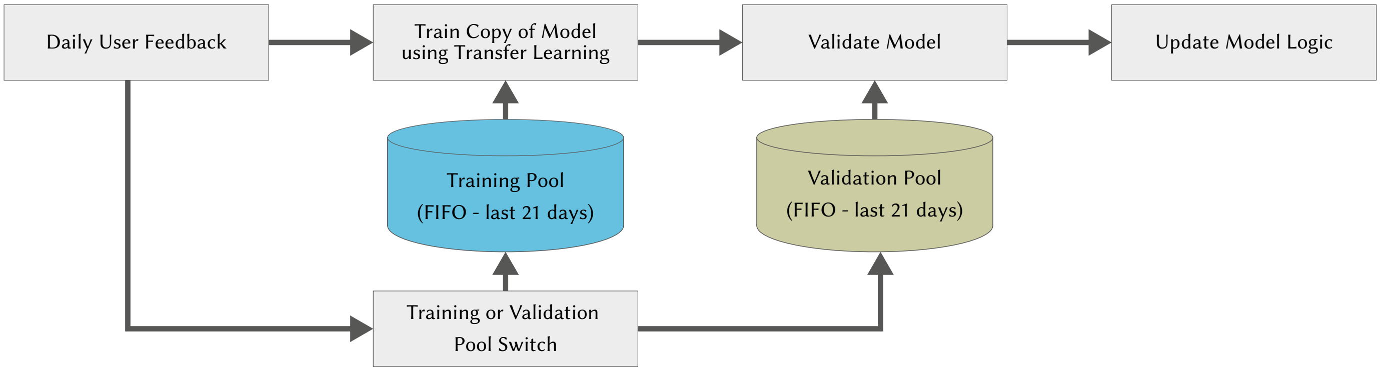


Fig. 3. Online learning through user feedback.

Algorithm 1: Training and Validation Switch Logic
Result: Assign sample to Training pool, Validation pool or Both

```
feedback = pop(feedback_stack)
prediction_error = validate_model(feedback)
if mode == Accelerated Training then
    if prediction_error < 0.3 and coin_toss() == Heads and positive_fraction in range [0.42,0.6] then
        switch_state = Both
        update(positive_fraction)
    end
else
    if prediction_error < 0.1 and coin_toss() == Heads and positive_fraction in range [0.42,0.6] then
        switch_state = Validation
        update(positive_fraction)
    else
        if positive_fraction in range [0.42,0.6] then
            switch_state = Training
            update(positive_fraction)
        end
    end
end
```

output layer is a single neuron fully connected layer activated by a sigmoid activation function. Notice that the final layer is activated by a sigmoid function as this is a binary classification problem trained on user preferences instead of significance. By performing this transfer learning, the model retains the features that correspond to the significance and simultaneously recognise the interestingness of insights based on user preference.

4. Learning Modes

The architecture of the online learning scheme is presented in Fig. 3. The scheme is executed in two modes, namely, accelerated learning mode and normal learning mode. These modes determine how well the models are trained. The accelerated learning mode, by default, starts from the first day of usage of the insight generator till the tenth day. Then, the normal mode begins. During the accelerated learning mode, the model learns quickly from the data and during the normal mode, it learns at a normal pace. This is achieved by varying the learning rate. Thus, the accelerated training mode has a higher learning rate.

5. Training and Validation of Switch Logic

Insights are generated on a daily basis. The insights contain a textual description of the behaviour and the back-tracking information of the corresponding data. Using this, we can get the data distributions corresponding to insights. Every day, the insights are assigned an interestingness value based on the user's feedback. The learning mode, prediction_error and positive_fraction help to determine if the feedback will be used to train or validate the model. This logic is represented in Algorithm 1.

The system collects the feedback and stores them in a FIFO stack named *feedback_stack*. The algorithm starts with popping a feedback from the stack and calculating its prediction error using the *validate_model* function. This function runs the neural model on the feedback data to predict an interestingness score and calculates its absolute difference from the true label. Then, the system follows subsequent steps to assign the data to one of the pool based on the *prediction_error*, learning mode, positive_fraction and a coin toss as show in Algorithm 1. Here, the *positive_fraction* is mean of all the interestingness score in the training pool, the *coin_toss* function generates a *Head* or a *Tail* randomly.

If the user does not give any feedback, the model does not update since the system implicitly assumes that the user's preference is unchanged.

6. Pool Maintenance Logic

Both the pools are maintained to hold only a maximum limit of days of data. We fixed this to be 14 days because we assume a user's interestingness remains fairly unchanged for a period of two weeks. Every 20 days, the model forcefully pops out 7 days of the oldest data in a FIFO fashion. This helps to avoid overloading the training and validation pools and forgetting older preferences. Additionally, the validation pool is completely emptied at the beginning of the first day of the normal learning phase.

7. Update Logic and Metrics

At the end of every day, a copy of the model is trained on the training pool and validated on the validation pool. If the validation accuracy exceeds a set limit (here 70%), the old model is replaced by the recently trained model. However, as an exception in the accelerated learning mode, the model is updated every day irrespective of its performance. This purposefully over-fits the model to the insights during accelerating learning mode. The performance of online learning is monitored using statistical measures, namely, sensitivity, specificity, and accuracy in predicting the interestingness of insights. Additionally, we introduce the significance preservation score, which is calculated as shown in Equation 5.

$$P_s = N_a / N_p \tag{5}$$

where, N_a and N_p are the number of actual interesting insights in the validation pool and the number of predicted interesting insights during validation, respectively. The P_s is not defined when N_p is zero. This is a limitation of the metric.

IV. Experimental Results

In this section, we present the results that we obtained at each stage.

A. Choosing The Base Model Architecture

An example of histograms of significant and insignificant pairs of normal distributions is shown in Fig. 4. It also demonstrates the variation of magnitude, range and cardinality (more samples have a smoother curve) of the synthetic data. Each of the base model architecture, namely, RNNA, RNNB, and SIAM were trained, validated and tested using the dataset containing only normal distributions. The performance of each model is presented in Table III. We observed that the RNNA model exhibits a test accuracy of 92% in predicting whether an insight is interesting or not. The performance of RNNA is thereby comparatively better than that of RNNB. This shows that more neurons do not always lead to improved performance. Also, RNNA exhibits slightly better performance than the SIAM network. This could be due to the sequential treatment of the data by the RNN which is part of the network. Additionally, since the SIAM network has fewer neurons, it also provides evidence that fewer neurons might not help either. In our view, the neural model should have an adequate number of neurons and parameters and an explainable architecture, which is, unfortunately, missing in recent works in this field. Hence, the RNNA architecture is chosen as the base model and considered for further analysis.

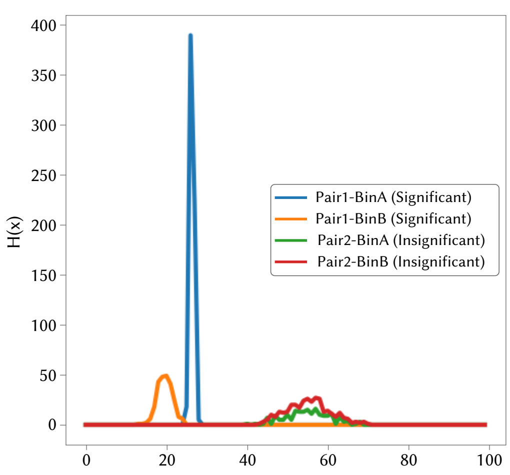


Fig. 4. Pair of normal distributions without significant difference.

TABLE III. Performance of Different Models while Training and Testing with Normal Distribution

Model	**Description**	**Accuracy $\alpha = 0.05$**
RNNA	Bidirectional RNN layer	0.92
RNNB	More neurons	0.86
SIAM	Siamese Network	0.87

B. Improving Base Model Training

We trained the base model using diverse pairs of distributions (histogram) such as Gamma, Gumbel, Laplace, Normal, Uniform and Wald. We observe that when we tested each distribution as shown in Fig. 5, we find out that the performance of the model to normal distribution remained at 0.92, but the uniform was even higher at 0.97. The worst performance was observed on Wald distribution. We have additional evidence that this is a limitation of the actual KS test that is being reflected in the neural model. It is also found that few distributions exhibit improved performances as alpha increases and few showed weaker performance as alpha increases.

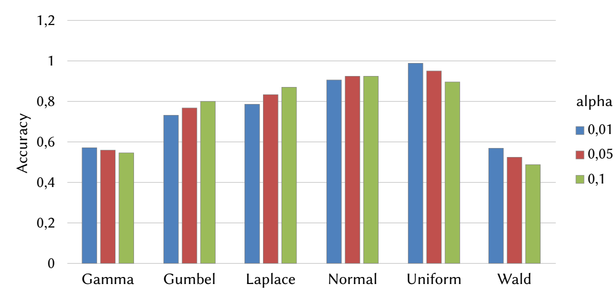


Fig. 5. Gaussian trained model on mixed distributions.

C. Generated Insights

We used the pipeline approach described in C to generate insights on the users sleep behaviour. A few examples are shown in Table IV.

TABLE IV. Representative Set of Generated User Behaviour Insights

S. No	**Insights**
1	In Q3, you slept shorter than in Q2
2	In October, it took longer for you to fall asleep than in September
3	In May, you spent less time in the bed than on other months
4	You slept longer, when the temperature during start of sleep was between 17°C and 30°C than when it was more than 30°C
5	It took longer for you to fall asleep, when the humidity during the start of sleep was high than when it was ideal
6	It took longer for you to fall asleep, when the illuminance during start of sleep was brighter than normal than when it was dim

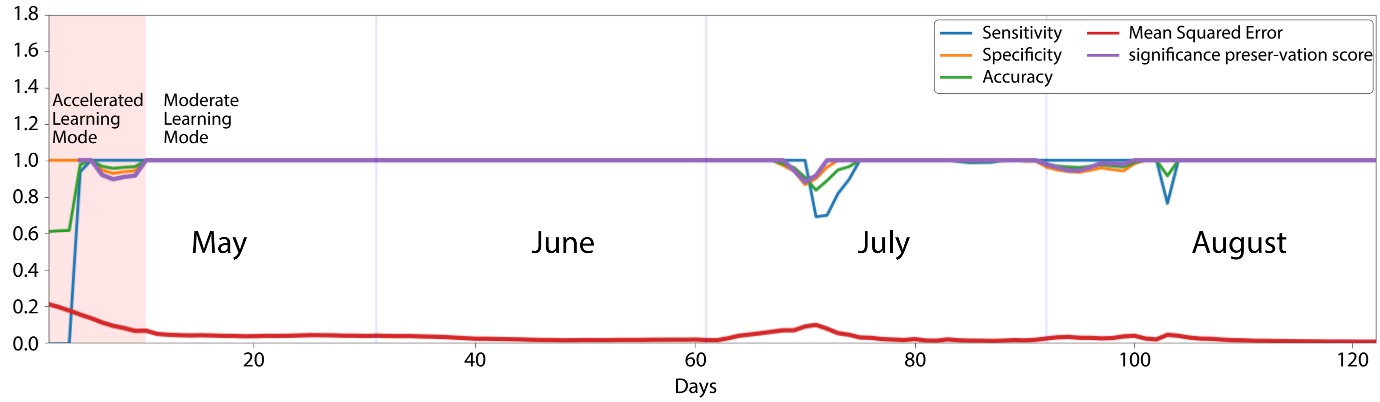


Fig. 6. Timeline of online learning with performance indicators.

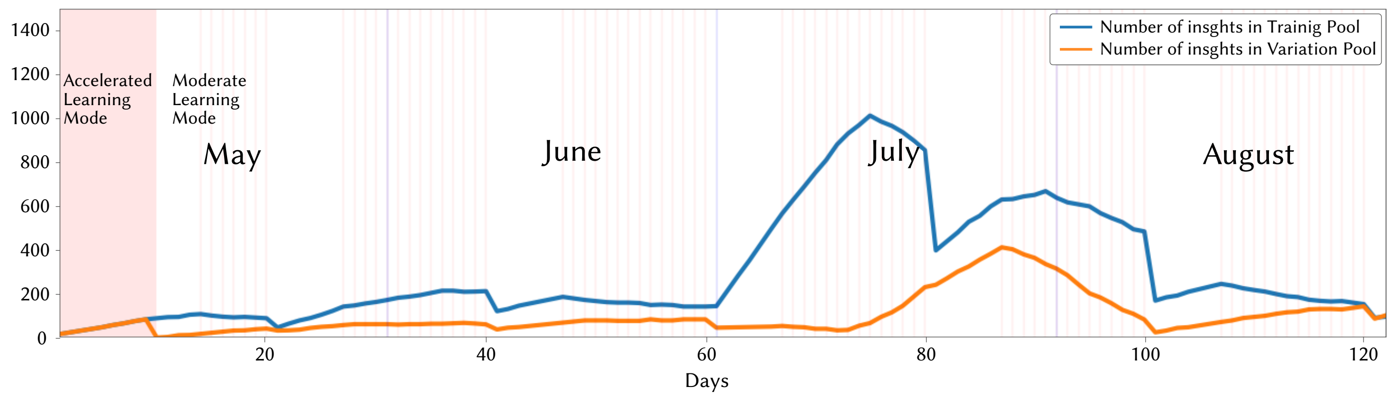


Fig. 7. Size of training and validation pool.

D. Online Learning on Single Preference

We simulated a user who has a preference for a single type of insight for a given month. We generated insights from another real user data and allowed the simulated user to provide feedback on each insight based on the implicitly defined preferences as follows

1. May: insights that talk about behaviour in weekdays
2. June: weekend insights
3. July: insights that describe how long the user sleeps
4. August: weekend insights

Subsequently, we initiated the online learning scheme and the performance metrics are presented in Fig. 6. We put the system on accelerated learning mode for the first 10 days. It is observed that the accuracy, sensitivity, and specificity were unstable during the first 4 days of the accelerated learning phase. From the fifth day onwards, the three measures show improvement and are in the range of 0.9 to 1. The Ps measure is not defined when there are no significantly valid insights that are interesting. This is observed till day 3 and on day 4, 100% Ps is observed. This implies that the model exhibits significance preservation starting at least from day 4 onwards. The performance is rather stable all the while during the remaining days of May and the entire June. Even though there is a transition between weekday insights and weekend insights, the model seems to adapt very well. In the months of July and August, there are visible drops in the performance around the 10th day of the month even though the preference changed on the 1st of both months. This could be an instability caused by the sudden rise in the training pool and reduction of validation pool data as shown in Fig. 7. In general, the pool maintenance logic is able to control the number of training and test data points. Although the first half of July saw a huge influx of training data, the maintenance logic prevented the training pool from overloading. Otherwise, there would have been a huge chance of exposing the model to noise in the data. The mean squared error (MSE) curve shows that the error between predictions and ground truth is not very high. The MSE decreased more steeply during the accelerated learning mode compared to the normal mode. There are periodic valleys in the training pool count and validation pool count denoting the reach of the 20-day window for cleanup of the pool. Also, additional cleanups are done every day when the number of days of insights in the pool exceeds 14. All cleanups on the training and validation pool are indicated by faint red vertical lines in Fig. 7. Additionally, we could observe that number of labelled insights (training + validation) at given point in time is in the range [20, 1087] however, the number of newly fed insights ranges from 0 to 74 per day with an average of 15.6 insights per day. Thus, it doesn't require to label all the insights, but only as much as required by the model.

E. Online Learning on Multiple Preferences

Usually, the user preference is not as simple as described in the previous section. It is a combination of multiple preferences. Hence, we simulated the user to have multiple insight preferences at a time. We first investigated in detail, the effect of a dual preference user model and secondly, discuss its general impact by simulating multi-preference scenarios up to 10 simultaneous user preferences.

1. Dual Preference

We considered a dual preference scenario where the user has following pairs of preferences:

1. May: insights that talk about user behaviour during weekdays or weekends
2. June: insights that talk about weekend behaviour or the user's sleep latency

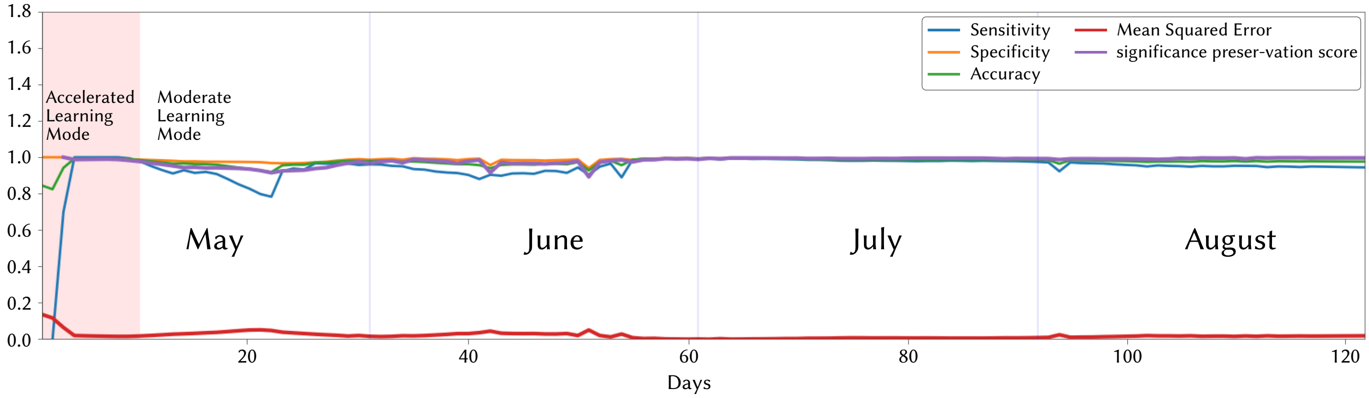


Fig. 8. Timeline of online learning on dual preference with performance indicators.

3. July: insights that describe how long the user sleeps and insights measured over the quarterly period
4. August: weekend-insights and insights talking about sleep latency

In the beginning, the model performs slightly better than on the single preference user model. There is a drop in performance around the twenty second day of the first month. However, the model is comparatively steady thereafter.

We had purposefully set the user preference in August to be the same as in June to see how well the model unlearns and relearns. From Fig. 8, it can be observed that the model learns in August, much more smoothly than in June. In overall, the dual preference scenario has slightly fewer and less intense performance drops than the single preference model.

F. Higher Order Preference

We defined a list of possible categories of insights as shown Table V. Each insight can belong to one or more categories. We simulated a multi-preference user by randomly choosing a combination of N out of the 14 insight categories for each month. We ran our learning algorithm under these conditions and measured the performance in terms of accuracy and preservation of significance as shown in Fig. 9.

TABLE V. List of Preference Profiles

S. No	Insights
1	Insights that talk about behaviour on weekdays
2	Weekend insights
3	Insights on duration of sleep
4	Time to fall asleep
5	Time spent on bed
6	Average time of getting into the bed
7	Average time of getting out of bed
8	Insights consolidating behaviour over a quarter
9	Insights talking about monthly behaviour
10	Yearly insights
11	Impact of humidity at start of sleep on sleep measures
12	Impact of humidity at end of sleep on sleep measures
13	Impact of temperature during start of sleep
14	Temperature during end of sleep

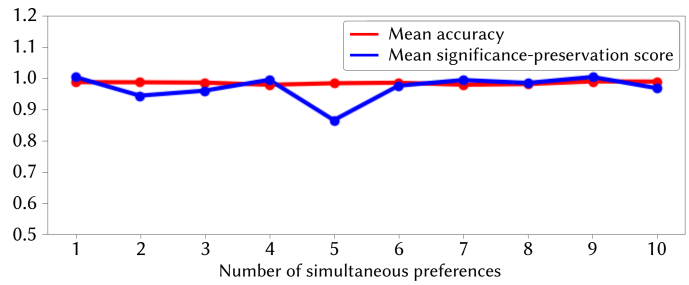


Fig. 9. Mean accuracy and mean significance preservation score (equation 5) by varying the number of simultaneous user preferences.

It is observed that the mean values of accuracy are consistently high. The difference between the highest and lowest mean accuracy score is as low as 0.010. The difference between the highest and lowest significance preservation score is 0.137. This is still good considering the fact that the means range from 0.86 to 1.

V. Discussion

A. Real User Feedback

In this work, we collected a real user's sleep signal and simulated their feedback to test the performance of the system. Our simulated-user was strictly compliant with the predefined or randomly chosen preference profiles. However real users might have conflicting preferences. For example, they might like an insight about sleep latency on weekdays, but at the same time might not be interested on their sleep latency on weekends. This means that the interestingness score of insights that talk about sleep latency is not 1, but a fraction of 1. There is no standard mechanism to simulating such conflicting scenarios. However, assuming multiple dimensions for the sleep profile avoids these confusions as the overlap between similar insights is reduced. Using the same example above, if we model the user preference as a two dimensional entity, we would define one interestingness profile to be a combination of sleep latency on weekdays and another to be sleep latency on weekends. Usually, random simulations have the risk of learning and unlearning within a short period of time thereby nullifying the notion of a strong user-preference. But, a few conflicting feedback can not bias the results as neural networks learn in small steps and are very robust to noisy labels [30].

B. Resource Consumption

The proposed neural models were trained and run on an NVIDIA V100 server. The final trained model has 3.1M param-eters. To update the model for one day of insights, the model takes on an average 7.9 seconds. However, this can be brought down with the help of pruning techniques. Even otherwise, this is fast enough for a server based mobile application in which the training will be performed remotely. The proposed algorithm is light enough to be run on an edge device (mobile phones, smart watches and tablets) and is a once-a-day task.

VI. Conclusions and Future Scope

In this work, we proposed an artificial neural network model to score pre-validated insights of user behaviour from data. We consider comparative insights that talk about how a quantity differs in two different contexts. We score these insights considering the significance of the user behaviour depicted in it and the user's preference towards the insight. We used ANN to build an insight scoring model for its ability to learn, unlearn, and relearn tasks. For this, we used self-supervised training to train an ANN to perform a statistical significance test, namely the Kolmogorov-Smirnov test. Next, we augmented the architecture to learn user preference with an interactive-learning scheme. We evaluated three different model architectures of ANNs and chose the best to be our base model: a simple neural network with recurrent neural network (RNN) layers with fewer neurons. However, the other two networks: a similar RNN network with more neurons and a slightly different siamese network also exhibited satisfactory performance. Subsequently, we improved our RNN model with more and more variety of input distributions following a self-supervised learning approach. We proceeded to relearn this model to also consider user preferences in an interactive setting with the help of transfer learning.

We subsequently learn user preference on the same model using their feedback. For this, our model requires three inputs, namely, the distribution of the quantity in one context, its distribution in the other context, and a binary encoding of the insight. We froze a part of the base model and augmented it with an additional input layer that reads the binary encoding vector. We trained it on a real dataset while simulating user preferences. We came up with single and multiple user preference scenarios. The model performs well with consistently good accuracy and preserves its knowledge about statistical significance while learning interestingness. This made the network unique in an intelligent way. Also, this is the first attempt in which a single neuron is shown to play two simultaneous roles. Our evaluations suggest that the model can learn complex and dynamic user preferences. In future, we would like to perform a field testing of the proposed technique and also device ways to obtain feedback from users with the least disturbance.

Acknowledgment

This paper is an extension of our previous work presented at the IntelLang workshop at the ECAI 2020 conference [31]. In that, we considered user preferences that are simple and concerns at-most one type of insights at a time. In this paper, we extend this to two or more types of similar or different insights being preferred at the same time.

This work was supported by the Horizon H2020 Marie Skłodowska-Curie Actions Initial Training Network European Industrial Doctorates project under grant agreement No. 812882 (PhilHumans).

References

[1] M. Hingle, H. Patrick, "There are thousands of apps for that: navigating mobile technology for nutrition education and behavior," *Journal of nutrition education and behavior*, vol. 48, no. 3, pp. 213–218, 2016.

[2] J. P. Higgins, "Smartphone applications for patients' health and fitness," *The American journal of medicine*, vol. 129, no. 1, pp. 11–19, 2016.

[3] A. Härmä, R. Helaoui, "Probabilistic scoring of validated in-sights for personal health services," in *2016 IEEE Symposium Series on Computational Intelligence (SSCI)*, 2016, pp. 1–6, IEEE.

[4] S. J. O'Malley, R. T. Smith, B. H. Thomas, "Data mining office behavioural information from simple sensors.," in *AUIC*, 2012, pp. 97–98.

[5] D. Braun, E. Reiter, A. Siddharthan, "Saferdrive: An nlg-based behaviour change support system for drivers," *Natural Language Engineering*, vol. 24, no. 4, pp. 551–588, 2018.

[6] S. G. Sripada, F. Gao, "Linguistic interpretations of scuba dive computer data," in *2007 11th International Conference Information Visualization (IV'07)*, 2007, pp. 436–441, IEEE.

[7] R. Agrawal, J. C. Shafer, "Parallel mining of association rules," *IEEE Transactions on knowledge and Data Engineering*, vol. 8, no. 6, pp. 962–969, 1996.

[8] A. A. Freitas, "On rule interestingness measures," in *Research and Development in Expert Systems XV*, Springer, 1999, pp. 147–158.

[9] U. Fayyad, G. Piatetsky-Shapiro, P. Smyth, "From data mining to knowledge discovery in databases," *AI magazine*, vol. 17, no. 3, pp. 37–37, 1996.

[10] N. Sudarsanam, N. Kumar, A. Sharma, B. Ravindran, "Rate of change analysis for interestingness measures," *Knowledge and Information Systems*, pp. 1–20, 2019.

[11] H. op den Akker, M. Cabrita, R. op den Akker, V. M. Jones, H. J. Hermens, "Tailored motivational message generation: A model and practical framework for real-time physical activity coaching," *Journal of biomedical informatics*, vol. 55, pp. 104–115, 2015.

[12] E. Reiter, R. Robertson, L. M. Osman, "Lessons from a failure: Generating tailored smoking cessation letters," *Artificial Intelligence*, vol. 144, no. 1-2, pp. 41–58, 2003.

[13] A. Susaiyah, A. Härmä, E. Reiter, R. Helaoui, M. Petković, et al., "Towards a generalised framework for behaviour insight mining," in *SmartPHIL: 1st Workshop on Smart Personal Health Interfaces*, 2020, ACM.

[14] "Connected sleep and wake up light hf3670/60," Mar 2020. [Online]. Available: https://www.usa.philips.com/shop/US_Air-FryerOnly/personal-care/light-therapy/smartsleep-connected-sleep-and-wake-up-light/p/HF3670_60.

[15] K. Antczak, "Deep recurrent neural networks for ecg signal denoising," *arXiv preprint arXiv:1807.11551*, 2018.

[16] S. Kaymak, A. Helwan, D. Uzun, "Breast cancer image classification using artificial neural networks," *Procedia computer science*, vol. 120, pp. 126–131, 2017.

[17] S. Chopra, D. Yadav, A. Chopra, "Artificial neural networks based indian stock market price prediction: before and after demonetization," *J Swarm Intel Evol Comput*, vol. 8, no. 174, p. 2, 2019.

[18] K. T. Islam, G. Mujtaba, R. G. Raj, H. F. Nweke, "Handwritten digits recognition with artificial neural network," in *2017 International Conference on Engineering Technology and Technopreneurship (ICE2T)*, 2017, pp. 1–4, IEEE.

[19] J. Tao, X. Fang, "Toward multi-label sentiment analysis: a transfer learning based approach," *Journal of Big Data*, vol. 7, no. 1, pp. 1–26, 2020.

[20] M. Long, Y. Cao, J. Wang, M. I. Jordan, "Learning transfer-able features with deep adaptation networks," *arXiv preprint arXiv:1502.02791*, 2015.

[21] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, J. Dean, "Distributed representations of words and phrases and their compositionality," in *Advances in neural information processing systems*, 2013, pp. 3111–3119.

[22] B. Settles, "Active learning literature survey," University of Wisconsin-Madison Department of Computer Sciences, 2009.

[23] B. Settles, "From theories to queries: Active learning in practice," in *Active Learning and Experimental Design workshop In conjunction with AISTATS 2010*, 2011, pp. 1-18.

[24] X. Zhai, A. Oliver, A. Kolesnikov, L. Beyer, "S4l: Self-supervised semi-supervised learning," in *Proceedings of the IEEE international conference on computer vision, 2019*, pp. 1476-1485.

[25] J. Devlin, M. W. Chang, K. Lee, K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[26] S. Pascual, M. Ravanelli, J. Serrà , A. Bonafonte, Y. Bengio, "Learning problem-agnostic speech representations from multiple self-supervised tasks," 2019.

[27] P. Neculoiu, M. Versteegh, M. Rotaru, "Learning text similarity with siamese recurrent networks," in *Proceedings of the 1st Workshop on*

Representation Learning for NLP, 2016, pp. 148-157.
[28] S. Berlemont, G. Lefebvre, S. Duffner, C. Garcia, "Siamese neural network based similarity metric for inertial gesture classification and rejection," in *International Conference on Automatic Face and Gesture Recognition*, 2015, pp. 1–6.
[29] W. Sun, S. Shao, R. Zhao, R. Yan, X. Zhang, X. Chen, "A sparse auto-encoder-based deep neural network approach for induction motor faults classification," *Measurement*, vol. 89, pp. 171-178, 2016, doi: https://doi.org/10.1016/j.measurement.2016.04.007.
[30] K. Choi, G. Fazekas, K. Cho, M. Sandler, "The effects of noisy labels on deep convolutional neural networks for music tagging," *IEEE Transactions on Emerging Topics in Computational Intelligence*, vol. 2, no. 2, pp. 139-149, 2018, doi: 10.1109/TETCI.2017.2771298.
[31] A. Susaiyah, A. Härmä, E. Reiter, M. Petković, "Iterative neural scoring of validated insight candidates," in *ECAI workshop on Intelligent Information Processing and Natural Language Generation*, Santiago de Compostela, Spain, Sep 2020.

Allmin Susaiyah

llmin completed his masters (by research) from the Indian Institute of Technology, Madras, and started his career as a Researcher at Philips, India in 2017. Currently, he is pursuing his doctoral studies at TU/e, Netherlands, and Philips Research, Eindhoven. His areas of interest are, artificial intelligence, insight mining, natural language generation, deep learning, and medical imaging. He is also a Marie Curie Fellow being a part of the PhilHumans consortium.

Aki Härmä

Dr. Aki Härmä did his PhD in audio and speech signal processing at HUT, Finland, in 2001. After positions at Lucent Bell Labs and HUT he joined Philips Research in 2004. He has published more than 100 conference and journal papers, numerous patents, and contributed to various product lines of Philips. He is currently at the position of Principal Scientist in the Data Science Department of Philips Research working on big data technologies, predictive analytics, and computational intelligence for automated personal and home health services. He has supervised several students at different levels.

Ehud Reiter

Ehud Reiter received the A.B. degree in mathematics and the Ph.D. degree in computer science from Harvard University, in 1982 and 1990, respectively., He is currently a Professor of computing science with the University of Aberdeen, and also a Chief Scientist with Arria NLG. He specializes in natural language generation, and has a Google Scholar h-index of 40. He also holds eight patents. Prof. Reiter is currently the Chair of ACL SIGGEN, the special interest group in (natural language) generation of the Association for Computational Linguistics.

Milan Petković

Milan Petković is a Professor at the Department of Mathematics and Computer Science, Eindhoven University of Technology. He is also the Head of the Department of Data Science, Philips Research, Eindhoven, the Netherlands. Among his research interests are information security, secure content management, privacy protection, multimedia information retrieval, and database systems.

Artificial Canaries: Early Warning Signs for Anticipatory and Democratic Governance of AI

Carla Zoe Cremer[1,2]*, Jess Whittlestone[2]

[1] Future of Humanity Institute, University of Oxford (United Kingdom)
[2] Centre for the Study of Existential Risk, University of Cambridge (United Kingdom)

Received 17 January 2021 | Accepted 11 February 2021 | Published 24 February 2021

Abstract

We propose a method for identifying early warning signs of transformative progress in artificial intelligence (AI), and discuss how these can support the anticipatory and democratic governance of AI. We call these early warning signs 'canaries', based on the use of canaries to provide early warnings of unsafe air pollution in coal mines. Our method combines expert elicitation and collaborative causal graphs to identify key milestones and identify the relationships between them. We present two illustrations of how this method could be used: to identify early warnings of harmful impacts of language models; and of progress towards high-level machine intelligence. Identifying early warning signs of transformative applications can support more efficient monitoring and timely regulation of progress in AI: as AI advances, its impacts on society may be too great to be governed retrospectively. It is essential that those impacted by AI have a say in how it is governed. Early warnings can give the public time and focus to influence emerging technologies using democratic, participatory technology assessments. We discuss the challenges in identifying early warning signals and propose directions for future work.

Keywords

AI Governance, Forecasting, Anticipatory Governance, Participatory Technology Assessments.

DOI: 10.9781/ijimai.2021.02.011

I. Introduction

Progress in artificial intelligence (AI) research has accelerated in recent years. Applications are already changing society [1] and some researchers warn that continued progress could precipitate transformative impacts [2]–[5]. We use the term "transformative AI" to describe a range of possible advances with potential to impact society in significant and hard-to-reverse ways [6]. For example, future machine learning systems could be used to optimise management of safety-critical infrastructure [7]. Advanced language models could be used in ways that corrupt our online information ecosystem [8] and future advances in AI systems could trigger widespread labour automation [9].

There is an urgent need to develop anticipatory governance approaches to AI development and deployment. As AI advances, its impacts on society will become more profound, and some harms may be too great to rely on purely 'reactive' or retrospective governance.

Anticipating future impacts is a challenging task. Experts show substantial disagreement about when different advances in AI capabilities should be expected [10], [11]. Policy-makers face challenges in keeping pace with technological progress: it is difficult to foresee impacts before a technology is deployed, but after deployment it may already be too late to shape impacts, and some harm may already have been done [12]. Ideally, we would focus preventative, anticipatory efforts on applications which are close enough to deployment to be meaningfully influenced today, but whose impacts we are not already seeing. Finding 'early warning signs' of transformative AI applications can help us to do this.

Early warning signs can also help democratise AI development and governance. They can provide time and direction for much-needed public discourse about what we want and do not want from AI. It is not enough for anticipatory governance to look out for supposedly 'inevitable' future impacts. We are not mere bystanders in this AI revolution: the futures we occupy will be futures of our own making, driven by the actions of technology developers, policymakers, civil society and the public. In order to prevent foreseeable harms towards those people who bear the effects of AI deployments, we must find ways for AI developers to be held accountable to the society which they are embedded in. If we want AI to benefit society broadly, we must urgently find ways to give democratic control to those who will be impacted. Our aim with identifying early warning signs is to develop anticipatory methods which can prompt a focussed civic discourse around significant developments and provide a wider range of people with the information they need to contribute to conversations about the future of AI.

We present a methodology for identifying early warning signs of potentially transformative impacts of AI and discuss how these can feed into more anticipatory and democratic governance processes. We call these early warning signs 'canaries' based on the practice of using canaries to provide early warnings of unsafe air pollution in coal mines in the industrial revolution. Others before us have used this term in the context of AI to stress the importance of early warning

* Corresponding author.

E-mail address: carla.cremer@philosophy.ox.ac.uk

signs [13], [14], but this is the first attempt to outline in detail how such 'artificial canaries' might be identified and used.

Our methodology is a prototype but we believe it provides an important first step towards assessing and then trialling the feasibility of identifying canaries. We first present the approach and then illustrate it on two high-level examples, in which we identify preliminary warning signs of AI applications that could undermine democracy, and warning signs of progress towards high-level machine intelligence (HLMI). We explain why early warning signs are needed by drawing on the literature of participatory technology assessments, and we discuss the advantages and practical challenges of this method in the hope of preparing future research that might attempt to put this method into practise. Our theoretical exploration of a method to identify early warning signs of transformative applications provides a foundation towards more anticipatory, accountable and democratic governance of AI in practice.

II. Related Work

We rely on two main bodies of work. Our methodology for identifying canaries relies on the literature on *forecasting and monitoring AI*. Our suggestions for how canaries might be used once identified build on work on *participatory technology assessments*, which stresses a more inclusive approach to technology governance. While substantial research exists in both these areas, we believe this is the first piece of work that shows how they could feed into each other.

A. AI Forecasting and Monitoring

Over the past decade, an increasing number of studies have attempted to forecast AI progress. They commonly use expert elicitations to generate probabilistic estimates for when different AI advances and milestones will be achieved [10], [15]–[17]. For example, [16] ask experts about when specific milestones in AI will be achieved, including passing the Turing Test or passing third grade. Both [15] and [10] ask experts to predict the arrival of high-level machine intelligence (HLMI), which the latter define as when "unaided machines can accomplish every task better and more cheaply than human workers".

However, we should be cautious about giving results from these surveys too much weight. These studies have several limitations, including the fact that the questions asked are often ambiguous, that expertise is narrowly defined, and that respondents do not receive training in quantitative forecasting [11], [18]. Experts disagree substantially about when crucial capabilities will be achieved [10], but these surveys cannot tell us who (if anyone) is more accurate in their predictions.

Issues of accuracy and reliability aside, forecasts focused solely on timelines for specific events are limited in how much they can inform our decisions about AI today. While it is interesting to know how much experts disagree on AI progress via these probabilistic estimates, they cannot tell us why experts disagree or what would change their minds. Surveys tell us little about what early warning signs to look out for or where we should place our focus today to shape the future development and impact of AI.

At the same time, several projects, e.g. [19]–[22], have begun to track and measure progress in AI. These projects focus on a range of indicators relevant to AI progress, but do not make any systematic attempt to identify which markers of progress are more important than others for the preparation of transformative applications. Time and attention for tracking progress is limited and it would be helpful if we were able to prioritise and monitor those research areas that are most relevant to mitigating risks.

Recognising some of the limitations of existing work, [23] aims for a more holistic approach to AI forecasting. This framework emphasises the use of the Delphi technique [24] to aggregate different perspectives of a group of experts, and cognitive mapping methods to study how different milestones relate to one another, rather than to simply forecast milestones in isolation. We agree that such methods might address some limitations of previous work in both AI forecasting and monitoring. AI forecasting has focused on timelines for particularly extreme events, but these timelines are subject to enormous uncertainty and do not indicate near-term warning signs. AI measurement initiatives have the opposite limitation: they focus on near-term progress, but with little systematic reflection on which avenues of progress are, from a governance perspective, more important to monitor than others. What is needed are attempts to identify areas of progress today that may be particularly important to pay attention to, given concerns about the kinds of transformative AI systems that may be possible in future.

B. Participatory Technology Assessments

Presently, the impacts of AI are largely shaped by a small group of powerful people with a narrow perspective which can be at odds with public interest [25]. Only a few powerful actors, such as governments, defence agencies, and firms the size of Google or Amazon, have the resources to conduct ambitious research projects. Democratic control over these research projects is limited. Governments retain discretion over what gets regulated, large technology firms can distort and avoid policies via intensive lobbying [26] and defence agencies may classify ongoing research.

Recognising these problems, a number of initiatives over the past few years have emphasised the need for wider participation in the development and governance of AI [27]–[29]. In considering how best to achieve this, it is helpful to look to the field of science and technology studies (STS) which has long considered the value of democratising research progress [30], [31]. Several publications refer to the 'participatory turn' [32] in STS and an increasing interest in the role of the non-expert in technology development and assessment [27]. More recently, in the spirit of "democratic experimentation" [33], various methods for civic participation have been developed and trialled, including deliberative polls, citizen juries and scenario exercises [33].

With a widening conception of expertise, a large body of research on "participatory technology assessment" (PTA) has emerged, aiming to examine how we might increase civic participation in how technology is developed, assessed and rolled out. We cannot summarise this wide-ranging and complex body of work fully here. But we point towards some relevant pieces for interested readers to begin with. [34] and [35] present a typology of the methods and goals of participating, which now come in many forms. This means that assessments of the success of PTAs are challenging [33] and ongoing because different studies evaluate different PTA processes against different goals [34]. Yet while scholars recognise remaining limitations of PTAs [31], several arguments for their advantages have been brought forward, ranging from citizen agency to consensus identification and justice. There are good reasons to believe that non-experts possess relevant end-user expertise. They often quickly develop the relevant subject-matter understanding to contribute meaningfully, leading to better epistemic outcomes due to a greater diversity of views which result in a cancellation of errors [36], [37]. To assess the performance of PTAs scholars draw from case studies and identify best practices [38]–[40].

There is an important difference between truly participatory, democratically minded, technology assessments, and consultations that use the public to help legitimise a preconceived technology [41]. The question of how to make PTAs count in established representational democracies is an ongoing challenge to the field [31], [33]. But [42], who present a recent example of collective technology policy-making, show that success and impact with PTAs is possible. [40] draw from 38 international case studies to extract best practices, building on [38], who showcase great diversity of possible ways in which to draw on the public. Comparing different approaches is difficult, but has been done [39], [43]. [41] present a conceptual framework with which to design and assess PTAs, [44] compares online versus offline methodologies and in [35] we find a typology of various design choices for public engagement mechanisms. See also [45] for a helpful discussion on how to determine the diversity of participants, [46] on what counts as expertise in foresight and [30], [32], [47] for challenges to be aware of in implementing PTAs.

Many before us have noted that we need wider participation in the development and governance of AI, including by calling for the use of PTAs in designing algorithms [48], [49]. We see a need to go beyond greater participation in addressing existing problems with algorithms and propose that wider participation should also be considered in conversations about future AI impacts.

Experts and citizens each have a role to play in ensuring that AI governance is informed by and inclusive of a wide range of knowledge, concerns and perspectives. However, the question of how best to marry expert foresight and citizen engagement is a challenging one. While a full answer to this question is beyond the scope of this paper, what we do offer is a first step: a proposal for how expert elicitation can be used to identify important warnings which can later be used to facilitate timely democratic debate. For such debates to be useful, we first need an idea of which developments on the horizon can be meaningfully assessed and influenced, for which it makes sense to draw on public expertise and limited attention. This is precisely what our method aims to provide.

III. Identifying Early Warning Signs

We believe that identifying canaries for transformative AI is a tractable problem and worth investing research effort in today. Engineering and cognitive development present a proof of principle: capabilities are achieved sequentially, meaning that there are often key underlying capabilities which, if attained, unlock progress in many other areas. For example, musical protolanguage is thought to have enabled grammatical competence in the development of language in homo sapiens [50]. AI progress so far has also seen such amplifiers: the use of multi-layered non-linear learning or stochastic gradient descent arguably laid the foundation for unexpectedly fast progress on image recognition, translation and speech recognition [51]. By mapping out the dependencies between different capabilities needed to reach some notion of transformative AI, therefore, we should be able to identify milestones which are particularly important for enabling many others - these are our canaries.

The proposed methodology is intended to be highly adaptable and can be used to identify canaries for a number of important potentially transformative events, such as foundational research breakthroughs or the automation of tasks that affect a wide range of jobs. Many types of indicators could be of interest and classed as canaries, including: algorithmic innovation that supports key cognitive faculties (e.g., natural language understanding); overcoming known technical challenges (such as improving the data efficiency of deep learning algorithms); or improved applicability of AI to economically-relevant tasks (e.g. text summarization).

Given an event for which we wish to identify canaries, our methodology has three essential steps: (1) identifying key milestones towards the event; (2) identifying dependency relations between these milestones; and (3) identifying milestones which underpin many others as canaries. See Fig. 1 for an illustration. We here deliberately refrain from describing the method with too much specificity, because we want to stress the flexibility of our approach, and recognise that there is currently no one-fits-all approach in forecasting. The method will require adaptation to the particular transformative event in question, but each step of this method is suited for such specifications. We outline example adaptations of the method to particular cases.

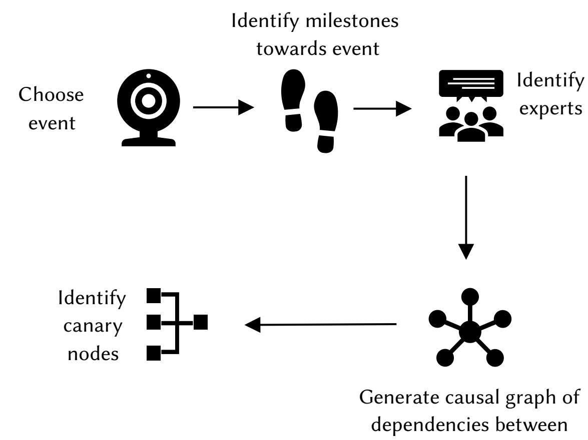


Fig. 1. Illustration of methodological steps to identify canaries of AI progress.

A. Identifying Milestones Via Expert Elicitation

The first step of our methodology involves using traditional approaches in expert elicitation to identify milestones that may be relevant to the transformative event in question. Which experts are selected is crucial to the outcome and reliability of studies in AI forecasting. There are unavoidable limitations of using any form of subjective judgement in forecasting, but these limitations can be minimised by carefully thinking through the group selection. Both the direct expertise of individuals, and how they contribute to the diversity of the overall group, must be considered. See [46] for a discussion of who counts as an expert in forecasting.

Researchers should decide in advance what kinds of expertise are most relevant and must be combined to study the milestones that relate to the transformative event. Milestones might include technical limitations of current methods (e.g. adversarial attacks) and informed speculation about future capabilities (e.g. common sense) that may be important prerequisites to the transformative event. Consulting across a wide range of academic disciplines to order such diverse milestones is important. For example, a cohort of experts identifying and ordering milestones towards HLMI should include not only experts in machine learning and computer science but also cognitive scientists, philosophers, developmental psychologists, evolutionary biologists, or animal cognition experts. Such a group combines expertise on current capabilities in AI, with expertise on key pillars of cognitive development and the order in which cognitive faculties develop in animals. Groups which are diverse (on multiple dimensions) are expected to produce better epistemic outcomes [37], [52].

We encourage the careful design and phrasing of questions to enable participants to make use of their expertise, but refrain from demanding answers that lie outside their area of expertise. For example, asking machine learning researchers directly for milestones towards HLMI does not draw on their expertise. But asking machine learning researchers about the limitations of the methods they use every day; or asking psychologists what human capacities they see lacking in machines today, draws directly on their day-to-day experience. Perceived limitations can be then be transformed into milestones.

There are several different methods available for expert elicitation including surveys, interviews, workshops and focus groups, each with advantages and disadvantages. Interviews provide greater opportunity to tailor questions to the specific expert, but can be time-intensive compared to surveys and reduce the sample size of experts. If possible, some combination of the two may be ideal: using carefully selected semi-structured interviews to elicit initial milestones, followed-up with surveys with a much broader group to validate which milestones are widely accepted as being key.

B. Mapping Causal Relations Between Milestones

The second step of our methodology involves convening experts to identify causal relations between identified milestones: that is, how milestones may underpin, depend on, or affect progress towards other milestones. Experts should be guided in generating directed causal graphs, a type of cognitive map that elicits a person's perceived causal relations between components. Causal graphs use arrows to represent perceived causal relations between nodes, which in this case are milestones [53].

This process primarily focuses on finding out whether or not a relationship exists at all; how precisely this relationship is specified can be adapted to the goals of the study. An arrow from A to B at minimum indicates that progress on A will allow for further progress on B. But this relationship can also be made more precise: in some cases indicating that progress on AI is *necessary* for progress on B, for example. The relationship between nodes may be either linear or non-linear; again this can be specified more precisely if needed or known.

Constructing and debating causal graphs can "help groups to convert tacit knowledge into explicit knowledge" [53]. Causal graphs are used as decision support for individuals or groups, and are often used to solve problems in policy and management involving complex relationships between components in a system by tapping into experts' mental models and intuitions. We therefore suggest that causal graphs are particularly well-suited to eliciting experts' models and assumptions about the relationship between different milestones in AI development.

As a method, causal graphs are highly flexible and can be adapted to the preferred level of detail for a given study: they can be varied in complexity and can be analysed both quantitatively and qualitatively [54], [55]. We neither exclude nor favour quantitative approaches here, due to the complexity and uncertainty of the questions around transformative events. Particularly for very high-level questions, quantitative approaches might not offer much advantage and might communicate a false sense of certainty. In narrower domains where there is more existing evidence, however, quantitative approaches may help to represent differences in the strength of relationships between milestones.

[56] notes that there are no ready-made designs that will fit all studies: design and analysis of causal mapping procedures must be matched to a clear theoretical context and the goal of the study. We highlight a number of different design choices which can be used to adapt the process. As more studies use causal graphs in expert elicitations about AI developments, we can learn from the success of different design choices over time and identify best practices.

[53] stress that interviews or collective brainstorming are the most accepted method for generating the data upon which to analyse causal relations. [57] list heuristics on how to manage the procedure of combining graphs by different participants, or see [58] for a discussion on evaluating different options presented by experts. [59] suggest visual, interactive tools to aid the process. [56] and [60] discuss approaches to analysing graphs and extracting the emergent properties, significant 'core' nodes as well as hierarchical clusters. Core or "potent" nodes are those that relate to many clusters in the graphs and thus have implications for connected nodes. In our proposed methodology, such potent nodes play a central role in pointing to canary milestones. For more detail on the many options on how to generate, analyse and use causal graphs we refer the reader to the volume of [57], or reviews such as [53], [59]. See [55] for an example of applying cognitive mapping to expert views on UK public policies; and [61] for group problem solving with causal graphs.

We propose that identified experts be given instruction in generating either an individual causal graph, after which a mediated discussion between experts generates a shared graph; or that the groups of experts as a whole generates the causal graph via argumentation, visualisations and voting procedures if necessary. As [62] emphasises, any group of experts will have both shared and conflicting assumptions, which causal graphs aim to integrate in a way that approaches greater accuracy than that contained in any single expert viewpoint. The researchers are free to add as much detail to the final maps as required or desired. Each node can be broken into subcomponents or justified with extensive literature reviews.

C. Identifying Canaries

Finally, the resulting causal graphs can be used to identify nodes of particular relevance for progress towards the transformative event in question. This can be a node with a high number of outgoing arrows, i.e. milestones which unlock many others that are prerequisites for the event in question. It can also be a node which functions as a bottleneck - a single dependency node that restricts access to a subsequent highly significant milestone. See Fig. 2 for an illustration. Progress on these milestones can thus represent a 'canary', indicating that further advances in subsequent milestones will become possible and more likely. These canaries can act as early warning signs for potentially rapid and discontinuous progress, or may signal that applications are becoming ready for deployment. Experts identify nodes which unlock or provide a bottleneck for a significant number of other nodes (some amount of discretion from the experts/conveners will be needed to determine what counts as 'significant').

Of course, in some cases generating these causal graphs and using them to identify canaries may be as complicated as a full scientific research project. The difficulty of estimating causal relationships between future technological advances must not be underestimated. However, we believe it to be the case that each individual researcher already does this to some extent, when they chose to prioritise a research project, idea or method over another within a research paradigm. Scientists also debate the most fruitful and promising research avenues and arguably place bets on implicit maps of milestones as they pick a research agenda. The idea is not to generate maps that provide a perfectly accurate indication of warning signs, but to use the wisdom of crowds to make implicit assumptions explicit, creating the best possible estimate of which milestones may provide important indications of future transformative progress.

IV. Using Early Warning Signs

Once identified, canary milestones can immediately help to focus existing efforts in forecasting and anticipatory governance. Given limited resources, early warning signs can direct governance attention to areas of AI progress which are soon likely to impact society and which can be influenced now. For example, if progress in a specific area of NLP (e.g. sentiment analysis) serves as a warning sign for the deployment of more engaging social bots to manipulate voters, policymakers and regulators can monitor or regulate access and research on this research area within NLP.

We can also establish research and policy initiatives to monitor and forecast progress towards canaries. Initiatives might automate the collection, tracking and flagging of new publications relevant to canary capabilities, and build a database of relevant publications. They might use prediction platforms to enable collective forecasting of progress towards canary capabilities. Foundational research can try to validate hypothesised relationships between milestones or illuminate the societal implications of different milestones.

These forecasting and tracking initiatives can be used to improve policy prioritisation more broadly. For example, if we begin to see substantial progress in an area of AI likely to impact jobs in a particular domain, policymakers can begin preparing for potential unemployment in that sector with greater urgency.

However, we believe the value of early warning signs can go further and support us in democratising the development and deployment of AI. Providing opportunities for participation and control over policy is a fundamental part of living in a democratic society. It may be especially important in the case of AI, since its deployment might indeed transform society across many sectors. If AI applications are to bring benefits across such wide-ranging contexts, AI deployment strategies must consider and be directed by the diverse interests found across those sectors. Interests which are underrepresented at technology firms are otherwise likely to bear the negative impacts.

There is currently an information asymmetry between those developing AI and those impacted by it. Citizens need better information about specific developments and impacts which might affect them. Public attention and funding for deliberation processes is not unlimited, so we need to think carefully about which technologies to direct public attention and funding towards. Identifying early warning signs can help address this issue, by focusing the attention of public debate and directing funding towards deliberation practises that centre around technological advancements on the horizon.

We believe early warning signs may be particularly well-suited to feed into participatory technology assessments (PTAs), as introduced earlier. Early warning signs can provide a concrete focal point for citizens and domain experts to collectively discuss concerns. Having identified a specific warning sign, various PTA formats could be suited to consult citizens who are especially likely to be impacted. PTAs come in many forms and a full analysis of which design is best suited to assessing particular AI applications is beyond the scope of this article. But the options are plenty and PTAs show much potential (see section 2). For example, Taiwan has had remarkable success and engagement with an open consultation of citizens on complex technology policy questions [42]. An impact assessment of PTA is not a simple task, but we hypothesise that carefully designed, inclusive PTAs would present a great improvement over how AI is currently developed, deployed and governed. Our suggestion is not limited to governmental bodies. PTAs or other deliberative processes can be run by research groups and private institutions such as AI labs, technology companies and think tanks who are concerned with ensuring AI benefits all of humanity.

V. Method Illustrations

We outline two examples of how this methodology could be adapted and implemented: one focused on identifying warning signs of a particular societal impact, the other on warning signs of progress towards particular technical capabilities. Both these examples pertain to high-level, complex questions about the future development and impacts of AI, meaning our discussion can only begin to illustrate what the process of identifying canaries would look like, and what questions such a process might raise. Since the results are only the suggestions of the authors of this paper, we do not show a full implementation of the method whose value lies in letting a group of experts deliberate. As mentioned previously, the work of generating these causal maps will often be a research project of its own, and we will return later to the question of what level of detail and certainty is needed to make the resulting graphs useful.

A. First Illustration: AI Applications in Voter Manipulation

We show how our method could identify warning signs of the kind of algorithmic progress which could improve the effectiveness of, or reduce the cost of, algorithmic election manipulation. The use of algorithms in attempts to manipulate election results incur great risk for the epistemic resilience of democratic countries [63]–[65].

Manipulations of public opinion by national and commercial actors are not a new phenomenon. [66] details the history of how newly emerging technologies are often used for this purpose. But recent advances in deep learning techniques, as well as the widespread use of social media, have introduced easy and more effective mechanisms for influencing opinions and behaviour. [8] and [67] detail the various ways in which political and commercial actors incur harm to the information ecosystem via the use of algorithms. Manipulators profile voters to identify susceptible targets on social media, distribute micro-targeted advertising, spread misinformation about policies of the opposing candidate and try to convince unwanted voters not to vote. Automation plays a large role in influencing online public discourse. Publications like [68], [69] note that manipulators use both human-run accounts and bots [70] or a combination of the two [71]. Misinformation [72] and targeted messaging [73] can have transformative implications for the resilience of democracies and very possibility of collective action [74], [75].

Despite attempts by national and sub-national actors to apply algorithms to influence elections, their impact so far has been contested [76]. Yet, foreign actors and national political campaigns will continue to have incentives and substantial resources to invest in such campaigns, suggesting their efforts are unlikely to wane in future. We may thus inquire what kinds of technological progress would increase the risk that elections can be successfully manipulated. We can begin this inquiry by identifying what technological barriers currently prevent full-scale election manipulation.

We would identify those technological limitations by drawing on the expertise of actors who are directly affected by these bottlenecks. Those might be managers of online political campaigns and foreign consulting firms (as described in [8]), who specialise in influencing public opinion via social media, or governmental organisations across the world who comment on posts, target individual influencers and operate fake accounts to uphold and spread particular beliefs. People who run such political cyber campaigns have knowledge of what technological bottlenecks still constrain their influence on voter decisions. We recommend running a series of interviews to collect a list of limitations.

This list might include, for example, that the natural language functionality of social bots is a major bottleneck for effective online influence (for the plausibility of this being an important technical factor see [8]). Targeted users often disengage from a chat conversation after detecting that they are exchanging messages with social bots. Low retention time is presumably a bottleneck for further manipulation, which suggests that improvements in natural language processing (NLP) would significantly reduce the cost of manipulation as social bots become more effective.

We will assume, for the purpose of this illustration that NLP were to be identified as a key bottleneck. We would then seek to gather experts (e.g. in a workshop) who can identify and map milestones (or current limitations) in NLP likely to be relevant to improving the functionality of social bots. This will include machine learning experts who specialise in NLP and understand the technical barriers to developing more convincing social bots; as well as experts in developmental

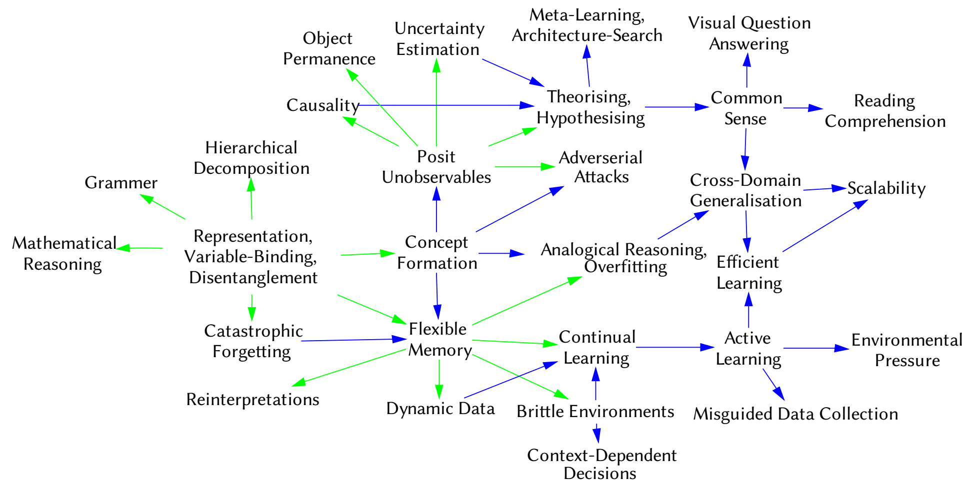


Fig. 2. Cognitive map of dependencies between milestones collected in expert elicitations. Arrows coloured in green signify those milestones that have most outgoing arrows. See appendix for description of each milestone and dependency relations between one 'canary' node and subsequent nodes.

linguistics and evolutionary biology, who can determine suitable benchmarks and the required skills, and who understand the order in which linguistic skills are usually developed in animals.

From these expert elicitation processes we would acquire a list of milestones in NLP which, if achieved, would likely lower the cost and increase the effectiveness of online manipulation. Experts would then order milestones into a causal graph of dependencies. Given the interdisciplinary nature of the question at hand, we suggest in this case that the graph should be directly developed by the whole group. A mediated discussion in a workshop context can help to draw out different connections between milestones and the reasoning behind them, ensuring participants do not make judgements outside their range of expertise. A voting procedure such as majority voting should be used if no consensus can be reached. In a final step, experts can highlight milestone nodes in the final graph which are either marked by many outgoing nodes or are bottlenecks for a series of subsequent nodes that are not accessed by an alternative pathway. These (e.g. sentiment analysis) are our canaries: areas of progress which serve as a warning sign of NLP being applied more effectively in voter manipulation.

Having looked at how this methodology can be used to identify warning signs of a specific societal impact, we next illustrate a different application of the method in which we aim to identify warning signs of a research breakthrough.

B. Second Illustration: High-level Machine intelligence

We use this second example to illustrate in more detail what the process of developing a causal map might look like once initial milestones have been identified, and how canary capabilities can be identified from the map.

We define high-level machine intelligence (HLMI) as an AI system (or collection of AI systems) that performs at the level of an average human adult on key cognitive measures required for economically relevant tasks. We choose to focus on HLMI since it is a milestone which has been the focus of previous forecasting studies [10], [15], and which, despite the ambiguity and uncertain nature of the concepts, is interesting to attempt to examine, because it is likely to precipitate widely transformative societal impacts.

To trial this method, we used interview results from [11]. 25 experts from a diverse set of disciplines (including computer science, cognitive science and neuroscience) were interviewed and asked what they believed to be the main limitations preventing current machine learning methods from achieving the capabilities of HLMI. These limitations can be translated into 'milestones': capabilities experts believe machine learning methods need to achieve on the path to HLMI, i.e. the output of step 1 of our methodology.

Having identified key milestones, step 2 of our methodology involves exploring dependencies between them using causal graphs. We use the software VenSim to illustrate hypothesised relationships between milestones (see Fig. 2). For example, we hypothesise that the ability to formulate, comprehend and manipulate abstract concepts may be an important prerequisite to the ability to account for unobservable phenomena, which is in turn important for reasoning about causality. This map of causal relations and dependencies was constructed by the authors alone, and is therefore far from definitive, but provides a useful illustration of the kind of output this methodology can produce.

Based on this causal map, we can identify three candidates for canary capabilities:

Representations that allow variable-binding and disentanglement: the ability to construct abstract, discrete and disentangled representations of inputs, to allow for efficiency and variable-binding. We hypothesise that this capability underpins several others, including grammar, mathematical reasoning, concept formation, and flexible memory.

Flexible memory: the ability to store, recognise, and re-use memory and knowledge representations. We hypothesise that this ability would unlock many others, including the ability to learn from dynamic data, to learn in a continual fashion, and to update old interpretations of data as new information is acquired.

Positing unobservables: the ability to recognise and use unobservable concepts that are not represented in the visual features of a scene, including numerosity or intentionality.

We might tentatively suggest that these are important capabilities to track progress on from the perspective of anticipating HLMI.

VI. Discussion and Future Directions

As the two illustrative examples show, there are many complexities and challenges involved in putting this method into practice. One particular challenge is that there is likely to be substantial uncertainty in the causal graphs developed. This uncertainty can come in many forms.

Milestones that are not well understood are likely to be composed of several sub-milestones. As more research is produced, the graph will be in need of revision. Some such revisions may include the addition of connections between milestones that were previously not foreseen, which in turn might alter the number of outgoing connections from nodes and turn them into potent nodes, i.e. 'canaries'.

The process of involving a diversity of experts in a multi-stage, collaborative process is designed to reduce this uncertainty by allowing for the identification of nodes and relationships that are widely agreed upon and so more likely to be robust. However, considerable uncertainty will inevitably remain due to the nature of forecasting. The higher the level of abstraction and ambiguity in the events studied (like events such as HLMI, which we use for our illustration) the greater the uncertainty inherent in the map and the less reliable the forecasts will likely be. It will be important to find ways to acknowledge and represent this uncertainty in the maps developed and conclusions drawn from them. This might include marking uncertainties in the graph and taking this into account when identifying and communicating 'canary' nodes.

Given the uncertainty inherent in forecasting, we must consider what kinds of inevitable misjudgements are most important to try to avoid. A precautionary perspective would suggest it is better to slightly overspend resources on monitoring canaries that turn out to be false positives, rather than to miss an opportunity to anticipate significant technological impacts. This suggests we may want to set a low threshold for what should be considered a 'canary' in the final stage of the method.

The uncertainty raises an important question: will it on average be better to have an imperfect, uncertain mapping of milestones rather than none at all? There is some chance that incorrect estimates of 'canaries' could be harmful. An incorrect mapping could focus undue attention on some avenue of AI progress, waste resources or distract from more important issues.

Our view is that it is nonetheless preferable to attempt a prioritisation. The realistic alternative is that anticipatory governance is not attempted or informed by scholars' individual estimates in an ad-hoc manner, which we should expect to be incorrect more often than our collective and structured expert elicitation. How accurate our method is can only be studied by trialling it and tracking its predictions as AI research progresses to confirm or refute the forecasts.

Future studies are likely to face several trade-offs in managing the uncertainty. For example, a large and cognitively diverse expert group may be better placed to develop robust maps eventually, but this may be a much more challenging process than doing it with a smaller, less diverse group -- making the latter a tempting choice (see [45] for a discussion of this trade-off). The study of broad and high-level questions (such as when we might attain HLMI or automate a large percentage of jobs) may be more societally relevant or intellectually motivating, but narrower studies focused on nearer-term, well-defined applications or impacts may be easier to reach certainty on.

A further risk is that this method, intended to identify warning signs so as to give time to debate transformative applications, may inadvertently speed up progress towards AI capabilities and applications. By fostering expert deliberation and mapping milestones, it is likely that important research projects and goals are highlighted and the field's research roadmap is improved. This means our method must be used with caution.

However, we do not believe this is a reason to abandon the approach, since these concerns must be balanced against the benefits of being able to deliberate upon and shape the impacts of AI in advance. In particular, we believe that the process of distilling information from experts in a way that can be communicated to wider society, including those currently underrepresented in debates about the future of AI, is likely to have many more benefits than costs.

The idea that we can identify 'warning signs' for progress assumes that there will be some time lag between progress on milestones, during which anticipatory governance work can take place. Of course, the extent to which this is possible will vary, and in some cases, unlocking a 'canary' capability could lead to very rapid progress on subsequent milestones. Future work could consider how to incorporate assessment of timescales into the causal graphs developed, so that it is easier to identify canaries which warn of future progress while allowing time to prepare.

Future work should also critically consider what constitutes relevant 'expertise' for the task of identifying canaries, and further explore ways to effectively integrate expert knowledge with the values and perspectives of diverse publics. Our method finds a role for the expert situated in a larger democratic process of anticipating and regulating emerging technologies. Expert judgement can thereby be beneficial to wider participation. However, processes that allow more interaction between experts and citizens could be even more effective. One limitation of the method presented in this paper is that it requires one to have already identified a particular transformative event of concern, but does not provide guidance on how to identify and prioritise between events. It may be valuable to consider how citizens that are impacted by technology can play a role in identifying initial areas of concern, which can then feed into this process of expert elicitation to address the concerns.

VII. Conclusion

We have presented a flexible method for identifying early warning signs, or 'canaries' in AI progress. Once identified, these canaries can provide focal points for anticipatory governance efforts, and can form the basis for meaningful participatory processes enabling citizens to steer AI developments and their impacts. Future work must now test this method by putting it into practice, which will more clearly reveal both benefits and limitations. Our artificial canaries offer a chance for forward-looking, democratic assessments of transformative technologies.

Appendix

It is worth noting there are apparent similarities and relationships between many of these milestones. For example, representation: the ability to learn abstract representations of the environment, seems closely related to variable binding: the ability to formulate place-holder concepts. The ability to apply learning from one task to another, cross-domain generalisation, seems closely related to analogical reasoning. Further progress in research will tell which of these are clearly separate milestones or more closely related notions.

Flexible memory, as described by experts in our sample, is the ability to recognize and store reusable information, in a format that is flexible so that it can be retrieved and updated when new knowledge is gained. We explain the reasoning behind the labelled arrows in Fig. 2 (see Fig. 3):

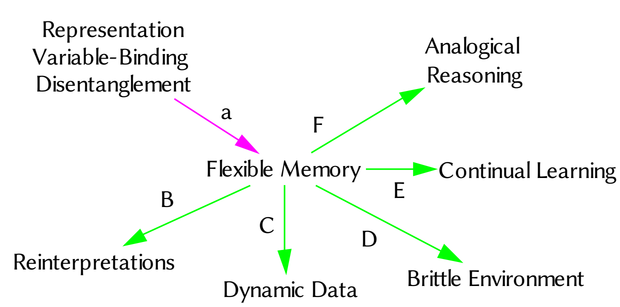


Fig. 3. Extract of Fig. 2, showing one candidate canary capability.

- (a): compact representations are a prerequisite for flexible memory since storing high-dimensional input in memory requires compressed, efficient and thus abstract representations.
- (B): the ability to reinterpret data in light of new information likely requires flexible memory, since it requires the ability to retrieve and alter previously stored information.
- (C) and (E): to make use of dynamic and changing data input, and to learn continuously over time, an agent must be able to store, correctly retrieve and modify previous data as new data comes in.
- (D): in order to plan and execute strategies in brittle environments with long delays between actions and rewards, an agent must be able to store memories of past actions and rewards, but easily retrieve this information and continually update its best guess about how to obtain rewards in the environment.
- (F): analogical reasoning involves comparing abstract representations, which requires forming, recognising, and retrieving representations of earlier observations.

Progress in flexible memory therefore seems likely to unlock or enable many other capabilities important for HLMI, especially those crucial for applying AI systems in real environments and more complex tasks. These initial hypotheses should be validated and explored in more depth by a wider range of experts.

Acknowledgements

We thank reviewers for their particularly detailed comments and engagement with this paper, the scholars at the Leverhulme Centre for the Future of Intelligence for fruitful discussions after our presentation,

TABLE I. Limitations of Deep Learning As Perceived and Named by Experts Found in [11]

Causal reasoning: the ability to detect and generalise from causal relations in data.	**Common sense:** having a set of background beliefs or assumptions which are useful across domains and tasks.
Meta-learning: the ability to learn how to best learn in each domain.	**Architecture search**: the ability to automatically choose the best architecture of a neural network for a task.
Hierarchical decomposition: the ability to decompose tasks and objects into smaller and hierarchical sub-components.	**Cross-domain generalization**: the ability to apply learning from one task or domain to another.
Representation: the ability to learn abstract representations of the environment for efficient learning and generalisation.	**Variable binding:** the ability to attach symbols to learned representations, enabling generalisation and re-use.
Disentanglement: the ability to understand the components and composition of observations, and recombine and recognise them in different contexts.	**Analogical reasoning:** the ability to detect abstract similarity across domains, enabling learning and generalisation.
Concept formation: the ability to formulate, manipulate and comprehend abstract concepts.	**Object permanence:** the ability to represent objects as consistently existing even when out of sight.
Grammar: the ability to construct and decompose sentences according to correct grammatical rules.	**Reading comprehension:** the ability to detect narratives, semantic context, themes and relations between characters in long texts or stories.
Mathematical reasoning: the ability to develop, identify and search mathematical proofs and follow logical deduction in reasoning.	**Visual question answering:** the ability to answer open-ended questions about the content and interpretation of an image.
Uncertainty estimation: the ability to represent and consider different types of uncertainty.	**Positing unobservables:** the ability to account for unobservable phenomena, particularly in representing and navigating environments.
Reinterpretation: the ability to partially re-categorise, re-assign or reinterpret data in light of new information without retraining from scratch.	**Theorising and hypothesising:** the ability to propose theories and testable hypotheses, understand the difference between theory and reality, and the impact of data on theories.
Flexible memory: the ability to store, recognise and retrieve knowledge so that it can be used in new environments and tasks.	**Efficient learning**: the ability to learn efficiently from small amounts of data.
Interpretability: the ability for humans to interpret internal network dynamics so that researchers can manipulate network dynamics.	**Continual learning:** the ability to learn continuously as new data is acquired.
Active learning: the ability to learn and explore in self-directed ways.	**Learning from inaccessible data:** the ability to learn in domains where data is missing, difficult or expensive to acquire.
Learning from dynamic data: the ability to learn from a continually changing stream of data.	**Navigating brittle environments:** the ability to navigate irregular, and complex environments which lack clear reward signals and short feedback loops.
Generating valuation functions: the ability to generate new valuation functions immediately from scratch to follow newly-given rules.	**Scalability:** the ability to scale up learning to deal with new features without needing disproportionately more data, model parameters, and computational power.
Learning in simulation: the ability to learn all relevant experience from a simulated environment.	**Metric identification:** the ability to identify appropriate metrics of success for complex tasks, such that optimising for the measured quantity accomplishes the task in the way intended.
Conscious perception: the ability to experience the world from a first-person perspective.	**Context-sensitive decision making:** the ability to adapt decision-making strategies to the needs and constraints of a given time or context.

as well as the attendees of the workshop Evaluating Progress in AI at the European Conference on AI (Aug 2020) for recognizing the potential of this work. We particularly thank Carolyn Ashurst and Luke Kemp for their efforts and commentary on our drafts.

References

[1] K. Crawford *et al.*, 'AI Now Report 2019', *AI 2019 Report*, p. 100, 2019.
[2] S. Russell, *Human Compatible.* Viking Press, 2019.
[3] C. Cath, S. Wachter, B. Mittelstadt, M. Taddeo, and L. Floridi, 'Artificial Intelligence and the "Good Society": the US, EU, and UK approach', *Sci. Eng. Ethics*, vol. 24, no. 2, pp. 505–528, Apr. 2018, doi: 10.1007/s11948-017-9901-7.
[4] J. Whittlestone, R. Nyrup, A. Alexandrova, K. Dihal, and S. Cave, 'Ethical and societal implications of algorithms, data, and artificial intelligence: a roadmap for research', *2019*, p. 59.
[5] Y. K. Dwivedi *et al.*, 'Artificial Intelligence (AI): Multidisciplinary perspectives on emerging challenges, opportunities, and agenda for research, practice and policy', *Int. J. Inf. Manag.*, p. 101994, Aug. 2019, doi: 10.1016/j.ijinfomgt.2019.08.002.
[6] R. Gruetzemacher and J. Whittlestone, 'The Transformative Potential of Artificial Intelligence', *ArXiv191200747 Cs*, Sep. 2020, Accessed: Jan. 09, 2021. [Online]. Available: http://arxiv.org/abs/1912.00747.
[7] M. Brundage *et al.*, 'The Malicious Use of Artificial Intelligence: Forecasting, Prevention, and Mitigation', *ArXiv180207228 Cs*, Feb. 2018, Accessed: Jan. 15, 2021. [Online]. Available: http://arxiv.org/abs/1802.07228.
[8] P. Howard, *Lie Machines, How to Save Democracy from Troll Armies, Deceitful Robots, Junk News Operations, and Political Operatives.* Yale: Yale University Press, 2020.
[9] C. B. Frey and M. A. Osborne, 'The future of employment: How susceptible are jobs to computerisation?', *Technol. Forecast. Soc. Change*, vol. 114, pp. 254–280, Jan. 2017, doi: 10.1016/j.techfore.2016.08.019.
[10] K. Grace, J. Salvatier, A. Dafoe, B. Zhang, and O. Evans, 'Viewpoint: When Will AI Exceed Human Performance? Evidence from AI Experts', *J. Artif. Intell. Res.*, vol. 62, pp. 729–754, Jul. 2018, doi: 10.1613/jair.1.11222.
[11] C. Z. Cremer, 'Deep Limitations? Examining Expert Disagreement over Deep Learning', *Prog. Artif. Intell. Springer*, to be published 2021.
[12] D. Collingridge, *The social control of technology.* London: Frances Pinter, 1980.
[13] O. Etzioni, 'How to know if artificial intelligence is about to destroy civilization', *MIT Technology Review.* https://www.technologyreview.com/s/615264/artificial-intelligence-destroy-civilization-canaries-robot-overlords-take-over-world-ai/ (accessed Mar. 12, 2020).
[14] A. Dafoe, 'The academics preparing for the possibility that AI will destabilise global politics', *80,000 Hours*, 2018. https://80000hours.org/podcast/episodes/allan-dafoe-politics-of-ai/ (accessed Jan. 15, 2021).
[15] V. C. Müller and N. Bostrom, 'Future Progress in Artificial Intelligence: A Survey of Expert Opinion', in *Fundamental Issues of Artificial Intelligence*, V. C. Müller, Ed. Cham: Springer International Publishing, 2016, pp. 555–572.
[16] S. D. Baum, B. Goertzel, and T. G. Goertzel, 'How long until human-level AI? Results from an expert assessment', *Technol. Forecast. Soc. Change*, vol. 78, no. 1, pp. 185–195, Jan. 2011, doi: 10.1016/j.techfore.2010.09.006.
[17] S. Beard, T. Rowe, and J. Fox, 'An analysis and evaluation of methods currently used to quantify the likelihood of existential hazards', *Futures*, vol. 115, p. 102469, Jan. 2020, doi: 10.1016/j.futures.2019.102469.
[18] P. E. Tetlock and D. Gardner, *Superforecasting: the art and science of prediction*, First edition. New York: Crown Publishers, 2015.
[19] N. Benaich and I. Hogarth, 'State of AI Report 2020', 2020. https://www.stateof.ai/ (accessed Jan. 15, 2021).
[20] P. Eckersley and Y. Nasser, 'AI Progress Measurement', *Electronic Frontier Foundation*, Jun. 12, 2017. https://www.eff.org/ai/metrics (accessed Jan. 15, 2021).
[21] 'Papers with Code', Available at: https://paperswithcode.com (accessed Feb. 08, 2021).
[22] R. Perrault *et al.*, 'The AI Index 2019 Annual Report', *AI Index Steer. Comm. Hum.-Centered AI Inst. Stanf. Univ. Stanf. CA*, 2019.
[23] Gruetzemacher, 'A Holistic Framework for Forecasting Transformative AI', *Big Data Cogn. Comput.*, vol. 3, no. 3, p. 35, Jun. 2019, doi: 10.3390/bdcc3030035.
[24] H. A. Linstone and M. Turoff, *The delphi method.* Addison-Wesley Reading, MA, 1975.
[25] S. M. West, M. Whittaker, and K. Crawford, 'Discriminating Systems: Gender, Race and Power in AI', AI Now Institute, 2019. [Online]. Available: Retrieved from https://ainowinstitute.org/discriminatingsystems.html.
[26] P. Nemitz and M. Pfeffer, *Prinzip Mensch - Macht, Freiheit und Demokratie im Zeitalter der Künstlichen Intelligenz.* Verlag J.H.W. Dietz Nachf., 2020.
[27] M. Ipsos, 'Public views of Machine Learning: Findings from public research and engagement conducted on behalf of the Royal Society', THe Royal Society, 2017. [Online]. Available: https://royalsociety.org/-/media/policy/projects/machine-learning/publications/public-views-of-machine-learning-ipsos-mori.pdf.
[28] The RSA, 'Artificial Intelligence: Real Public Engagement.', oyal Society for the encouragement of Arts, Manufactures and Commerce, London, 2018.
[29] T. Cohen, J. Stilgoe, and C. Cavoli, 'Reframing the governance of automotive automation: insights from UK stakeholder workshops', *J. Responsible Innov.*, vol. 5, no. 3, pp. 257–279, Sep. 2018, doi: 10.1080/23299460.2018.1495030.
[30] M. Lengwiler, 'Participatory Approaches in Science and Technology: Historical Origins and Current Practices in Critical Perspective', *Sci. Technol. Hum. Values*, vol. 33, no. 2, pp. 186–200, Mar. 2008, doi: 10.1177/0162243907311262.
[31] M. Rask, 'The tragedy of citizen deliberation – two cases of participatory technology assessment', *Technol. Anal. Strateg. Manag.*, vol. 25, no. 1, pp. 39–55, Jan. 2013, doi: 10.1080/09537325.2012.751012.
[32] J. Chilvers, 'Deliberating Competence: Theoretical and Practitioner Perspectives on Effective Participatory Appraisal Practice', *Sci. Technol. Hum. Values*, vol. 33, no. 2, pp. 155–185, Mar. 2008, doi: 10.1177/0162243907307594.
[33] G. Abels, 'Participatory Technology Assessment And The "Institutional Void": Investigating Democratic Theory And Representative Politics" published on 01 Jan 2010 by Brill.', in *Democratic Transgressions of Law*, vol. 112, Brill, 2010, pp. 237–268.
[34] P. Biegelbauer and A. Loeber, 'The Challenge of Citizen Participation to Democracy', *Inst. Für Höhere Stud. - Inst. Adv. Stud. IHS*, p. 46, 2010.
[35] G. Rowe and L. J. Frewer, 'A Typology of Public Engagement Mechanisms', *Sci. Technol. Hum. Values*, vol. 30, no. 2, pp. 251–290, Apr. 2005, doi: 10.1177/0162243904271724.
[36] L. Hong and S. E. Page, 'Groups of diverse problem solvers can outperform groups of high-ability problem solvers', *Proc. Natl. Acad. Sci.*, vol. 101, no. 46, pp. 16385–16389, Nov. 2004, doi: 10.1073/pnas.0403723101.
[37] H. Landemore, *Democratic Reason.* Princeton: Princeton University Press, 2017.
[38] S. Joss and S. Bellucci, *Participatory Technology Assessment: European Perspectives.* London: Center for the Study of Democracy, 2002.
[39] Y. Zhao, C. Fautz, L. Hennen, K. R. Srinivas, and Q. Li, 'Public Engagement in the Governance of Science and Technology', in *Science and Technology Governance and Ethics: A Global Perspective from Europe, India and China*, M. Ladikas, S. Chaturvedi, Y. Zhao, and D. Stemerding, Eds. Cham: Springer International Publishing, 2015, pp. 39–51.
[40] M. T. Rask *et al.*, *Public Participation, Science and Society: Tools for Dynamic and Responsible Governance of Research and Innovation.* Routledge - Taylor & Francis Group, 2018.
[41] J. Burgess and J. Chilvers, 'Upping the ante: a conceptual framework for designing and evaluating participatory technology assessments', *Sci. Public Policy*, vol. 33, no. 10, pp. 713–728, Dec. 2006, doi: 10.3152/147154306781778551.
[42] Y. T. Hsiao, S.-Y. Lin, A. Tang, D. Narayanan, and C. Sarahe, 'vTaiwan: An Empirical Study of Open Consultation Process in Taiwan', SocArXiv, preprint, Jul. 2018. doi: 10.31235/osf.io/xyhft.
[43] J. Hansen, 'Operationalising the public in participatory technology assessment: A framework for comparison applied to three cases', *Sci. Public Policy*, vol. 33, no. 8, pp. 571–584, Oct. 2006, doi: 10.3152/147154306781778678.
[44] T.-P. Ertiö, P. Tuominen, and M. Rask, 'Turning Ideas into Proposals: A Case for Blended Participation During the Participatory Budgeting Trial in Helsinki', in *Electronic Participation: ePart 2019*, Jul. 2019, pp. 15–25,

doi: 10.1007/978-3-030-27397-2_2.

[45] M. Rask, 'Foresight — balancing between increasing variety and productive convergence', *Technol. Forecast. Soc. Change - TECHNOL FORECAST SOC CHANGE*, vol. 75, pp. 1157–1175, Oct. 2008, doi: 10.1016/j.techfore.2007.12.002.

[46] S. Mauksch, H. A. von der Gracht, and T. J. Gordon, 'Who is an expert for foresight? A review of identification methods', *Technol. Forecast. Soc. Change*, vol. 154, p. 119982, May 2020, doi: 10.1016/j.techfore.2020.119982.

[47] J. Saldivar, C. Parra, M. Alcaraz, R. Arteta, and L. Cernuzzi, 'Civic Technology for Social Innovation: A Systematic Literature Review', *Comput. Support. Coop. Work CSCW*, vol. 28, no. 1–2, pp. 169–207, Apr. 2019, doi: 10.1007/s10606-018-9311-7.

[48] T. Kariotis and J. Darakhshan, 'Fighting Back Algocracy: The need for new participatory approaches to technology assessment', in *Proceedings of the 16th Participatory Design Conference 2020 - Participation(s) Otherwise - Volume 2*, Manizales Colombia, Jun. 2020, pp. 148–153, doi: 10.1145/3384772.3385151.

[49] M. Whitman, C. Hsiang, and K. Roark, 'Potential for participatory big data ethics and algorithm design: a scoping mapping review', in *Proceedings of the 15th Participatory Design Conference: Short Papers, Situated Actions, Workshops and Tutorial - Volume 2*, New York, NY, USA, Aug. 2018, pp. 1–6, doi: 10.1145/3210604.3210644.

[50] C. Buckner and K. Yang, 'Mating dances and the evolution of language: What's the next step?', *Biol. Philos.*, vol. 32, 2017, doi: 10.1007/s10539-017-9605-z.

[51] Y. LeCun, Y. Bengio, and G. Hinton, 'Deep learning', *Nature*, vol. 521, no. 7553, pp. 436–444, May 2015, doi: 10.1038/nature14539.

[52] S. E. Page, *The Difference: How the Power of Diversity Creates Better Groups, Firms, Schools and Societies.* Princeton: Princeton University Press, 2008.

[53] A. J. Scavarda, T. Bouzdine-Chameeva, S. M. Goldstein, J. M. Hays, and A. V. Hill, 'A Review of the Causal Mapping Practice and Research Literature', in *Abstract number: 002-0256*, Cancun, Mexico, 2004, p. 21.

[54] L. Markíczy, and J. Goldberg, 'A method for eliciting and comparing causal maps', *J. Manag.*, vol. 21, no. 2, pp. 305–333, Jan. 1995, doi: 10.1016/0149-2063(95)90060-8.

[55] C. Eden and F. Ackermann, 'Cognitive mapping expert views for policy analysis in the public sector', *Eur. J. Oper. Res.*, vol. 152, no. 3, pp. 615–630, Feb. 2004, doi: 10.1016/S0377-2217(03)00061-4.

[56] C. Eden, 'ON THE NATURE OF COGNITIVE MAPS', 1992, doi: 10.1111/J.1467-6486.1992.TB00664.X.

[57] F. Ackerman, J. Bryson, and C. Eden, *Visible Thinking, Unlocking Causal Mapping for Practical Business Results.* John Wiley & Sons, 2004.

[58] G. Montibeller and V. Belton, 'Causal maps and the evaluation of decision options—a review', *J. Oper. Res. Soc.*, vol. 57, no. 7, pp. 779–791, Jul. 2006, doi: 10.1057/palgrave.jors.2602214.

[59] A. J. Scavarda, T. Bouzdine-Chameeva, S. M. Goldstein, J. M. Hays, and A. V. Hill, 'A Methodology for Constructing Collective Causal Maps*', *Decis. Sci.*, vol. 37, no. 2, pp. 263–283, May 2006, doi: 10.1111/j.1540-5915.2006.00124.x.

[60] C. Eden, F. Ackermann, and S. Cropper, 'The Analysis of Cause Maps', *J. Manag. Stud.*, vol. 29, no. 3, pp. 309–324, 1992, doi: https://doi.org/10.1111/j.1467-6486.1992.tb00667.x.

[61] F. Ackermann and C. Eden, 'Using Causal Mapping with Group Support Systems to Elicit an Understanding of Failure in Complex Projects: Some Implications for Organizational Research', *Group Decis. Negot.*, vol. 14, no. 5, pp. 355–376, Sep. 2005, doi: 10.1007/s10726-005-8917-6.

[62] C. Eden, F. Ackermann, J. Bryson, G. Richardson, D. Andersen, and C. Finn, 'Integrating Modes of Policy Analysis and Strategic Management Practice: Requisite Elements and Dilemmas', p. 13, 2009.

[63] L.-M. Neudert and P. Howard, 'Ready to vote: elections, technology and political campaigning in the United Kingdom', Oxford Technology and Elections Commission, Report, Oct. 2019. Accessed: Jan. 11, 2021. [Online]. Available: https://apo.org.au/node/263976.

[64] G. Bolsover and P. Howard, 'Computational Propaganda and Political Big Data: Moving Toward a More Critical Research Agenda', *Big Data*, vol. 5, no. 4, pp. 273–276, Dec. 2017, doi: czz.

[65] M. J. Mazarr, R. Bauer, A. Casey, S. Heintz, and L. J. Matthews, 'The Emerging Risk of Virtual Societal Warfare: Social Manipulation in a Changing Information Environment', Oct. 2019, Accessed: Jan. 14, 2021. [Online]. Available: https://www.rand.org/pubs/research_reports/RR2714.html.

[66] T. Wu, *The Attention Merchants: From the Daily Newspaper to Social Media, How Our Time and Attention is Harvested and Sold.* London: Antlantic Books, 2017.

[67] K. Starbird, 'Disinformation's spread: bots, trolls and all of us', *Nature*, vol. 571, no. 7766, pp. 449–450, Jul. 2019.

[68] R. Gorwa and D. Guilbeault, 'Unpacking the Social Media Bot: A Typology to Guide Research and Policy', *Policy Internet*, vol. 12, no. 2, pp. 225–248, Jun. 2020, doi: 10.1002/poi3.184.

[69] E. Ferrara, 'Disinformation and Social Bot Operations in the Run Up to the 2017 French Presidential Election', Social Science Research Network, Rochester, NY, SSRN Scholarly Paper ID 2995809, Jun. 2017. doi: 10.2139/ssrn.2995809.

[70] C. Shao, G. L. Ciampaglia, O. Varol, K.-C. Yang, A. Flammini, and F. Menczer, 'The spread of low-credibility content by social bots', *Nat. Commun.*, vol. 9, no. 1, Art. no. 1, Nov. 2018, doi: 10.1038/s41467-018-06930-7.

[71] P. N. Howard, S. Woolley, and R. Calo, 'Algorithms, bots, and political communication in the US 2016 election: The challenge of automated political communication for election law and administration', *J. Inf. Technol. Polit.*, vol. 15, no. 2, pp. 81–93, Apr. 2018, doi: 10.1080/19331681.2018.1448735.

[72] M. Chessen, 'The MADCOM Future: How Artificial Intelligence Will Enhance Computational Propaganda, Reprogram Human Culture, and Threaten Democracy... and What can be Done About It.', in *Artificial Intelligence Safety and Security*, Chapman and Hall/CRC Press, 2018, pp. 127–144.

[73] K. Kertysova, 'Artificial Intelligence and Disinformation: How AI Changes the Way Disinformation is Produced, Disseminated, and Can Be Countered', 2018, doi: 10.1163/18750230-02901005.

[74] J. Brainard and P. R. Hunter, 'Misinformation making a disease outbreak worse: outcomes compared for influenza, monkeypox, and norovirus', *SIMULATION*, vol. 96, no. 4, pp. 365–374, Apr. 2020, doi: 10.1177/0037549719885021.

[75] E. Seger, S. Avin, G. Pearson, M. Briers, S. O Heigeartaigh, and H. Bacon, 'Tackling threats to informed decision-making in democratic societies: Promoting epistemic security in a technologically-advanced world', Allan Turing Institute, CSER, dstl, 2020. Accessed: Jan. 15, 2021. [Online]. Available: https://www.turing.ac.uk/sites/default/files/2020-10/epistemic-security-report_final.pdf.

[76] K. H. Jamieson, *Cyberwar: How Russian Hackers and Trolls Helped Elect a President: What We Don't, Can't, and Do Know.* Oxford University Press, 2020.

Carla Zoe Cremer

Carla Zoe is a Research Scholar at the Future of Humanity Institute at the University of Oxford and a Research Affiliate at the Centre for the Study of Existential Risk at the University of Cambridge. Her background is in neurobiology, acquired at Ludwig-Maximilian University in Munich and ETH Zurich. She works on comparative cognition, the limitations of deep learning, and on estimating tail-risks of emerging technologies.

Jess Whittlestone

Jess Whittlestone is a Senior Research Associate at Centre for the Study of Existential Risk and the Leverhulme Centre for the Future of Intelligence at the University of Cambridge. She works on various aspects of AI ethics and policy, with a particular focus on what we can do today to ensure AI is safe and beneficial in the long-term. She holds a PhD in Behavioural Science from the University of Warwick and a degree in Mathematics and Philosophy from Oxford University.

Efficient and Robust Model Benchmarks with Item Response Theory and Adaptive Testing

Hao Song, Peter Flach*

University of Bristol (United Kingdom)

Received 21 January 2021 | Accepted 10 February 2021 | Published 23 February 2021

Abstract

Progress in predictive machine learning is typically measured on the basis of performance comparisons on benchmark datasets. Traditionally these kinds of empirical evaluation are carried out on large numbers of datasets, but this is becoming increasingly hard due to computational requirements and the often large number of alternative methods to compare against. In this paper we investigate adaptive approaches to achieve better efficiency on model benchmarking. For a large collection of datasets, rather than training and testing a given approach on every individual dataset, we seek methods that allow us to pick only a few representative datasets to quantify the model's goodness, from which to extrapolate to performance on other datasets. To this end, we adapt existing approaches from psychometrics: specifically, Item Response Theory and Adaptive Testing. Both are well-founded frameworks designed for educational tests. We propose certain modifications following the requirements of machine learning experiments, and present experimental results to validate the approach.

Keywords

Item Response Theory, Adaptive Testing, Model Evaluation, Benchmarks.

DOI: 10.9781/ijimai.2021.02.009

I. Introduction

Thanks to the recent popularity of machine learning and artificial intelligence techniques, researchers and practitioners now have a very considerable choice of models and learning algorithms when facing a given task. However, as choices come with deliberations, selecting an appropriate model is also becoming more challenging. Traditionally, model selection involves two steps:

1. Gather related work and explore existing comparisons.
2. Prepare a shortlist and run the models within the target task for more detailed and local comparisons.

However, given the number of research areas and datasets available now, few research papers provide a fully comprehensive benchmark on all related datasets. There is also a considerable risk of confirmation bias. People tend to focus on datasets where the proposed approach leads to improvements, making it even harder to obtain a fair and comprehensive view of different methods [1]. Regarding the second step above, given the rapid rise in computational demands among recent approaches, it is often impractical to simultaneously cover a broad set of experiments.

Despite the emergence of platforms such as OpenML [2] that aim to collect experimental results via standard configurations, it still requires relatively large numbers of new experiments once a novel task/method is introduced. These additional experiments could take a non-trivial time to run given OpenML's crowdsourcing nature. Although certain research areas and methods can come with formal guarantees, these only cover limited scenarios and most practices in the field still rely on experiments and empirical evaluations. Therefore, in this paper, we consider the problem of efficiently obtaining fair and reliable benchmarks on a set of models and datasets.

To get started, in this paper we focus on the typical setting of predictive machine learning. We assume some labelled datasets and several classifiers can be trained and tested on any possible combination. An experiment includes a set of evaluation measures, and we read the measurements to reflect the performance on any given model-dataset pair. We want to investigate approaches that can accurately quantify performance on a large variety of models and datasets while limiting the overall computational costs. For this purpose, we refer to the fields of psychometrics and testing in education and borrow the frameworks of Item Response Theory [3], [4] and Computerised Adaptive Testing [5], [6]. Both frameworks assume the same scenario, where a participant is assigned several items to answer (response). A typical example would be educational tests, where each student is a participant, and each test question is an item.

Item Response Theory (IRT) is a collection of probabilistic models built on the participants' responses to the items. In IRT, a representative setting assumes each participant has an ability parameter, and each item has a difficulty parameter. Both sets of parameters can affect the collected responses, but are not directly observable. IRT aims to learn these parameters from the collected responses, after which we can quantitatively interpret each participant's level and item with the parameter magnitudes. We can further use these parameters to perform statistical transformations, such as to rank students on their estimated abilities (rather than ranking them on the observed responses).

Computerised Adaptive Testing (CAT) is a framework further built on top of IRT. IRT expects the availability of many responses from different participant-item combinations. Sometimes a specific combination might not be necessary. For instance, it is less informative to give a more challenging question to a student who just failed to

* Corresponding author.

E-mail address: peter.flach@bristol.ac.uk

answer a much simpler one. The purpose of the CAT is to adaptively select the items according to previous responses so that the total number of items used in the test is kept at a relatively low level. Our work's central hypothesis is that IRT and CAT can be used – with some essential modifications – for benchmarking machine learning models.

This paper focuses on predictive machine learning tasks, where every dataset is an item, and each model class is a participant. We aim to investigate the possibilities of using the IRT and CAT frameworks to obtain accurate benchmarks on each model-dataset combination while limiting the total number of experiments. The main contributions of the paper include: (1) We adapt and modify the IRT and CAT frameworks to incorporate the need for model benchmarks as in machine learning. (2) We establish a set of experiments to investigate and compare a set of IRT and CAT approaches in a machine learning context, and demonstrate the benefits of having adaptive testing in typical predictive tasks. The outline of the paper is as follows. We first give a brief introduction of the existing approaches from both IRT and CAT in section II, following proposed modifications on them for our benchmarking requirements in section III. Experiments on some standard models and datasets will be presented in section IV, and finally, additional discussions and insights are provided in section V.

II. Background

This section gives an overview of basic concepts and methods in IRT and CAT and introduces necessary notation. We also discuss some existing work on applying IRT in machine learning.

A. Item Response Theory

Item Response Theory refers to a collection of methods that measure individual abilities, item (question) difficulties, and other potential attributes by checking individual responses to a set of items. IRT models are probabilistic models with latent variables, where the responses are the observations, and abilities, difficulties and other related parameters are the latent variables to be estimated. IRT models are of particular use when the responses distribute differently according to different items, and only averaging the responses does not adequately represent a participant's ability. IRT is therefore particularly suitable for analysing the results of educational exams and many physiological tests. When it comes to machine learning experiments, where different datasets typically come with varying baseline performance, IRT provides an opportunity to treat the performance gains among these datasets fairly.

In the following, we introduce two conventional IRT models and discuss their parameter settings and applications. We use θ to denote the parameter of a particular candidate, and δ and a for item parameters (some IRT models have more than two item parameters). The notation R denotes the random variable of the responses.

1. Two-parameter Logistic Model

The two-parameter (per item) logistic model is defined as follows:

$$R \mid \Theta = \theta, \Delta = \delta, A = a \sim \mathbf{Bernoulli}(\mu_{(\theta,\delta,a)}) \tag{1}$$

$$\mu_{(\theta,\delta,a)} = \frac{1}{1 + \exp\big(-a \cdot (\theta - \delta)\big)} \tag{2}$$

from which expectation and variance of R are obtained as follows:

$$\mathrm{E}\Big[R \mid \Theta = \theta, \Delta = \delta, A = a\Big] = \mu_{(\theta,\delta,a)} \tag{3}$$

$$\mathrm{Var}\Big[R \mid \Theta = \theta, \Delta = \delta, A = a\Big] = \mu_{(\theta,\delta,a)} \cdot \Big(1 - \mu_{(\theta,\delta,a)}\Big) \tag{4}$$

Here $R \in \{0, 1\}$ is a binary response variable indicating whether a particular individual answered a particular item correctly, $\theta \in \mathbb{R}$ is the individual's ability parameter, and $\delta \in \mathbb{R}$ is the item's difficulty parameter. The two-parameter logistic model additionally has a discrimination parameter a on the items, which controls how rapidly the response distribution changes when candidate ability varies. Therefore, assume we have two participants with different abilities, an item with high discrimination tends to have higher differences between the responses from the two participants, respectively. Positive discrimination indicates that higher ability leads to higher expectation on the responses, and vice versa. Besides the two-parameter setting, there also exists a few variants on Logistic IRT. The three-parameter setting further adds a guessing parameter which lower-bounds the response expectation. A multinomial setting can also be adapted to support categorical responses beyond the binary setting.

2. Three-parameter Beta Model

While the logistic model supports binary responses, a recently proposed IRT model extends the support to continuous response [7]:

$$R \mid \Theta = \theta, \Delta = \delta, A = a \sim \mathbf{Beta}(\alpha_{(\theta,\delta,a)}, \beta_{(\theta,\delta,a)}) \tag{5}$$

$$\alpha_{(\theta,\delta,a)} = \left(\frac{\theta}{\delta}\right)^a \tag{6}$$

$$\beta_{(\theta,\delta,a)} = \left(\frac{1-\theta}{1-\delta}\right)^a \tag{7}$$

It can then be shown that:

$$\mathrm{E}\Big[R \mid \Theta = \theta, \Delta = \delta, A = a\Big] = \frac{\alpha_{(\theta,\delta,a)}}{\alpha_{(\theta,\delta,a)} + \beta_{(\theta,\delta,a)}} \tag{8}$$

$$\mathrm{Var}\Big[R \mid \Theta = \theta, \Delta = \delta, A = a\Big] = \frac{\alpha_{(\theta,\delta,a)} \cdot \beta_{(\theta,\delta,a)}}{(\alpha_{(\theta,\delta,a)} + \beta_{(\theta,\delta,a)})^2 \cdot (\alpha_{(\theta,\delta,a)} + \beta_{(\theta,\delta,a)} + 1)} \tag{9}$$

Here $R \in [0, 1]$ is a bounded continuous response, $\theta \in [0, 1]$, $\delta \in [0, 1]$ and $a \in \mathbb{R}$. Similar to the logistic case, a is a discrimination parameter that controls the change rate of responses according to the ratio between ability and discrimination. In addition to supporting continuous responses, one advantage is that the item characteristic curve of the three-parameter Beta model can have a variety of shapes beyond the usual sigmoid shape (for $a > 1$), including inverse-sigmoid ($0 < a < 1$), parabolic ($a = 1$) and even identity ($a = 1, \delta = 1/2$). For the cases with $a < 0$, the Beta model can give a symmetry shape to the cases with $a > 0$ with respect to the vertical line of $r = 0.5$.

3. Estimation of IRT Parameters

The estimation of IRT parameters proceeds as follows. We assume to have a bag of L items, denoted as $\mathbb{D} = \{1, \ldots, L\}$, and a bag of M participants, denoted as $\mathbb{F} = \{1, \ldots, M\}$. With a given experiment protocol, we can collect a set of N item-participant-response tuples, denoted as $\{(d_1, f_1, r_1), \ldots, (d_N, f_N, r_N)\}$. Here $d_i \in \mathbb{D}$, $f_i \in \mathbb{F}$ represents a particular item / participant respectively, and r_i is the corresponding response. Denote $\boldsymbol{\theta} = \{\theta_1, \ldots, \theta_M\}$ as the parameter vector of abilities of all participants, $\boldsymbol{\omega} = \{\omega_1, \ldots, \omega_L\}$ as the vector of item parameters, and $g(r; \theta, \omega)$ as the likelihood function of a selected IRT model. The maximum likelihood estimation can then be stated as:

$$(\hat{\boldsymbol{\omega}}, \hat{\boldsymbol{\theta}}) = \arg\max_{(\boldsymbol{\omega},\boldsymbol{\theta})} \sum_{i=1}^{N} \ln\, g(r_i; \omega_{d_i}, \theta_{f_i}) \tag{10}$$

Among specific applications, we can also see a Bayesian treatment [7], [8], where the aim is to calculate the full posterior of the parameters, hence capture the corresponding uncertainties. In this work, we primarily use maximum likelihood fitting in order to keep the computational cost manageable.

B. Computerised Adaptive Testing

The fundamental idea of CAT is that, rather than testing a participant with all the questions or a random sequence of questions, the participant is given questions with practical difficulty selected in real-time based on the current estimate of ability. We can then update the ability estimate with the response to the selected question and select the next question. Therefore, it is quite common to apply CAT based on a pre-trained IRT model, where we have estimated the difficulties (and other parameters) and abilities on a pool of items/participants.

As a result, most CAT approaches include three main components: an IRT model, an item selection method, and an item exposure method. As the name suggests, an item selection method determines, given the current ability estimate, how we select an item with appropriate difficulty to be the next question to estimate the ability better. Intuitively, we do not want the item to be too complicated or too simple for the actual ability, as in both cases the responses do not give much additional information about the ability. We introduce two common item selection methods in the following sections.

On the other hand, the item exposure method controls the marginal probability that a particular item is selected for the participant. The motivation is that we do not want a small number of questions to be exposed to the participants often. Such high exposure can potentially leak these questions to further participants hence affect later responses. In this work, we focus on the item selection criterion and discuss item exposure methods at the end of the paper.

1. Fisher Item Information

We start with the most commonly adopted approach for item selection, which uses Fisher information [9], [10]. Given the current candidate ability θ, a fitted IRT model with the likelihood function $g(r; \omega, \theta)$, and a set of L items with parameters $\{\omega_1, \ldots, \omega_L\}$, the Fisher item information (FII) on the jth item is then calculated as:

$$\mathtt{FII}(\theta; g, \omega_j) = \mathrm{E}_{R\sim\mathbb{P}(\omega_j,\theta)}\Big[\big(\frac{\partial \ln\ g(R;\omega_j,\theta)}{\partial\theta}\big)^2\Big] \tag{11}$$

$$= \int_r \Big(\frac{\partial \ln\ g(r;\omega_j,\theta)}{\partial\theta}\Big)^2 g(r;\omega_j,\theta)\mathrm{d}\ r \tag{12}$$

Here $\mathbb{P}(\omega_j, \theta)$ refers to the corresponding probability measure of the IRT model. The Fisher item information calculates the variance of the likelihood gradient, so that we can find the item(s) that can potentially change the likelihood function to a more considerable extent.

2. Kullback-Leibler Item Information

As illustrated above, FII only depends on the current estimate of the ability parameter θ according to the local gradient. Alternatively, one can consider calculating the information based on both the current estimate θ and a potential estimate $\theta_\star$. By considering different potential $\theta_\star$, we might obtain more global information for the item selection process. This idea motivates the KL information (KLI) [10], [11], which is constructed based on the Kullback-Leibler divergence between the IRT likelihood g with current ability θ and the one with an updated ability $\theta_\star$. The divergence on the jth item with parameter ω_j is defined as:

$$\mathrm{KL}_{\omega_j}(\theta_* \mid\mid \theta) = \mathrm{E}_{R\sim\mathbb{P}(\omega_j,\theta)}\Big[\ln\frac{g(R;\omega_j,\theta)}{g(R;\omega_j,\theta_*)}\Big] \tag{13}$$

However, during application time we do not have access to the updated parameter $\theta_\star$, and hence cannot calculate the KL-divergence directly. As a solution, we consider the potential information from the jth item to be the integrated divergence around the current ability θ, given the fact that the KL divergence is non-negative:

$$\mathtt{KLI}(\theta, g, \omega_j) = \int_{\theta_*=\theta-\epsilon}^{\theta+\epsilon} \mathrm{KL}_{\omega_j}(\theta_* || \theta)\mathrm{d}\theta_* \tag{14}$$

Hence, this KL item information is an aggregated gain around the current ability estimate, hence can be used to select the item with maximal information.

As mentioned, the main difference between FII and KLI is that the former only uses the local parameter estimates while the later obtains the information globally across different parameters [11]. The main benefit of the KLI approach is that it captures the changes in the ability parameter in both directions with a targeted range. Thus, it provides a way to merge the contributions from nearby regions on the item characteristic curve. On the other hand, FII is always based on the local gradient, requiring no extra configuration, which is more suitable when the ability estimate is closer to the actual value. KLI and FII can also prefer the same selection, particularly when the IRT model quantifies the responses well and has optimised likelihood on them. Later in the experiments, we adopt both of these two approaches to investigate their effectiveness for adaptive testing in machine learning empirically.

C. Applications in Machine Learning

There has been some recent work adopting the IRT framework for machine learning model analysis [7], [12], [13]. All three apply IRT on a model-instance level, seeing a model as a participant and treating a data instance within a given dataset as an item. In [12], [13] the authors use the Logistic model and discuss the interpretation of the learnt IRT parameters, including models like the always-correct model (e.g. predicts the ground truth). The response reflects whether a model correctly predicts the target class. In [7], the authors propose the three-parameter Beta model and learn its parameter in a Bayesian setting (e.g. posterior of the parameters). As the Beta IRT model supports bounded continuous response, in [7], the authors selected the correct class's predicted probability as the response.

III. Proposed Methods

The benchmarking methods we propose in this paper require some modifications on top of existing IRT and CAT methods to apply them to the problem of model-dataset evaluation. In general, we consider the following two requirements for the IRT and CAT methods. (1) They should support modelling continuous gain/loss measures standard in machine learning. (2) The corresponding item information should be obtainable analytically or through efficient approximations. Furthermore, we discuss the preference for non-negative discrimination in the scenario of a model-dataset benchmarking.

A. Modified Logistic IRT

The first modification is on the Logistic IRT family. Due to its original application scenario, the Logistic IRT family was used to model binary responses. As introduced above, to support CAT with a continuous response, the IRT needs to model a continuous response and provide the corresponding likelihood. The original Logistic IRT works on a Bernoulli assumption and the model estimates a mean parameter in the closed interval [0, 1]. While in Bernoulli distribution, the mean parameter is sufficient to calculate the likelihood, we need to consider another parameterisation for the continuous case. Although the Beta-3 IRT model uses the Beta likelihood and supports continuous response by default, it would also be valuable to keep an IRT model with sigmoid shape for better comparison. To achieve this, we replace the Bernoulli assumption with a logit-normal assumption in the IRT model. We use the original logistic function to calculate the mean of the response, and add a extra parameter s as the standard deviation:

$$R \mid \Theta = \theta, \Delta = \delta, A = a, S = s \sim \textbf{Logit-normal}(\mu_{(\theta,\delta,a)}, \sigma_{(s)}) \tag{15}$$

$$\mu_{(\theta,\delta,a)} = -a \cdot (\theta - \delta) \tag{16}$$

$$\sigma_s = s \tag{17}$$

The likelihood is then given as:

$$p(r \mid \theta, \delta, a, s) = \frac{1}{\sqrt{2\pi s^2}} \frac{1}{r \cdot (1-r)} \exp\Big(-\frac{\big(\ln(\frac{1-r}{r}) + a \cdot (\theta - \delta)\big)^2}{2s^2}\Big) \tag{18}$$

However, as the logit transform is not linear, the expectation (mean) and variance don't have closed forms:

$$\mathrm{E}[R \mid \theta, \delta, a, s] = \int_r p(r \mid \theta, \delta, a, s) \frac{1}{1 + \exp(r_i)} \mathrm{d}\, r \tag{19}$$

$$\mathrm{Var}[R \mid \theta, \delta, a, s] = \int_r p(r \mid \theta, \delta, a, s) \Big(\frac{1}{1 + \exp(r)} - \mathrm{E}[R \mid \theta, \delta, a, s]\Big) \mathrm{d}\, r \tag{20}$$

As both integrations involve the probability density function, the most straightforward solution here is to consider Monte-Carlo numeric integration (e.g., importance sampling):

$$\mathrm{E}[R \mid \theta, \delta, a, s] = \frac{1}{Q} \sum_{i=1}^{Q} \frac{1}{1 + \exp(r_i)} \tag{21}$$

$$\mathrm{Var}[R \mid \theta, \delta, a, s] = \frac{1}{Q} \sum_{i=1}^{Q} \Big(\frac{1}{1 + \exp(r)} - \mathrm{E}[R \mid \theta, \delta, a, s]\Big) \tag{22}$$

$$r_i \sim \textbf{Normal}(-a \cdot (\theta - \delta), s) \tag{23}$$

Here Q is the number of samples used in the calculation. In general, the approximation will be more accurate when using a larger Q. Although there is also a certain analytic approximation for the expected value (e.g. the probit approximation), we keep the sampling approach as it is also required to calculate item information as discussed later. With these modifications, the IRT model and corresponding CAT approaches can work with any bounded continuous response. While other possible extensions support continuous response [14], [15], we experimented particularly with the logistic and Beta-3 models given their close connection.

B. Approximate Item Information

The second modification also aims to incorporate continuous responses. While using binary responses, both Fisher item information and KL item information can be derived analytically [11]. Such closed forms generally are no longer possible when switching to IRT models with continuous response. However, as the integration in both Fisher item information and part of KL item information calculates an expectation over a density function, we can approximate them again with Monte-Carlo sampling. For FII, the approximation is given as:

$$\mathtt{FII}(\theta; g, \omega_j) = \frac{1}{Q} \sum_{i=1}^{Q} \Big(\frac{\partial \ln\, g(r_i; \omega_j, \theta)}{\partial \theta}\Big)^2$$

And similarly for KL divergence:

$$\mathrm{KL}_{\omega_j}(\theta_* \,||\, \theta) = \frac{1}{Q} \sum_{i=1}^{Q} \Big[\ln \frac{g(r_i; \omega_j, \theta)}{g(r_i; \omega_j, \theta_*)}\Big]$$

For both approximations we have $r_i \sim \mathbb{P}(\omega_j, \theta)$ to be random samples from the corresponding distribution.

While the calculation of FII is done with this single step, KLI still requires a further approximation to solve the second integration around θ, where we can consider a simple trapezoidal rule given ϵ is relatively small and $\mathrm{KL}_{\omega j}(\theta||\theta) = 0$:

$$\begin{aligned}\mathtt{KLI}(\theta, g, \omega_j) &= \frac{\epsilon}{2}\Big(\mathrm{KL}_{\omega_j}(\theta - \epsilon||\theta) + \mathrm{KL}_{\omega_j}(\theta||\theta)\Big) + \\ &\quad \frac{\epsilon}{2}\Big(\mathrm{KL}_{\omega_j}(\theta - \epsilon||\theta) + \mathrm{KL}_{\omega_j}(\theta||\theta)\Big) \\ &= \frac{\epsilon}{2}\Big(\mathrm{KL}_{\omega_j}(\theta - \epsilon||\theta) + \mathrm{KL}_{\omega_j}(\theta + \epsilon||\theta)\Big)\end{aligned}$$

With these approximation approaches, both item information quantities can be calculated efficiently, which is relevant as item information needs to be calculated for every candidate dataset at every step of the adaptive testing process.

C. The Constraint of Non-negative Discrimination

For typical IRT models, positive discrimination indicates the item has better average responses from candidates with higher ability estimates. In contrast, items with negative discrimination can be seen as tricky ones that cause stronger candidates to be more likely to give the wrong response than lower-ability candidates. In [13], the authors discuss the interpretation of negative discrimination in machine learning with each data instance being an item. One of their observations is that negative discrimination is often observed on instances within the regions where their opposite label dominates. A similar discussion can also be found in [7] with the Beta-3 IRT model. In this setting, the correct response from a model (candidate) when facing a data instance (item) is the instance's correct label. Assume we have a bag of instances with a Bayes optimal probability of 0.9 to be a positive class, and we can then conclude that models with a higher ability estimate should be more likely to give the correct response (positive). However, as there is still a probability of 0.1 for an instance to be negative, an optimistic prediction from a good model becomes the wrong response for these instances. It is clear that negative discrimination indeed describes the situation for these minority instances, and having negative discrimination parameters is essential for the IRT model to fit the responses correctly.

We now switch to the dataset configuration addressed in this paper, where each participant is still a model, but each item is changed to be a particular dataset. We consider a response to be a single performance measurement obtained via fitting the model on a random training fold and measuring the model with the remainder of the dataset. We assume all performance metrics to be calculated as gain measures so that a higher measurement indicates a better response for the IRT models. Therefore, if a model has a stronger ability, we expect it to have a higher averaged performance on most of the datasets, meaning it statistically fits well with a variety of training sets (i.e., it can capture a large function space) and also generalises to unseen test sets (i.e., no over-fitting). The question is then if we can design a dataset so that stronger models tend to have a lower (expected) performance, which is the requirement for negative discrimination to occur. The first possibility to have a averaged lower performance on a given dataset is that the dataset is hard to separate, that is, there is little pattern to be learnt from any part of the dataset. However, for such a dataset we expect most models to perform similarly as the labels are not dependent on the features, indicating a 0 discrimination is more suitable than negative values. The second possibility for a model to perform poorly on a dataset is that the model over-fits the training set. While this can happen with a particular combination of the training set and test set, it is less likely to occur when considering the averaged performance from a large number of random training sets and test sets. Furthermore, as discussed above, a model needs to be robust against over-fitting on most datasets to be estimated with higher ability. Therefore, it does not appear realistic to postulate that a specific dataset can cause more robust models to be more vulnerable to over-fitting.

In accordance with this discussion, in this paper we assume the discrimination parameter to be non-negative. In practice, we can achieve this either via constrained optimisation during the estimation of IRT parameters, or directly by estimating the logarithm of the discrimination parameters via unconstrained approaches. We adopt the latter in our implementation, within a stochastic gradient descent and automatic differentiation framework. Alternatively, one can also do it the Bayesian way, which assumes a prior distribution that makes positive discrimination more likely. However, as we only consider the maximum likelihood case in this paper, we leave this option as future work.

IV. Empirical Evaluation

This section experimentally investigates the performance of the IRT and CAT-based benchmarking methods introduced in this paper. We assess their performance with the following two experiments.

1. To compare different IRT models, we evaluate their performance to make inferences over unseen responses (several standard machine learning evaluation measures).
2. To assess the utility of the CAT-based method, we examine the efficiency of different item selection methods, in terms of the amount of computation costs it saves from testing the entire collection of datasets.

We first introduce the experimental setup. For the first IRT experiment, we compare the inference errors on responses using a standard train-test split. Regarding the CAT methods, we compare the final root mean squared error (RMSE) on the inferred response and the convergence speed, given the test sequences and the validation sets.

A. Setup

As response targets, we selected six evaluation measures commonly used in predictive machine learning: (1) multi-class accuracy, (2) Brier score, (3) log-loss, (4) weighted averaged binary accuracy, (5) weighted averaged binary AUC, and (6) weighted averaged binary F-measure. All these losses are bounded within $[0, 1]$ except the log-loss, which requires post-processing. We rescaled the averaged log-loss to the range of $[0, 1]$ with the exponential operation, which is an invertible calculation and ensures the final density function is valid on both scales. Furthermore, we use the negative value of Brier score and log-loss to fit the IRT models, so that they become gain measures (i.e., larger values indicate better results), in line with the other evaluation measures.

We select a set of datasets and model classes (described below) and run each model-dataset combination with an even random train-test split ten times. We use these results to train both Beta-3 and Logistic IRT models.

We use the 165 datasets provided by PMLB [16], which is a pre-processed collection of UCI datasets on various classification tasks. For computational efficiency, for all the datasets with more than $10,000$ instances, we sample it down to $10,000$ instances while approximately keeping the marginal distribution of the target variable.

We selected 9 model classes from the sklearn package: (1) multi-layer perceptron (MLP), (2) K nearest neighbours (KNN), (3) support vector machine (SVM), (4) pseudo Gaussian process (GP), (5) decision tree (TREE), (6) random forest (RF), (7) Ada boosting (ADA), (8) naive Bayes (NB), and (9) logistic regression (LR).

We selected eight different parameter settings for each model class to form different model instances, resulting in a total number of 72 models. For instance, for the MLP we choose a range of hidden units in a two-layer setting. Regarding the GP, here we call it pseudo models as the sklearn implementation does not support sparse covariance matrix hence can not scale to large datasets. We hence perform a simple random sampling on the training set. We first randomly select one data point for each class, then further sample random data points from the entire training set.

B. Evaluation of IRT Approaches

The first experiment we performed was to investigate whether the IRT models can accurately model and infer the performance measurements. As introduced in section II and III, the IRT models can estimate a distribution of the responses given each dataset and model combination. Therefore, we can evaluate each IRT model's goodness by evaluating the quality of these estimated distributions. Here we consider evaluating each distribution's mean, which is the estimated average performance measurement between the corresponding model and dataset. In general, we expect the estimated average response from a good IRT model to be close to some previously unseen measurements during future tests.

For this purpose, we perform ten times random split experiments on the collected responses from the 165 datasets and 72 models. We then divided the collected responses into a training set and a test set. We use the training set to estimate the IRT models' parameters, and the test set to verify the expected responses from each IRT model. Given the continuous responses, we use the root mean squared errors (RMSE) as the metric to evaluate the IRT models. Table I gives the results; notice here the RMSE is calculated after re-scaling all the evaluation metrics (e.g. the log-loss is re-scaled to $[0, 1]$). Additionally, the raw global mean and standard deviations of all the evaluation measures are also given. As the results show, both IRT models infer the evaluation measures well, with most RMSE values smaller than 0.05, which is considerably lower than the population standard deviation. For most evaluation measures, Logistic IRT and Beta-3 IRT perform similarly. Fig. 1 shows the item characteristic curves of both IRT approaches on the chess dataset and labour dataset with multi-class accuracy as an evaluation metric. Both IRT approaches tend to assign the same order to the ability parameters, as we can observe a similar pattern with the responses marked by the black points. Although in the bottom figures the two item characteristic curves are quite different from each other around the edges of the figures, it is noticeable that the curves behave similarly around the region with dense black points. This observation can help illustrate how the two different IRT approaches share similar RMSE values in the final results.

TABLE I. The Inference Errors (RMSE) of Both IRT Models on Different Evaluation Measures (top Two Rows), and the Global Mean and Standard Deviation of the Original Evaluation Measures (Bottom Two Rows)

	Acc	BS	LL	W-Acc	W-AUC	W-F1
Logistic	0.01349	0.00555	0.04379	0.00922	0.01763	0.05150
Beta-3	0.01569	0.00625	0.04151	0.02367	0.01498	0.05060
Global Mean	0.71878	0.79850	0.42725	0.79339	0.76487	0.58888
Global Std	0.21352	0.12954	0.29986	0.16509	0.18878	0.34585

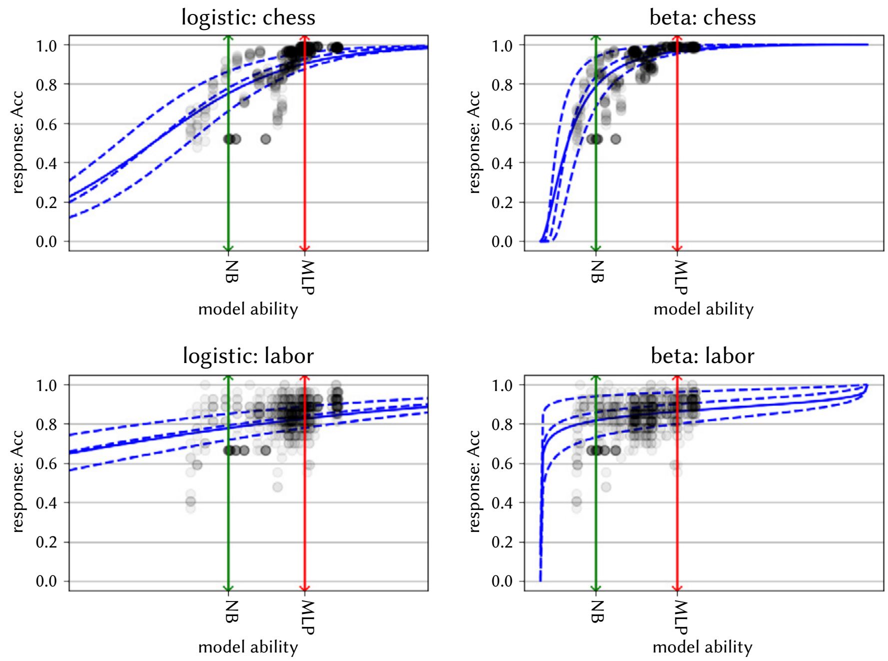


Fig. 1. The estimated item characteristic curves of the Logistic IRT and Beta-3 IRT models on two datasets evaluated with multi-class accuracy. The blue line indicates the mean value of responses, and the three dashed lines mark the 25%, 50% and 75% percentile of the responses. The grey dots mark the collected responses of the XXX models. The green line and the red line indicate the estimated ability of a naive Bayes classifiers and a multi-layer perceptron classifier respectively.

TABLE II. The Index of the Selected Dataset At Some Locations of the Adaptive Testing Sequence (Gradient Boosting Classifier and Multi-class Accuracy). For Example, the Logistic-FII Approach Select the 93 Dataset for the First Test, and Proceed With the 114 Dataset for the Second Test, This Selection Process Continues Till All the Datasets Have Been Tested

	Test 1	**Test 2**	**Test 3**	**Test 10**	**Test 50**	**Test 100**	**Test 150**
Logistic (Fisher)	93	114	147	135	79	140	42
Logistic (KL)	114	93	147	135	79	148	70
Logistic (Random)	20	80	89	39	27	164	61
Beta-3 (Fisher)	107	115	156	55	113	116	45
Beta-3 (KL)	107	1	43	124	113	23	45
Beta-3 (Random)	116	38	34	87	15	3	92

C. Evaluation of IRT and CAT Pairs

For the second experiment, we use different IRT and item selection approaches to test a set of different classifiers. We selected five classifiers from the sklearn package with their default settings as the candidate model: (1) gradient boosting classifier (GBC), (2) multi-layer perceptron (MLP), (3) support vector machine (SVM), (4) random forest (RF), and (5) logistic regression (LR). Here the GBC classifier is not used during the fitting of IRT models, while other classifiers have different parameter settings compared to those in the IRT model estimation process. While this group of classifiers doesn't cover all the model types as seen in the previous experiment, we select them due to their differences (e.g. linear v.s. nonlinear, ensemble v.s. standalone).

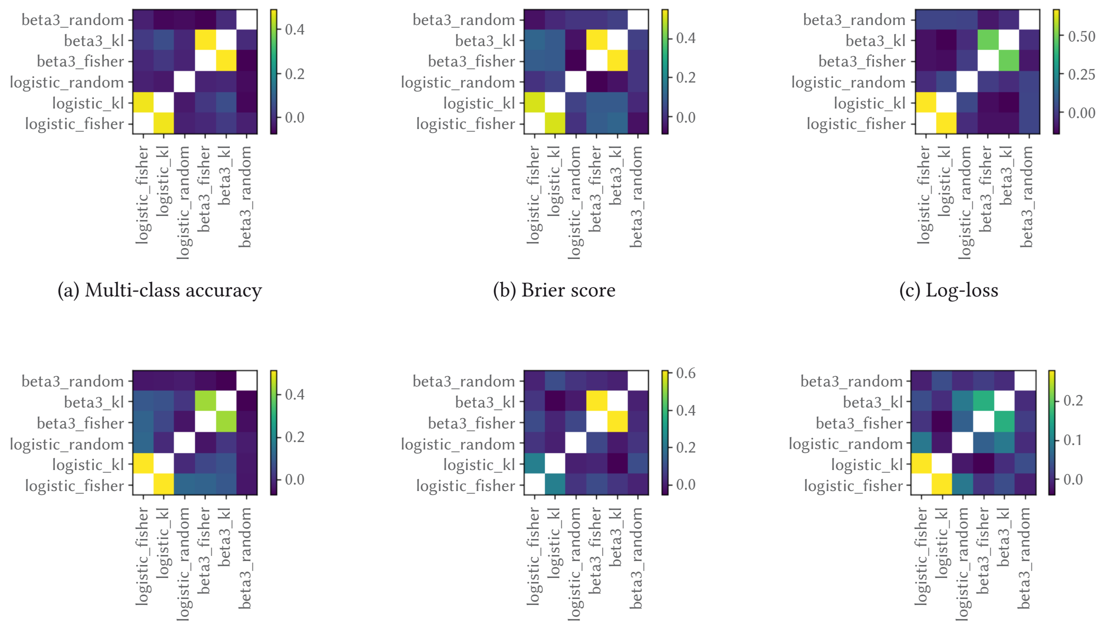
We run these models with all the datasets ten times using the same setting as in the previous experiment. The performance measurements are collected and used as a validation set. During adaptive testing, each time we update the model ability, we use the trained IRT to infer the expected value of responses (performance measures). We then calculate the corresponding RMSE the validation set to evaluate different IRT and CAT approaches. In principle, a better IRT-CAT combination should eventually have a lower RMSE and a faster convergence speed to the final RMSE.

We start by assuming the candidate model has average ability, then keep testing the model and updating its ability until we have tested all the datasets. We record the selected dataset at each test step, and the RMSE calculated using the validation set. Here we first analyse the results on the gradient boosting classifier (GBC) with multi-class accuracy as an example. Table II and Table III show the indices of the selected dataset and the RMSE on the averaged response on some locations of the testing sequences, respectively. It can be seen that both item information approaches pick similar datasets around the beginning of the sequence. This result can be observed with the logistic case, where test 1, 2, 3 , and 10 all select the same combination of datasets, and the order only differs between the first two tests. As discussed, FII and KLI can give similar selections when the IRT approach models the responses well. Hence our observation here agrees with the low RMSE as shown in the previous experiment. To further verify this observation, we calculate the pair-wise correlation with Kendall's Tau among the entire testing sequences for the GBC with all six performance metrics, and the results are shown in Fig. 2. The correlation between the two item information quantities with the same IRT approach can be clearly observed for the entire test sequence of 165 datasets across all metrics.

We can observe a similar pattern on the RMSE sequence decay on averaged responses. Both FII and KLI led to quite similar RMSE values around the beginning of the sequence with the first 3 tests. While the two item information approaches have lower RMSE values at the

TABLE III. Root Mean Squared Error of the Expected Response At Some Locations of the Adaptive Testing Sequence (Gradient Boosting Classifier and Multi-class Accuracy)

	Initial	**Test 1**	**Test 2**	**Test 3**	**Test 10**	**Test 50**	**Test 100**	**Test 150**
Logistic (Fisher)	0.12102	0.08409	0.08411	0.08354	0.08368	0.08466	0.08438	0.08375
Logistic (KL)	0.12103	0.12099	0.08525	0.08462	0.08393	0.08480	0.08452	0.08380
Logistic (Random)	0.12103	0.09959	0.09138	0.09038	0.08542	0.08380	0.08362	0.08362
Beta-3 (Fisher)	0.10100	0.07818	0.07828	0.07821	0.07817	0.07814	0.07809	0.07811
Beta-3 (KL)	0.10100	0.07818	0.07808	0.07803	0.07832	0.07855	0.07841	0.07843
Beta-3 (Random)	0.10100	0.08526	0.08236	0.08264	0.08444	0.07814	0.07834	0.07811



(d) Weighted averaged binary accuracy (e) Weighted averaged binary AUC (f) Weighted averaged binary F-measure

Fig. 2. Kendall's Tau between the adaptive testing sequences of the gradient boosting classifier on all six evaluation metrics, a brighter yellow colour indicates a stronger correlation and a deeper blue colour corresponds to a weaker correlation.

early stage, the random selection also performs remarkably well and gets to a relatively low RMSE value at test 10 in the logistic case. All RMSE values are very similar within each IRT approach from test 50 onwards.

To quantitatively adaptive testing sequence on the RMSE, we now examine the number of tests required before the inference error converges to a certain level according to the end of the test sequence. To calculate this number, given a test sequence of inference errors (RMSE), denoted as $(r_1, \ldots, r_T)$, we first select the final inference error r_T at the end of the entire sequence and construct a consequence region of $|r_T - \epsilon|$. In this experiments we set ϵ to be 0.05 of the minimum RMSE in the sequence, then we can obtain the convergence point c so that for $\forall i \geq c$ we have $|r_i - \epsilon| \leq \epsilon$.

Table IV lists the convergence points for the five different candidate classifiers and six evaluation measures. It is noteworthy that there are various cases where it took only 1 or 2 tests (out of 165 datasets) before the testing sequence reaches the convergence point. While both Fisher item information and KL item information require a smaller number of tests than random selection, we can still observe a few cases where random selection gives the fastest convergence. We hypothesise the randomness causes this within the model testing procedure. As the evaluation measurements can differ even on the same combination of dataset and model configuration, specific measurements cause a high bias on the item selection information, which leaves random selection a suitable backup choice. To obtain the best efficiency of adaptive testing, it is therefore suggested to calculate both item information and perform random selection while adaptive testing is required, so that the fastest convergence can always be achieved.

V. Conclusion

This paper introduced a novel framework to effectively benchmark a set of predictive models on an extensive collection of datasets. Instead of performing experiments on all possible datasets, we propose to model the similarity and dependency among different models and datasets to infer their experimental results without actually running all train-test cycles. Furthermore, we adopt the adaptive testing technique and uses the uncertainties on the unknown measurements to automatically decide a testing sequence for any unseen model based on the previous observations.

We performed a range of experiments, from which some general conclusions can be drawn. First of all, the choice of the IRT model plays an essential role in the benchmark. A suitable IRT model can indeed lead to better inference on the test results, without spending much effort on further testing. Which IRT model is most suited for which machine learning evaluation metric warrants further research.

TABLE IV. Number of Tests Before the Inferred Response Error (RMSE) Converges to the Overall Test Results

	GBC	MLP	SVM	RF	LR
Logistic (Fisher)	2	15	36	2	4
Logistic (KL)	3	17	36	3	4
Logistic (Random)	8	3	10	8	5
Beta-3 (Fisher)	2	2	5	2	15
Beta-3 (KL)	2	2	5	2	16
Beta-3 (Random)	13	16	27	5	8
Median	2.5	9	18.5	2.5	6.5

(a) Accuracy, Median: 5

	GBC	MLP	SVM	RF	LR
Logistic (Fisher)	2	1	10	88	18
Logistic (KL)	2	1	5	42	2
Logistic (Random)	7	1	13	4	3
Beta-3 (Fisher)	2	1	5	2	24
Beta-3 (KL)	2	1	2	2	19
Beta-3 (Random)	12	3	2	9	147
Median	2	1	5	6.5	18.5

(b) Brier Score, Median: 3

	GBC	MLP	SVM	RF	LR
Logistic (Fisher)	61	1	80	2	2
Logistic (KL)	21	1	54	3	30
Logistic (Random)	2	45	32	21	17
Beta-3 (Fisher)	13	97	31	147	9
Beta-3 (KL)	46	4	31	151	13
Beta-3 (Random)	43	125	2	97	6
Median	32	24.5	31.5	59	11

(c) Log loss, Median: 25.5

	GBC	MLP	SVM	RF	LR
Logistic (Fisher)	2	1	7	2	1
Logistic (KL)	2	1	4	2	1
Logistic (Random)	2	7	4	5	1
Beta-3 (Fisher)	2	1	1	2	7
Beta-3 (KL)	2	1	1	2	7
Beta-3 (Random)	2	12	18	2	14
Median	2	1	4	2	4

(d) Weighted averaged binary accuracy, Median: 2

	GBC	MLP	SVM	RF	LR
Logistic (Fisher)	2	2	3	6	2
Logistic (KL)	2	2	3	4	2
Logistic (Random)	2	7	17	2	3
Beta-3 (Fisher)	15	2	1	2	2
Beta-3 (KL)	15	2	4	10	2
Beta-3 (Random)	1	5	6	24	18
Median	2	2	3.5	5	2

(e) Weighted averaged binary AUC, Median: 2.5

	GBC	MLP	SVM	RF	LR
Logistic (Fisher)	3	1	1	24	4
Logistic (KL)	53	4	1	2	5
Logistic (Random)	11	3	7	2	3
Beta-3 (Fisher)	119	2	167	26	8
Beta-3 (KL)	2	167	88	8	116
Beta-3 (Random)	7	166	167	51	120
Median	9	3.5	47.5	16	6.5

(f) Weighted averaged binary F-measure, Median: 7.5

Secondly, we have demonstrated that adaptive testing can effectively reduce the total number of experiments. For most evaluation measures, we can observe a significant decay on the inference error with a small number of tests, leading to a significant reduction of model benchmarking costs.

One of the most promising directions for future research is to incorporate this adaptive testing framework into the development cycle of machine learning approaches. Modern data-driven approaches usually require many train-test runs to optimise their configuration and hyper-parameters. Although a range of approaches have been proposed in auto-ML and neural architecture search [17], most approaches still require to perform large-scale experiments on the given datasets to obtain the search points. With the assistance of adaptive testing, we can further attempt to reduce such search costs by selecting the most promising datasets. Another direction is to look beyond predictive machine learning tasks. Recent work has made significant progress on non-predictive tasks such as random data generation and neural-based density estimation. Both areas can potentially benefit from adaptive testing considering their significant computational demands during training. Item exposure control [18] is also worth further consideration in the benchmarking process, which allows us to further control the rate that a particular dataset is used.

References

[1] M. Hutson, "Artificial intelligence faces reproducibility crisis," 2018. [Online]. Available: https://science.sci-encemag.org/content/359/6377/725, doi: 10.1126/sci-ence.359.6377.725.

[2] J. Vanschoren, J. N. Van Rijn, B. Bischl, L. Torgo, "OpenML: networked science in machine learning," *ACM SIGKDD Explorations Newsletter*, vol. 15, no. 2, pp. 49–60, 2014.

[3] G. A. Morris, L. Branum-Martin, N. Harshman, S. D. Baker, E. Mazur, S. Dutta, T. Mzoughi, V. McCauley, "Testing the test: Item response curves and test quality," *American Journal of Physics*, vol. 74, no. 5, pp. 449–453, 2006.

[4] W. J. van der Linden, R. K. Hambleton, *Handbook of modern item response theory*. Springer Science & Business Media, 2013.

[5] B. F. Green, R. D. Bock, L. G. Humphreys, R. L. Linn, M. D. Reckase, "Technical guidelines for assessing computerized adaptive tests," *Journal of Educational Measurement*, vol. 21, no. 4, pp. 347–360, 1984.

[6] D. J. Weiss, G. G. Kingsbury, "Application of computerized adaptive testing to educational problems," *Journal of Educa-tional Measurement*, vol. 21, no. 4, pp. 361–375, 1984.

[7] Y. Chen, T. S. Filho, R. B. Prudencio, T. Diethe, P. Flach, "β3-IRT: A new item response model and its applications," in *AISTATS 2019*, vol. 89 of *Proceedings of Machine Learning Research*, 2019, pp. 1013–1021.

[8] B. P. Veldkamp, M. Matteucci, "Bayesian computerized adaptive testing," *Ensaio: Avaliação e Políticas Públicas em Educação*, vol. 21, no. 78, pp. 57–82, 2013.

[9] R. R. Meijer, M. L. Nering, "Computerized adaptive testing: Overview and introduction," 1999. [Online]. Available: https://doi.org/10.1177/01466219922031310, doi: 10.1177/01466219922031310.

[10] H.-H. Chang, "Psychometrics behind computerized adaptive testing," *Psychometrika*, vol. 80, no. 1, pp. 1–20, 2015.

[11] H.-H. Chang, Z. Ying, "A global information approach to computerized adaptive testing," *Applied Psychological Measurement*, vol. 20, no. 3, pp. 213–229, 1996.

[12] R. B. Prudêncio, J. Hernández-Orallo, A. Martínez-Usó, "Analysis of instance hardness in machine learning using item response theory," in *Second International Workshop on Learning over Multiple Contexts in ECML 2015. Porto, Portugal, 11 September 2015*, 2015.

[13] F. Martínez-Plumed, R. B. Prudêncio, A. Martínez-Usó, J. Hernández-Orallo, "Making sense of item response theory in machine learning," in *Proceedings of the Twentysecond European Conference on Artificial Intelligence*, 2016, pp. 1140–1148, IOS Press.

[14] K. SHOJIMA, "A noniterative item parameter solution in each em cycle of the continuous response model," *Educational technology research*, vol. 28, no. 1-2, pp. 11–22, 2005.

[15] F. Samejima, "Graded response model," in *Handbook of modern item response theory*, Springer, 1997, pp. 85–100.

[16] R. S. Olson, W. La Cava, P. Orzechowski, R. J. Urbanowicz, J. H. Moore, "Pmlb: a large benchmark suite for ma-chine learning evaluation and comparison," *BioData mining*, vol. 10, no. 1, p. 36, 2017.

[17] M. Feurer, A. Klein, K. Eggensperger, J. Springenberg, M. Blum, F. Hutter, "Efficient and robust automated machine learning," in *Advances in neural information processing systems*, 2015, pp. 2962–2970.

[18] J. Sympson, R. Hetter, "Controlling item-exposure rates in computerized adaptive testing," pp. 973–977, 1985.

Hao Song

HS obtained his PhD at the University of Bristol in 2017, and is currently a postdoctoral researcher at the University of Bristol. His research interests include quantifying different types of uncertainties within the machine learning pipeline, particularly for different probabilistic out-puts and corresponding evaluation metrics.

Peter Flach

PF is Professor of Artificial Intelligence at the University of Bristol and has over 30 years experience in machine learning and data mining, with particular expertise in mining highly-structured data and the evaluation and improvement of machine learning models. He was PC co-chair of KDD'09 and ECML-PKDD'12 and has edited and authored several books, including *Machine Learning: the Art and Science of Algorithms that Make Sense of Data*.

Antimicrobial Resistance Prediction in Intensive Care Unit for *Pseudomonas Aeruginosa* using Temporal Data-Driven Models

Àlvar Hernàndez-Carnerero[1]*, Miquel Sànchez-Marrè[1], Inmaculada Mora-Jiménez[2], Cristina Soguero-Ruiz[2], Sergio Martínez-Agüero[2], Joaquín Álvarez-Rodríguez[3]

[1] Intelligent Data Science and Artificial Intelligence Research Centre (IDEAI-UPC), Dept. of Computer Science, Universitat Politècnica de Catalunya, 08034 Barcelona (Spain)

[2] Department of Signal Theory and Communications, Telematics and Computing Systems, Rey Juan Carlos University, Madrid 28943 (Spain)

[3] Intensive Care Department, University Hospital of Fuenlabrada, Madrid 28942 (Spain)

Received 4 January 2021 | Accepted 29 January 2021 | Published 25 February 2021

Abstract

One threatening medical problem for human beings is the increasing antimicrobial resistance of some microorganisms. This problem is especially difficult in Intensive Care Units (ICUs) of hospitals due to the vulnerable state of patients. Knowing in advance whether a concrete bacterium is resistant or susceptible to an antibiotic is a crux step for clinicians to determine an effective antibiotic treatment. This usual clinical procedure takes approximately 48 hours and it is named antibiogram. It tests the bacterium resistance to one or more antimicrobial families (six of them considered in this work). This article focuses on cultures of the *Pseudomonas Aeruginosa* bacterium because is one of the most dangerous in the ICU. Several temporal data-driven models are proposed and analyzed to predict the resistance or susceptibility to a determined antibiotic family previously to know the antibiogram result and only using the available past information from a data set. This data set is formed by anonymized electronic health records data from more than 3300 ICU patients during 15 years. Several data-driven classifier methods are used in combination with several temporal modeling approaches. The results show that our predictions are reasonably accurate for some antimicrobial families, and could be used by clinicians to determine the best antibiotic therapy in advance. This early prediction can save valuable time to start the adequate treatment for an ICU patient. This study corroborates the results of a previous work pointing that the antimicrobial resistance of bacteria in the ICU is related to other recent resistance tests of ICU patients. This information is very valuable for making accurate antimicrobial resistance predictions.

Keywords

Antimicrobial Resistance, Intensive Care Unit, Prediction, *Pseudomonas Aeruginosa*, Temporal Data-Driven Modeling.

DOI: 10.9781/ijimai.2021.02.012

I. Introduction

Antimicrobial resistance occurs when a germ develops the capacity to not respond to the drugs designed to combat them [1]. Nowadays, antimicrobial resistance is one of the greatest threats to the global health system [2]. Apart from the health consequences, the economic impact deriving from antimicrobial resistance is not a trivial issue, resulting in a 7% reduction in the Gross Domestic Product by 2050 [3]. Indeed, it has become more acute in recent years due to the excessive use of antibiotics in many facets of daily life [4].

The acquisition of antimicrobial resistance is favoured in hospital environments, being even worsened for patients admitted to the Intensive Care Unit (ICU). This could be motivated by the duration and intensity of the drug treatment, as well as by the use of life support devices. The critical health status of ICU patients pushes actions to anticipate the result of the cultures provided by the microbiology laboratory, which usually takes 48 hours. A culture is a biological sample collected to isolate a bacterium, aiming to analyze its susceptibility to different antibiotics. The test used to measure this susceptibility is called antibiogram, and its result (susceptible/resistant) is commonly used by clinicians to determine the antibiotic treatment [5]. It is interesting to note that several families of antibiotics may have similar susceptibility when tested on a given germ species [6]. There are several species with high prevalence, for example, Acinetobacter spp.; *Enterococcous fecalis* and *Enterococcus faecium*; *Escherichia coli*; *Klebsiella pneumoniae*; *Pseudomonas aeruginosa*; and *Staphylococcus aureus*, among others. In this paper, we focus on *Pseudomonas aeruginosa* for the following reasons: (1) its virulence, specially in the ICU;(2) its ability to cause chronic infectious diseases; and (3) its ability to develop multi-drug resistance [7], [8].

For all these reasons, anticipation to the culture result in case of resistance, is vital to isolate the patient and control the spread of

* Corresponding author.

E-mail address: alvar.hernandez@upc.edu

antimicrobial resistance among other ICU patients. Computational tools inspired on data-driven models may be supportive to clinical decisions previous to the antibiogram result. The article [6] introduces the *concept drift* observed in antimicrobial resistance data sets, and it uses a windowing scheme together with dynamic classifiers to perform resistance prediction. It classifies cultures as susceptible or resistant to some antibiotics using a database of EHR which includes years from 2002 to 2004, considering cases of meningitis. A high number of the state-of-the-art studies use whole genome sequencing [9]–[12]. Because of its considerable cost, in this study we propose to predict resistant bacteria based on Electronic Health Records (EHR) data from ICU, together with historic antibiogram results. This data is already available in most hospitals, and therefore the methodology proposed in this paper can be straightforward extrapolated. Comparable approaches are studied in previous works [6], [13]–[18]. In [17], bacterial infection in the ICU using EHR data is predicted (binary classification task) by applying a set of machine learning (ML) methods. The prediction is carried out at the patient level in order to determine which patients no longer require more antimicrobial treatment. Longitudinal data from 2001 to 2012, extracted over the 24-hour, 48-hour or 72-hour window following their first antibiotic dose, are considered. No temporal modelling was explicitly taken into account. The work in [18] presents an study for predicting bacterial resistance also using EHR data, from 2013 to 2015. An ensemble of ML methods is used to classify isolated bacterial cultures as susceptible or resistant to a particular antibiotic. The temporal relation among instances is considered here, with features indicating the proportion of past antibiotic resistance infections identified as having the highest average impact. This study also concludes that the feature encoding the date of the culture has some effect on the prediction, probably due to the fluctuating resistance frequencies through time.

Owing to the dynamics of antimicrobial resistance, we analyze in this paper electronic health records collected during 15 years, from 2004 to 2019, by the University Hospital of Fuenlabrada (UHF) in Madrid, Spain. This data have been partly considered in previous studies carried out by the authors [14], [15], [16],[19]. In particular, authors in [14] used a reduced dataset taking into account two years less (from 2004 to 2017) than in the current work. All patients admitted in the ICU in this period were considered in [14], regardless of their length stay. Additionally, authors in [14] used ML to determine whether a *Pseudomona Aeruginosa* bacterium will be resistant or not (binary target) to different families of antimicrobials without considering information about historic antibiogram results. In [15], we analyzed for the first time the dynamics on *Pseudomona Aeruginosa* by considering incremental time windows on a period of time from 2004 to 2013, with two families of antibiotics. It was also our first incursion on the use of features taking into account the result provided by previous antibiograms of other ICU patients. This current paper extends the work in [15] while considering the predictive window length (one month) that best results provided in [15]. Specifically, to carry out predictions, the Random Forest (RF) method has been added to previously considered method, Logistic Regression (LR). We have increased both the number of years under study and the number of antibiotics (from 2 to 6). We have also considered as features the result provided by previous antibiograms of each patient, weighted by a factor depending on the time elapsed since the last antibiogram was tested. Furthermore, two approaches have been explored to analyze the dynamic of antimicrobial resistance by evaluating the models in several time horizons.

The rest of the paper is as follows. In Section II, we describe the data set analyzed in this paper and provide a graphical exploration of it. Section III introduces the data preprocessing as well as the methods used for temporal modelling. Results and discussion and provided in IV. Finally, the conclusions are presented in Section V.

II. Materials

A. Data Set Description

Data considered in this work correspond to 3812 admissions of 3346 ICU patients, collected at the UHF during a period of 15 consecutive years (from July 2004 to May 2019). Note that, since the number of ICU admissions exceeds the number of patients, there are patients with more than one ICU admission during this period. A total of 43658 cultures were collected. Although there are more than 290 different types of bacteria and 27 antimicrobial families, we only take into account here the cultures where *Pseudomonas* have been detected, ending up in a total of 764 cultures. For this bacterium, the antibiograms considered in this work test the response (encoded as susceptible (s) or resistant (r)) against the following set of family of antibiotics *a* = {amg, car, cf4, pap, pol, qui}. Elements in the set *a* refer to: Aminoglycosides (AMG), Carbapenems (CAR), 4th Generation Cephalosporins (CF4), Extended-spectrum penicillins (PAP), polymyxins (POL) and Quinolones (QUI), respectively.

Since data-driven models are based on learning from instances, we consider here the target c&a_i, as the antibiogram result for a specific family of antibiotic a_i, for every culture collected to any patient. The feature vector associated with each target is represented by the 40 features described in Table I. We define here the instance as the pair composed by the feature vector (input features to the data-driven models) and the target (outcome of the data-driven models).

TABLE I. Name and Description of the Features Characterizing Each Instance for Every Family of Antibiotics (AMG, CAR, CF4, PAP, POL, QUI), Tested on a particular Patient P. The Result for the Antibiogram Family a_i Is Encoded in the Binary Target Feature c&a_i (Not Presented in this Table)

Feature name	Description
age	age of the patient
gender	gender of the patient
origin	clinical origin before ICU admission
goi_*	7 features, each linked to a different group of illness *: A, B, C, D, E, F, G
pluripathology	number of groups of illness
patient_category	clinical category of the patient
reason_admission	reason of admission at ICU
start_date	date the patient was admitted
day_week_admission	day of the week the patient was admitted to the ICU
day_month_admission	day of the month the patient was admitted to the ICU
month_admission	month the patient was admitted
year_admission	year the patient was admitted
date_culture	date of the culture
day_week_culture	weekday the culture was collected
day_month_culture	day of month the culture was collected
month_culture	month the culture was collected
year_culture	year the culture was collected
culture_type	type of culture
culture_type_group1	1st level grouping for the culture type
culture_type_group2	2nd level grouping for the culture type
days_to_culture	number of days elapsed from start_date to date_culture
p&a_i	6 features, each linked to one family a_i of previous antibiograms of patient p: amg, car, cf4, pap, pol and qui
r&a_i	6 features, each linked to one family a_i of previous antibiograms for other recent patients different from p: amg, car, cf4, pap, pol and qui

As for the input features, we first analyze **demographic data**: age, gender, group of illness A (cardiovascular events), B (kidney failure, arthritis), C (respiratory problems), D (pancreatitis, endocrine), E (epilepsy, dementia), F (diabetes, arteriosclerosis) and G (neoplasms), and pluripathology (indicating whether the patient has more than two comorbidities). The median age of patients admitted to the ICU was 64 years (interquartile range 55-73, range 18-87), with a majority of men (70%). Pluripathological patients are 40.6% of the patients, with comorbidities mostly related to respiratory problems (33.4%), diabetes (26.3%) and neoplasms (33.1%).

We then focus on the **information about the ICU admission**: date of admission to the ICU, department of origin before ICU admission (surgery, internal medicine, urology,...), reason for admission (serious infection, acute respiratory failure, hypovolaemia,...) and patient category (medical or surgical). The clinical origin before the ICU admission more common was surgery (31.1% patients) and emergency department (18.4%). The reason of admission more common was serious infection (22.5% patients) and acute respiratory failure (18.4% patients). The most common patient category was medical (52.2 %).

This work also analyses the information related to the **cultures**. Specifically, we consider the culture type (exudate, drainage, biopsy, sputum, bronchoaspirate, etc.); first level grouping for the type of culture, which classifies the cultures into surface, liquids, respiratory, etc.; and the second level grouping for the type of culture, used to identify a clinical sample or a surface culture. Besides, the date of the culture, the weekday the culture was collected , as well as the month and the year.

Finally, to collect temporal information in each instance associated to patient *p*, the current study proposes to generate two kind of features linked to previous resistant antibiograms. In particular, we consider: (1) previous resistant results of the same patient, and (2) previous resistant results of all patients who recently stayed in the ICU.

Own past cultures features. The first kind of features is associated with the detection of resistant bacteria in previous antibiograms for a specific patient *p*, and aims to quantify the current "*intensity*" of these bacteria. These features consider the result of antibiograms of *Pseudomonas Aeruginosa* during an interval between 21 days and 48 hours previous to the current culture being studied for patient *p*, $c^{(p)}$. The 48-hour limit is considered since it is usually the time the results of the antibiogram take to be available. Furthermore, cultures are gathered until 21 days before the date *d* of current culture *c*, because if the antibiogram result is positive, from a clinical point of view, it is kept as positive for the following 21 days.

Thus, when a culture is collected, a total of six features, one per antimicrobial family, are generated: `p&amg`, `p&car`, `p&cf4`, `p&pap`, `p&pol` and `p&qui`. Each feature takes into account the antibiogram results for the corresponding antimicrobial family, e.g. `p&pap` just consider previous results associated with patient *p* for the family of antibiotics PAP. Because of that, the group of own past cultures of patient *p*, named $C^{(p)}$, is divided into six data sets $C_{a_i}^{(p)}$. To illustrate how the value for each feature $\texttt{p\&a}_i$, $i = 1, 2, \cdots, 6$ is obtained, let us consider that the subset $C_{a_i}^{(p)}$ has $n_{a_i}^{(p)}$ cultures, i.e. $C_{a_i}^{(p)} = \{ c_{a_i,k}^{(p)} \}_{k=1}^{n_{a_i}^{(p)}}$. Each culture $c_{a_i,k}^{(p)} \in C_{a_i}^{(p)}$ has associated: (1) a date $d_{a_i,k}^{(p)}$ when it was collected ; and (2) a susceptibility test result $r_{a_i,k}^{(p)}$, which is *susceptible* or *resistant* depending on whether the bacterium is susceptible or resistant to a_i, respectively. To calculate the potential contribution of a culture $C_{a_i}^{(p)}$ to the feature $\texttt{p\&a}_i$, $i = 1, 2, \dots, 6$, the Negative Exponential Function (NEF) is applied as follows:

$$\text{NEF}\left(c_{a_i,k}^{(p)}, c^{(p)}\right) = \begin{cases} 0 & \text{if } r_{a_i,k}^{(p)} = \text{susceptible} \\ e^{-\lambda\left(d - d_{a_i,k}^{(p)}\right)} & \text{if } r_{a_i,k}^{(p)} = \text{resistant} \end{cases} \tag{1}$$

where the value of parameter λ is experimentally set to 0.095. To compute the feature value $\texttt{p\&a}_i$ for the instance associated with culture $c^{(p)}$ of patient *p*, the maximum outcome in Equation (1) is obtained according to Equation (2):

$$\texttt{p\&a}_i = \max_{k=1,\cdots,n^{(p)}} \text{NEF}\left(c_{a_i,k}^{(p)}, c^{(p)}\right) \tag{2}$$

ICU-patients past cultures features. The second kind of features are named `r&amg`, `r&car`, `r&cf4`, `r&pap`, `r&pol` and `r&qui`. These features aim to encode the "*intensity*" of resistant bacteria in the ICU during the time previous to the date d of the current instance and culture. Differently from the previous set of six features $\texttt{p\&a}_i$, now the "*intensity*" takes into account the number of patients (different from current patient *p*) that were infected by a resistant bacterium and, for each of them, the time elapsed since the bacterium was detected. For a particular feature, a single value is calculated by considering the result of past susceptibility tests of *Pseudomonas Aeruginosa* for the *P* patients, denoted as p_j with $j = 1 \cdots P$, in the ICU during the time interval between 21 days and 48 hours previous to date d of culture $c^{(p)}$ of patient *p*. An exponential decay is again considered to weight the result of each susceptibility test.

The group $C^{(p')}$ of past cultures of other patients is divided into six subsets $C_{a_i}^{(p')}$ too. Every particular subset $C_{a_i}^{(p')}$ is split into *n* disjoint subsets, as many as patients:

$$C_{a_i}^{(p')} = \bigcup_{j=1}^{P} C_{a_i}^{(p_j)}, \qquad C_{a_i}^{(p_j)} = \left\{c_{a_i,k}^{(p_j)}\right\}_{k=1}^{n_{a_i}^{(p_j)}} \tag{3}$$

where $C_{a_i}^{(p_j)}$ is composed of the $n_{a_i}^{(p_j)}$ antibiogram results for a_i in patient p_j. As previously mentioned, the set of cultures of patient *p* are excluded from $C_{a_i}^{(p')}$.

Since every culture $c_{a_i,k}^{(p_j)}$ has a susceptibility test result $r_{a_i,k}^{(p_j)}$ and a date $d_{a_i,k}^{(p_j)}$, the application of the NEF expression equivalent to that in Equation (1), just replacing $c_{a_i,k}^{(p)}$, $d_{a_i,j}^{(p)}$ and $r_{a_i,k}^{(p)}$ by $c_{a_i,k}^{(p_j)}$, $d_{a_i,k}^{(p_j)}$ and $r_{a_i,k}^{(p_j)}$, respectively. Then, each feature $\texttt{r\&a}_i$ is obtained by adding up the maximum value of Equation (1) for each patient p_j, as indicated in Equation (4).

$$\texttt{r\&a}_i = \sum_{j=1}^{P} \max_{k=1,\cdots,n_{a_i}^{(p_j)}} \text{NEF}\left(c_{a_i,k}^{(p_j)}, c^{(p)}\right) \tag{4}$$

B. Graphical Exploration

Owing to the high number of features, we start by identifying the most relevant features per family of antibiotics. For this purpose, we consider a filter approach with the Mutual Information (MI) score [20]. Thus, for each family of antibiotics, Fig. 1 shows the five features with the highest MI values, comprising among them the date of culture and the information about the previous cultures both for the own patient and for the UCI environment. According to the mutual information score, the most relevant feature is `date_culture` for each of the antimicrobial families considered. This results supports the importance of the antimicrobial resistance dynamics, which is common for all families of antibiotics.

To get a deeper insight on this issue, Fig. 2 graphically illustrates the evolution of the number of susceptible antibiograms (a) and resistant antibiograms (b) for each family of antimicrobials tested on *Pseudomonas* along time. Not all families of antibiotics were tested during the whole period considered. Specifically, clinicians first agreed to modify the range of tested antibiotics in the ICU of the UHF, first by including POL in 2007 and then by stop susceptibility testing antibiograms of QUI in 2018, due to its high resistance. Furthermore, there is a very noticeable fall in the number of resistant and susceptible

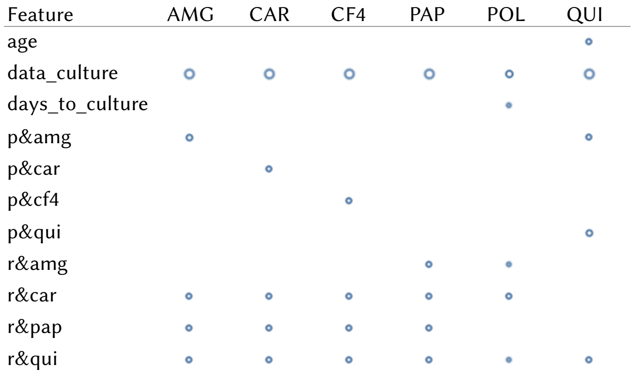


Fig. 1. For each antimicrobial family, the five features with the highest MI scores, indicated by the circle size from MI=0.56 (biggest size, pair date_culture-AMG) to MI=0.09 (smallest size, pair `r&amg`-POL).

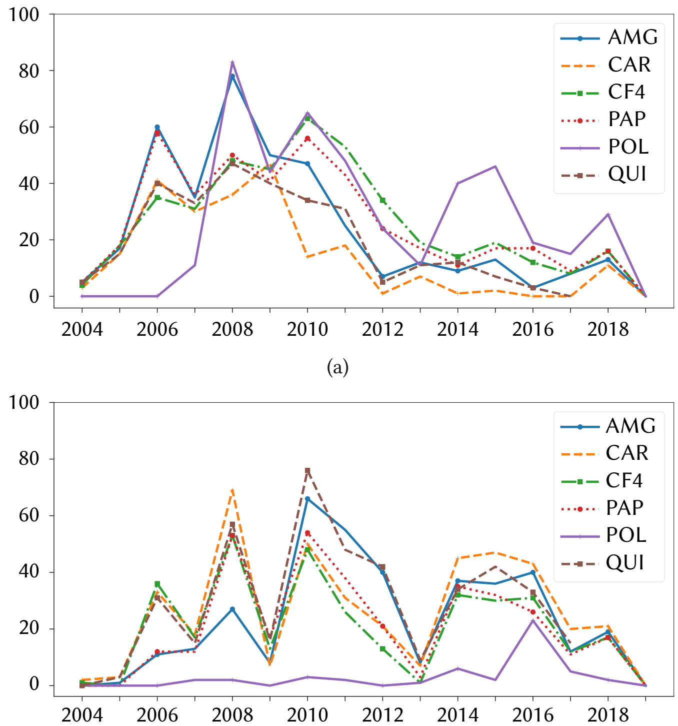


Fig. 2. Temporal evolution for the number of annual susceptible (a) and resistant (b) antibiograms when tested on *Pseudomonas* cultures for each family of antimicrobials.

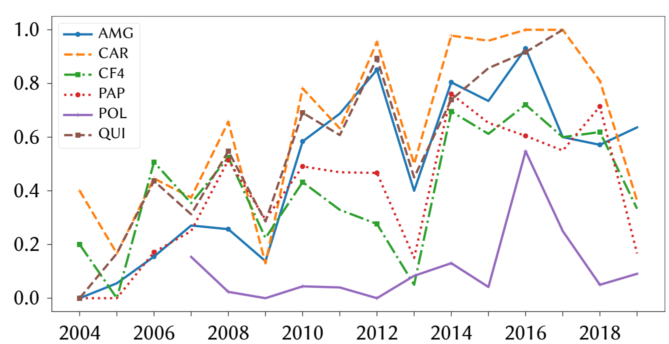


Fig. 3. Temporal evolution of the ratio between the number of annual resistant antibiograms tested on *Pseudomonas* cultures and the total annual number of cultures on *Pseudomonas* for each family of antimicrobials.

antibiograms in 2013. This decrease is probably motivated because of integration problems due to software update in the ICU health information system in 2013. As stated in the literature, the number of susceptible antibiograms tend to decrease in the most recent years.

In this line, we also analyze the annual ratio of resistant antibiograms results for each family of antimicrobials. To obtain this ratio, the number of resistant cultures per year has been divided by the number of total cultures per year (both resistant and susceptible cultures). The general trend is that, as time progresses (and therefore the value of `date_culture` increases), a higher percentage of instances tend to be resistant.

The second most relevant feature for the antimicrobial families AMG, CAR and QUI are `p&amg`, `p&car` and `p&qui`, respectively. This shows the importance of the outcome of previous antibiograms of the same patient for the family under consideration. In the case of CF4, `p&cf4` is the 4th most important feature. Though not presented in Fig. 1, `p&pap` is ranked on the 7th position for PAP, and `p&pol` in the 11th position for POL. It is interesting to remark here that, in all cases, the MI score for a particular family of antibiotics is higher for the $\texttt{p\&a}_i$ feature corresponding to that particular family than to any of the other five $\texttt{p\&a}_i$ features. This points out the relevance of considering the particular antimicrobial family when using results of previous antibiograms.

Fig. 4 shows the boxplots for each of the six features named $\texttt{p\&a}_i$, associated to the antibiogram results of the same patient for each family of antibiotics (in rows). Blue boxplots refer to $\texttt{p\&a}_i$ for resistant results, while black ones refer to $\texttt{p\&a}_i$ for susceptible results. In general, we observe that the median of $\texttt{p\&a}_i$ is higher when the culture c was resistant than when it was susceptible. The results shown in Fig. 4 for CAR and QUI are particularly interesting for susceptible cultures (black boxplots) for all the families, with most of the previous antibiogram results being susceptible. However, for CF4 and PAP, most of antibiogram results are susceptible for `p&cf4`, `p&pap` and `p&pol`, whereas for POL it only happens for `p&pol`. Note that, regardless the family of antibiotics tested, the boxplot of `p&car` and `p&qui` for resistant cultures (blue bloxplots) is very similar to the boxplot associated to the corresponding family of antibiotics considered (e.g, see `p&amg`, `p&car` and `p&qui` in Fig. 4 for AMG, or `p&pap`, `p&car` and `p&qui` for PAP.

The $\texttt{r\&a}_i$ features are also among the most relevant features according to the MI score. In this case there is no clear distinction on the ranking depending on the antimicrobial family. It supports the importance of taking into account the existence of any resistant germ in the ICU. The feature `r&pol` (not included among the top five features in Table I) seems to be the one providing less information, probably because of low number of antibiograms with a resistant result for this family. Fig. 5 presents the bloxplots for the $\texttt{r\&a}_i$ features. In comparison with boxplots in Fig. 4, note that boxplots of the $\texttt{r\&a}_i$ features are not limited to a maximum of one, since the number of patients contributing in Equation (4) is n (usually greater than 1). For each antibiogram a_i, the median values of the $\texttt{r\&a}_i$ features resistant and susceptible results is much closer between them than when comparing the $\texttt{p\&a}_i$ features. It is also remarkable that boxplots associated with `r&pol` show a median value very close to zero both for resistant and susceptible cultures, in line with previous comments. Furthermore, when analyzing POL, the median value is higher for susceptible than for resistant cultures, excepting for `r&pol`, showing a different behavior of this antibiotic.

Finally, among the features in the top five with a higher MI score, we also find `days_to_culture` (for POL) and *age* (for QUI). Both features are also among the top ten for the rest of the antimicrobial families. From a clinical viewpoint, it is known that both age and a longer ICU stay are risk factors to become infected [14].

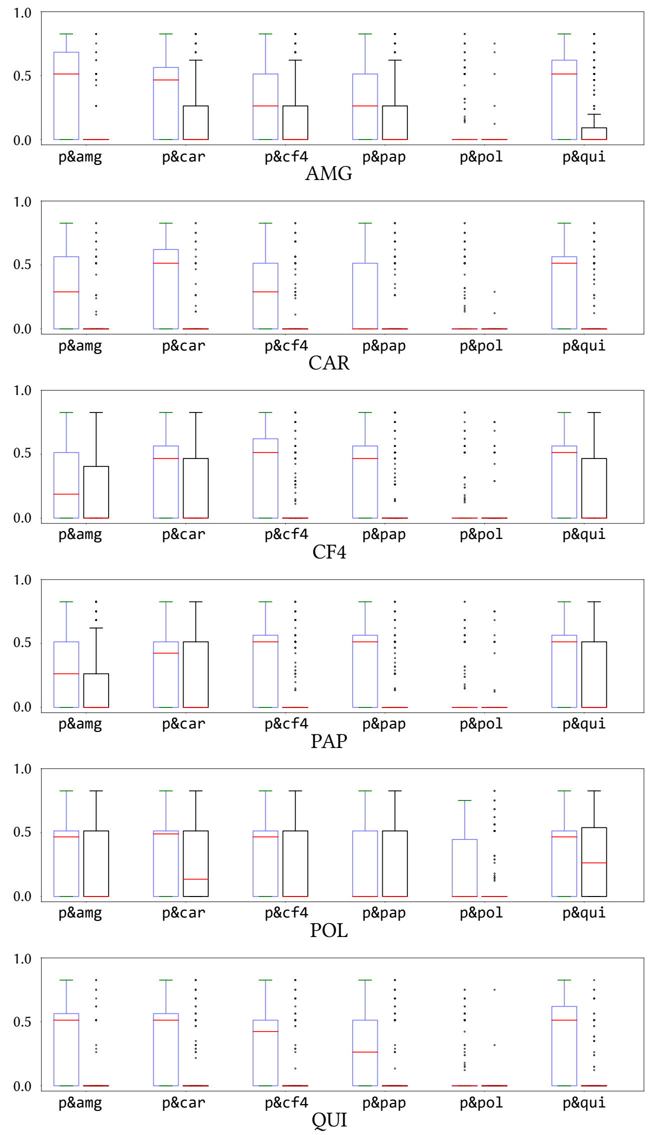


Fig. 4. Boxplots for the six features named p&a$_i$ when considering both resistant (left boxpplot, in blue) and susceptible (right bloxplot, in black) antibiograms for culture *c*.

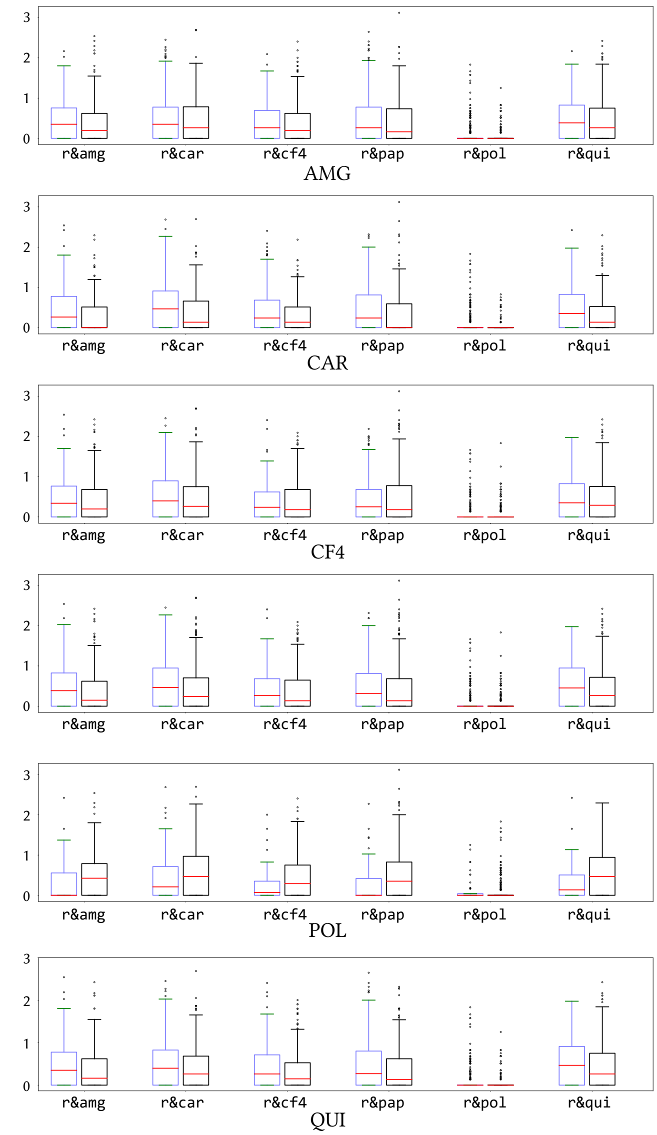


Fig. 5. Boxplots for the six features named r&a$_i$ when considering both resistant (left boxpplot, in blue) and susceptible (right bloxplot, in black) antibiograms for culture c.

III. Methods

A. Data Preprocessing

Before using the data set to predict the result of the susceptibility test, a previous stage of preprocessing is needed. The first aspect to be considered is that six binary classifiers are going to be built in order to predict whether a culture is susceptible or resistant to each of the six different antimicrobial families. A different approach to tackle this problem would be to train a multi-class classifier. However, generating different classifiers allows to individually tune the hyperparameters of each of them and also makes the interpretation and analysis of results easier. To train them, the main data set is divided in six smaller data sets, each of them just considering one binary target c&a$_i$. After that, all the instances representing cultures from patients that have stayed less than 48 hours in the ICU, are removed from every of the six data sets.

As indicated in Table I, the number of features is 40 for every data set, considering the respective target feature. The number of instances are 755, 643, 749, 749, 483 and 708 for AMG, CAR, CF4, PAP, POL and QUI data sets, respectively. Since instances represent cultures, and cultures have an intrinsic temporal ordering, instances are sorted in a temporal manner, with older instances at the beginning of the data set and the newer ones towards the end.

The missing values of the data sets are found in the 12 generated features (r&a$_i$ and p&a$_i$). The percentages of missing values for each of the data sets and features are detailed in Table II and Table III.

It is remarkable that the percentages of missing values for p&a$_i$ features are higher than those of r&a$_i$ features. This happens because, in general, during the same time interval the number of cultures associated to a group of patients will be higher than the number cultures associated to just one patient. It is also notable that, overall, p&pol and r&pol have a high percentage of missing values with respect to the rest of the features of their respective type. This is caused by the very few resistant instances there are for POL family , probably because POL started to be tested in 2007 and the rest of antimicrobial families in 2004.

TABLE II. Percentage of Missing Values of the $p\&a_i$ Features for each of the Antimicrobial Families

Fam	p&amg	p&car	p&cf4	p&pap	p&pol	p&qui
AMG	34.97	41.72	34.83	34.83	51.39	38.15
CAR	33.28	36.86	33.13	33.13	48.68	37.01
CF4	35.38	42.06	35.25	35.25	51.80	38.58
PAP	35.25	41.92	35.11	35.11	51.67	38.45
POL	29.81	34.78	29.61	29.61	34.16	34.58
QUI	34.75	41.95	34.75	34.75	52.40	35.03

TABLE III. Percentage of Missing Values of the $r\&a_i$ Features for each of the Antimicrobial Families

Fam	r&amg	r&car	r&cf4	r&pap	r&pol	r&qui
AMG	15.36	20.66	15.23	15.23	33.91	17.88
CAR	16.64	19.60	16.49	16.49	36.24	19.60
CF4	15.35	20.69	15.22	15.22	34.05	18.02
PAP	15.35	20.56	15.22	15.22	33.78	18.02
POL	15.11	19.67	14.91	14.91	20.91	19.05
QUI	14.41	20.20	14.41	14.41	34.32	14.69

In the clinical setting, dealing with missing values is an interesting and challenging topic which may have different implications. In this study, missing values are replaced by zeros because of the clinical meaning of $p\&a_i$ and $r\&a_i$ features. The reason for a $p\&a_i$ feature not having a value is that, for the particular patient and time interval considered, it is not found a resistance test result for the specific antimicrobial family studied. If that is the case, it means that, probably, clinicians have considered that the patient may not be infected by a bacterium resistant to the antimicrobial family. Therefore, it can be inferred that likely, in the time prior of the culture being analyzed, the patient was not infected with a resistant bacterium. It seems reasonable to assign a zero in this case, since the feature gets a higher value the more recent a resistant bacterium was detected. Regarding $r\&a_i$ features, a similar reasoning can be done. If in the time interval observed, none of the patients in the ICU were tested for resistance to the particular antimicrobial family, it implies clinicians considered it was unlikely to find this kind of resistant bacterium. Thus, it is probable that, prior to the culture, there were no patients infected with a bacterium resistant to the feature's antimicrobial family, causing zero to be an appropriate value.

The categorical features in the data sets are converted into numerical before using them with the machine learning methods considered in this work. The two features representing dates (`date_culture` and `start_date`) are categorical and ordered. Because of that, dates are encoded with integers, assigning lower values to older dates, and higher values to recent dates, indicating, in that way, the ordering among them. The value of a particular date is calculated as the difference, in number of days, between the particular date to be encoded and the first date in the data sets of the specific feature.

Having all features expressed as numerical, Pearson correlation is applied to detect the most correlated ones. If two features (both different from the target feature) are highly correlated, they are adding redundant information to the prediction, and therefore one of them should be removed. In this study it is considered that two features are highly correlated if their correlation coefficient is higher than 0.9 or lower than -0.9. In all of the six data sets, the same four features (`date_culture`, `year_culture`, `start_date` and `year_admission`) are highly correlated among them. Because of that, just `date_culture` is maintained and the other three are removed from the data sets. After that, the number of features in every data set is 37 including the target feature.

B. Predictive Methods

In this section, we describe briefly the data-driven classifier considered in this work. Specifically, LR is tested as base line method, and it was also used in our previous work [15]. In this study, RF has been added to carry out predictions since its interpretability capabilities.

The LR method, very common in the clinical literature, allows us to conduct a linear analysis when the dependent variable is binary. It was used in our previous study [15] because of its simplicity to serve as a baseline, and to evaluate the feasibility of learning from data. In this work, it is again used to classify the instances, now with a greater amount of data and a higher number of antimicrobial families to be analyzed. This is done in order to have a more solid insight on whether the target can be predicted with the available features and the performance this method can provide. Before using LR, each feature is standardized by removing the mean and scaling to unit variance.

The another data-driven method explored here is RF, a machine learning approach commonly used for regression and classification [21], [22]. It is an ensemble method, that is, a RF model is built from multiple decision trees named estimators, which are able to generate individual predictions. RF combines the different predictions of its decision trees (which, individually, tend to over fitting to the training set) to provide a better prediction, providing a better generalization to data not considered in training. The RF method is very robust, since it can handle data sets with an extensive number of features, high dimensionality and heterogeneous features, while having very few hyperparameters. Because of this, RF is often used as a first approach to develop machine learning systems, as it enables to get an overview of the performance on a particular task.

C. Temporal Modeling

Analyzing the problem to be solved, some special characteristics have to be considered when designing the experiments.

The first one is the temporal ordering among instances of the data sets. Since instances are associated with cultures with a susceptibility test, they have an inherent order marked by the date when they were collected. This forces to maintain this same order when predicting instances, that is, past instances cannot be predicted with instances in their respective future. This particularity arises from the fact that, in the real world, when predicting an antibiogram result, future results are not available.

Antimicrobial resistance is a phenomenon that changes over time as bacteria mutates. It allows bacteria to be more resistant to antibiotics as time progresses. As previously mentioned, the features considered include demographic data, information about the patient's admission, and information about the culture and antibiogram results. Since bacteria's mutations are not among the available features, the feature's values telling apart one class from another may change along time. This fact has been previously described as the *concept drift* in which the concept being studied depends on some hidden context, not explicitly given in the form of predictive features [6]. An approach that is normally used to tackle this type of problems is the so called *windowing*, which generalizes from a *sliding window* that moves over the data set instances and applies the knowledge gathered to predict only in the immediate future.

The other particularity is the data scarcity. As previously mentioned, the maximum number of cultures (755) is observed for the AMG antimicrobial family. With the time interval considered (15 years, from 2004 to 2019), there is at most an average of 50 cultures per year. Data scarcity is a trouble spot when using windowing, because in this paradigm, usually, just a small fraction of the data set (the one considered by the sliding window at each particular time) is used for training.

A solution proposed in the previous work [15] was to build an *incremental training window* as the one depicted in panel (d) of Fig. 6. This type of window, which grows in length, contains instances that are as temporarily close as possible to the test instances. Then, the concept drift can be avoided by predicting temporarily close instances to the training set, but it also contains instances far in the past, so that the number of available instances for training is higher than when using *sliding window*. In addition to the *incremental training window*, this work considers a more commonly used *sliding training window with fixed size* to compare their prediction performance. Below we first describe the characteristics of the test window, which is the same for both types of training windows. After that, we present the characteristics of the two types of training windows considered in this work.

The test window consists in a sliding window with a fixed size of just 1 month. Considering just a small amount of time, it is ensured that test instances are as close as possible to the training set. In the experiments of this study, this window begins just considering the first month (January) of 2016. After that, in each prediction step, the test window shifts one month towards later dates. In Fig. 6, steps are indicated at the end of each row as (1), (2), (3), ... (N) for every approach. In the last step, this window considers the last month of the data set. The test window, when shifted, does not overlap with its previous position, that is, in each step predicted instances are different from instances predicted in any other step.

The *incremental training window*, as previously mentioned, is a window of increasing size. In the experiments, this window starts containing instances from 2004 to 2015. In the following steps, the window increases in size one month at a time. In the last step, the training window includes all the instances in the data set except the last month, which is the one considered by the test window.

The *sliding training window with a fixed size* consists in a window just considering 4 years of instances. In every step, this window shifts 1 month towards last instances of the data set, in the same way as the test window does. Since the train and test windows always shift the same amount of time, the distance between them, if any, is always the same. The last step, as previously explained, is the one in which the test window considers the last month of the data set. This kind of window is tested with three different configurations, 0 years approach, 2 years approach and 4 years approach, which are represented in panels (a), (b) and (c) of Fig. 6. In the 0 years approach, the distance between the training and test windows is 0 years, that is, the training window is next to the test one. In this case, the training window considers years from 2012 to 2015 in the initial step. In the 2 years approach the distance among windows is 2 years, therefore taking into account that the test window initially contains the first month of year 2016, the training window includes years from 2010 to 2013, so that the desired distance is respected. Similarly, in the 4 years approach, the window starts considering years from 2008 to 2011, because of the same reason. These three different configurations are considered in order to observe how the prediction evolves as the windows move away from each other, and therefore, the concept drift is more noticeable.

For both types of training windows, at each step, a classifier is trained, and the performance is evaluated on a test set with each of the two methods considered (LR and RF). It is relevant to take into account that patients from training and test windows are different. That is, when predicting a particular patient's susceptibility test

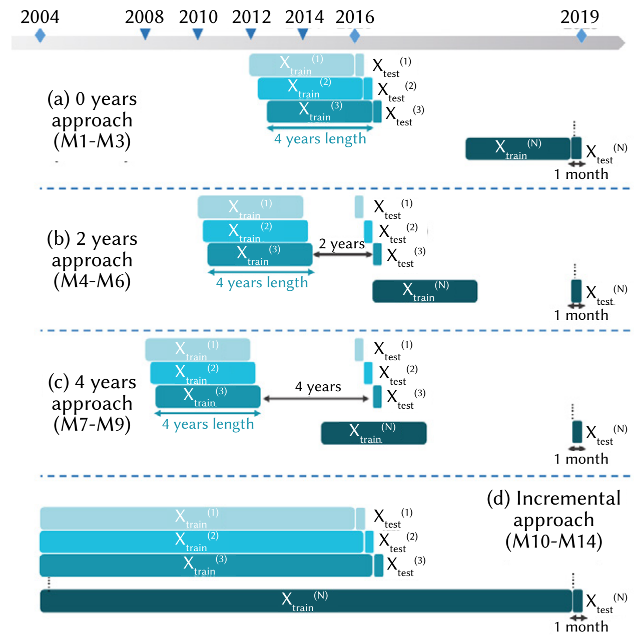


Fig. 6. Sketch for the proposed 14 models (M1 to M14). All models consider a test window of 1 month. Panels (a), (b) and (c) consider a training window of 4 years, with a 1-month sliding training and test windows. Different time slots are considered between the training and the test set: 0 (a), 2 (b) and 4 (c) years. Panel (d) shows an incremental approach for the training set (starting from an initial length of 12 years and incremental steps of 1 month), with the test set immediately after the training set.

result, it is ensured that there are not other susceptibility results of the same patient in the training set. Also, in the approaches where training and test windows are next to each other (as in the *incremental training window and the 0 years approach*), a margin of 48 hours is considered between them, since it is the time required for getting the antibiogram's results.

As the windows traverse the data set, they encounter *class imbalance*, due to the temporal evolution of bacterial resistance. This causes that, in the time interval considered by test windows, there is a higher number of instances from one class. Because of that, in order to evaluate the prediction of the classifiers, is not enough to consider the global accuracy. To get a realistic approximation of the classifier performance, the success in susceptible instances and the success in resistant instances are also calculated. The names assigned to these figures of merit are *Total Accuracy* (A_{Tot}), *Resistant Accuracy* (A_{Rst}) and *Susceptible Accuracy* (A_{Scb}), respectively. For a test window with *Ns* susceptible instances and *Nr* resistant instances, if the method succeeds in predicting *Ss* susceptible instances and *Sr* resistant instances, these figures of merit are computed as follows:

$$A_{Tot} = \frac{Ss + Sr}{Ns + Nr} \quad (5)$$

$$A_{Rst} = \frac{Sr}{Nr} \quad (6)$$

$$A_{Scb} = \frac{Ss}{Ns} \quad (7)$$

These three figures of merit are calculated for the test set of the particular approach considered. In order to get the mean value of these measurements, for every step, the values of *Ns*, *Nr*, *Ss* and *Sr* are accumulated and, at the end, the three figures of merit are obtained. This accumulation is carried out because test windows may have a different amount of instances, due to the fact that not all 1-month time intervals contain the same number of antibiograms. For that reason, an average would not be adequate, since some instances would have more weight than others depending on the number of instances in their test window.

In addition to the experiments using the different windows, a series of experiments are carried out considering different aspects of the prediction. First, it is analyzed the prediction contribution of the most relevant features according to the MI score. In particular, the features studied are `date_culture` and the two groups of features related to $\texttt{p\&a}_i$ and $\texttt{r\&a}_i$. To assess their contribution, the target is predicted with and without considering these features, and the two outcomes are compared.

Secondly, since the *incremental training window* considers a high amount of instances (from the beginning of the data set) it is proposed to assign weights to its training instances. The purpose is to give a higher importance to the training instances that are temporarily closer to the test, which theoretically would have a more similar distribution to the test instances, and lower importance to instances far from the test. Equation (8) details how the weight is generated for each instance.

$$e^{-\lambda(d_l - d_c)} \tag{8}$$

where d_l represents the date of the last culture in the training window, and d_c is the culture date for the instance which weight is being calculated. In the equation, the difference of these two dates is expressed in days. The parameter λ is empirically chosen for each experiment as the one providing the best results among the following: 0, 1e-05, 1e-04, 1e-03, 1e-02, 0.1 and 1. If λ is very small, all instances get a very similar weight, regardless of how far they are from the end of the training window. For instance, for λ = 0, all instances has a weight of 1. On the other hand, if the value of λ is high, only a very few instances very close to the end of the training set get a weight close to 1, and the great majority of instances get a weight very close to 0. Note that when the value of λ is zero, it is the same case as the incremental training window without weights. In the case of high values for λ, it is more similar to the 0 *years approach of the sliding training window with a fixed size*. So, in the end, these weights allow to regulate the amount of past instances considered for prediction.

To encode the models obtained from different combinations of windowing and features, a number is assigned to each model, with the following description:

M1. *Sliding training window with a fixed size* and following the **0 years approach**. It uses **neither** $\texttt{r\&a}_i$ **nor** $\texttt{p\&a}_i$ features.

M2. *Sliding training window with a fixed size* and following the **2 years approach**. It uses **neither** $\texttt{r\&a}_i$ **nor** $\texttt{p\&a}_i$ features.

M3. *Sliding training window with a fixed size* and following the **4 years approach**. It uses **neither** $\texttt{r\&a}_i$ **nor** $\texttt{p\&a}_i$ features.

M4. *Sliding training window with a fixed size* and following the **0 years approach**. It uses $\texttt{r\&a}_i$ features but **not** $\texttt{p\&a}_i$ features.

M5. *Sliding training window with a fixed size* and following the **2 years approach**. It uses $\texttt{r\&a}_i$ features but **not** $\texttt{p\&a}_i$ features.

M6. *Sliding training window with a fixed size* and following the **4 years approach**. It uses $\texttt{r\&a}_i$ features but **not** $\texttt{p\&a}_i$ features.

M7. *Sliding training window with a fixed size* and following the **0 years approach**. It uses **both** $\texttt{r\&a}_i$ **and** $\texttt{p\&a}_i$ features.

M8. *Sliding training window with a fixed size* and following the **2 years approach**. It uses **both** $\texttt{r\&a}_i$ **and** $\texttt{p\&a}_i$ features.

M9. *Sliding training window with a fixed size* and following the **4 years approach**. It uses **both** $\texttt{r\&a}_i$ **and** $\texttt{p\&a}_i$ features.

M10. *Incremental training window*. It uses **neither** $\texttt{r\&a}_i$ **nor** $\texttt{p\&a}_i$ features.

M11. *Incremental training window*. It uses $\texttt{r\&a}_i$ features but **not** $\texttt{p\&a}_i$ features.

M12. *Incremental training window*. It uses **both** $\texttt{r\&a}_i$ **and** $\texttt{p\&a}_i$ features.

M13. *Incremental training window* with instance weighting. It uses $\texttt{r\&a}_i$ features but **not** $\texttt{p\&a}_i$ features.

M14. *Incremental training window* with instance weighting. It uses **both** $\texttt{r\&a}_i$ **and** $\texttt{p\&a}_i$ features.

Each of the above kind of models are designed with and without considering the `date_culture` feature, also with the two aforementioned machine learning methods, LR and RF.

After studying the outcomes of the different experiments, the feature relevance is calculated again, now with an embedded method from the RF model. Also, `date_culture` and the $\texttt{p\&a}_i$ set of features is analyzed in more depth by making the predictions with just one of these features at a time.

TABLE IV. Prediction Accuracy Results for the **AMG** Antimicrobial Family. Column *M* Indicates the Model, Column *DC* Refers to Whether `date_culture` Is Used (✓) or Not (X). In the Three Left/Right Grouped Columns, LR/RF Is Used

M	DC	LR			RF		
		A_{Tot}	A_{Rst}	A_{Scb}	A_{Tot}	A_{Rst}	A_{Scb}
1	✓	62.5	75.76	22.73	54.55	66.67	18.18
2	✓	**75.0**	**83.33**	**50.0**	64.77	83.33	9.09
3	✓	75.0	96.97	9.09	53.41	60.61	31.82
4	✓	60.23	74.24	18.18	60.23	78.79	4.55
5	✓	73.86	81.82	50.0	70.45	90.91	9.09
6	✓	76.14	96.97	13.64	54.55	62.12	31.82
7	✓	59.09	71.21	22.73	76.14	93.94	22.73
8	✓	73.86	80.30	54.55	73.86	90.91	22.73
9	✓	81.82	96.97	36.36	73.86	81.82	50.0
10	✓	73.86	98.48	0.0	51.14	62.12	18.18
11	✓	73.86	98.48	0.0	64.77	83.33	9.09
12	✓	77.27	96.97	18.18	**76.14**	**87.88**	**40.91**
13	✓	62.5	77.27	18.18	62.5	74.24	27.27
14	✓	75.0	89.39	31.82	**81.82**	**90.91**	**54.55**
1	X	60.23	69.7	31.82	62.5	77.27	18.18
2	X	67.05	72.73	50.0	63.64	80.3	13.64
3	X	29.55	27.27	36.36	36.36	34.85	40.91
4	X	60.23	74.24	18.18	65.91	87.88	0.0
5	X	65.91	72.73	45.45	68.18	86.36	13.64
6	X	31.82	27.27	45.45	34.09	34.85	31.82
7	X	65.91	75.76	36.36	80.68	98.48	27.27
8	X	**73.86**	**77.27**	**63.64**	75.0	90.91	27.27
9	X	69.32	69.7	68.18	65.91	63.64	72.73
10	X	34.09	22.73	68.18	39.77	30.30	68.18
11	X	37.5	24.24	77.27	38.64	24.24	81.82
12	X	73.86	71.21	81.82	**77.27**	**72.73**	**90.91**
13	X	60.23	69.7	31.82	56.82	69.7	18.18
14	X	76.14	78.79	68.18	**79.55**	**75.76**	**90.91**

IV. Results and Discussion

The Results and Discussion section is divided in two different subsections. In the Subsection A, the performance of the predictive methods is assessed by considering different experiments. In the Subsection B, the features identified as the most relevant along the study are further analyzed.

A. Prediction

The prediction results are detailed in Tables IV, V, VI, VII, VIII and

TABLE V. Prediction Accuracy Results for the **CAR** Antimicrobial Family. Column *M* Indicates the Model, Column *DC* Refers to Whether `date_culture` Is Used (✓) or Not (X). In the Three Left/Right Grouped Columns, LR/RF Is Used

M	DC	LR A_{Tot}	LR A_{Rst}	LR A_{Scb}	RF A_{Tot}	RF A_{Rst}	RF A_{Scb}
1	✓	**93.18**	**98.78**	**16.67**	93.18	100.0	0.0
2	✓	90.91	97.56	0.0	92.05	98.78	0.0
3	✓	88.64	95.12	0.0	80.68	86.59	0.0
4	✓	93.18	98.78	16.67	93.18	100.0	0.0
5	✓	88.64	95.12	0.0	90.91	97.56	0.0
6	✓	89.77	96.34	0.0	77.27	82.93	0.0
7	✓	93.18	98.7805	16.67	93.18	100.0	0.0
8	✓	89.77	96.34	0.0	89.77	96.34	0.0
9	✓	88.64	95.12	0.0	72.73	78.05	0.0
10	✓	**93.18**	**100.0**	**0.0**	93.18	100.0	0.0
11	✓	93.18	100.0	0.0	93.18	100.0	0.0
12	✓	93.18	100.0	0.0	93.18	100.0	0.0
13	✓	**94.32**	**100.0**	**16.67**	93.18	98.78	16.67
14	✓	94.32	100.0	16.67	92.05	97.56	16.67
1	X	90.91	97.56	0.0	93.18	100.0	0.0
2	X	84.09	90.24	0.0	93.18	100.0	0.0
3	X	61.36	65.85	0.0	77.27	82.93	0.0
4	X	89.77	95.12	16.67	93.18	100.0	0.0
5	X	81.82	87.80	0.0	90.91	97.56	0.0
6	X	55.68	59.76	0.0	56.82	60.98	0.0
7	X	88.64	93.90	16.67	93.18	100.0	0.0
8	X	79.55	85.37	0.0	85.23	91.46	0.0
9	X	**68.18**	**70.73**	**33.33**	69.32	73.17	16.67
10	X	60.23	60.98	50.0	61.36	65.85	0.0
11	X	50.0	52.44	16.67	51.14	54.88	0.0
12	X	**75.0**	**74.39**	**83.33**	79.55	81.71	50.0
13	X	**94.32**	**100.0**	**16.67**	93.18	98.78	16.67
14	X	92.05	97.56	16.67	93.18	98.78	16.67

TABLE VI. Prediction Accuracy Results for the **CF4** Antimicrobial Family. Column *M* Indicates the Model, Column *DC* Refers to Whether `date_culture` Is Used (✓) or Not (X). In the Three Left/Right Grouped Columns, LR/RF Is Used

M	DC	LR A_{Tot}	LR A_{Rst}	LR A_{Scb}	RF A_{Tot}	RF A_{Rst}	RF A_{Scb}
1	✓	53.93	74.14	16.13	50.56	60.34	32.26
2	✓	52.81	48.28	61.29	46.07	39.66	58.06
3	✓	35.96	10.34	83.87	41.57	13.79	93.55
4	✓	57.30	74.14	25.81	46.07	53.45	32.26
5	✓	49.44	43.10	61.29	39.33	32.76	51.61
6	✓	34.83	5.17	90.32	37.08	5.17	96.77
7	✓	64.04	82.76	29.03	67.42	84.48	35.48
8	✓	**60.67**	**55.17**	**70.97**	50.56	53.45	45.16
9	✓	46.07	18.97	96.77	49.44	36.21	74.19
10	✓	52.81	62.07	35.48	55.06	68.97	29.03
11	✓	46.07	56.9	25.81	38.20	50.0	16.13
12	✓	**61.8**	**74.14**	**38.71**	59.55	74.14	32.26
13	✓	58.43	81.03	16.13	55.06	67.24	32.26
14	✓	61.8	74.14	38.71	**61.8**	**72.41**	**41.94**
1	X	58.43	65.52	45.16	47.19	48.28	45.16
2	X	47.19	37.93	64.52	48.31	24.14	93.55
3	X	47.19	24.14	90.32	35.96	3.45	96.77
4	X	58.43	65.52	45.16	49.44	51.72	45.16
5	X	51.69	34.48	83.87	33.71	17.24	64.52
6	X	42.7	17.24	90.32	31.46	3.45	83.87
7	X	62.92	75.86	38.71	61.8	74.14	38.71
8	X	**64.04**	**53.45**	**83.87**	50.56	41.38	67.74
9	X	49.44	25.86	93.55	47.19	31.03	77.42
10	X	35.96	24.14	58.06	40.45	10.34	96.77
11	X	37.08	31.03	48.39	33.71	13.79	70.97
12	X	52.81	46.55	64.52	**56.18**	**51.72**	**64.52**
13	X	51.69	56.9	41.94	57.30	70.69	32.26
14	X	59.55	63.79	51.61	**60.67**	**58.62**	**64.52**

IX for AMG, CAR, CF4, PAP, POL and QUI families, respectively. The best results are in bold. For each table and models considering or not the `date_culture` feature, three results are marked: the best result among models from M1 to M9, the best result from M10 to M12 and the best result from M13 to M14. Table X shows the chosen values for the λ hyperparameter (instance weighting). The prediction models, identified in column *M* in Tables from IV to IX, are analyzed in three different groups according to the type of temporal window. Firstly, the experiments with an *sliding training window with fixed size* are discussed, with the impact of the distance between training and test windows becoming manifest. Secondly, the results obtained using an *incremental training window* are studied. Finally, we evaluate whether results of the *incremental training window* can be improved by an instance weighting approach.

1. Sliding Training Windows with Temporal Distance Variation Among Training and Test Windows

The figures of merit provided by models considering the temporal distance between the training and test sets are in rows with numbers 1 to 9 in the *M* column of Tables from IV to IX.

In the case of the LR method when considering the feature `date_culture`, the evolution of the figures of merit is not consistent among antimicrobial families when analyzing the separation between training and test windows. In some families, the *Total Accuracy* increases as the training window approaches the test window, while the opposite happens for other families. The same is observed with *Resistant Accuracy* and *Susceptible Accuracy*, its behavior varies depending on the antimicrobial family being predicted.

Predicting with RF and using feature `date_culture`, the evolution of the figures of merit is more similar among the different antimicrobial families. In general, *Total Accuracy* increases, *Resistant Accuracy* increases and *Susceptible Acurracy* decreases as the training window approaches test window. When this pattern is less evident, it may be helpful to analyze when both $\texttt{r\&a}_i$ and $\texttt{p\&a}_i$ features are considered. Also, the general performance of the three figures of merit appears to be better when both $\texttt{r\&a}_i$ and $\texttt{p\&a}_i$ features are used.

For LR and **not using** the feature `date_culture`, the aforementioned pattern appears, in which *Total Accuracy* increases, *Resistant Accuracy* increases and *Susceptible Accuracy* decreases when reducing the distance between windows. Comparing these results with those provided by LR and `date_culture`, two remarks deserve to be underscored: for the families in which this pattern was not previously evident (such as AMG, CAR and QUI), now windows 4 and 2 years apart have lower *Total Accuracy* and lower *Resistant Accuracy*, with similar figures of merit in the 0 years-apart windows; on the other hand, for the families where this pattern was reasonably evident (such as CF4, PAP and POL), the figures of merit usually improve, while maintaining the same pattern. Also using both the $\texttt{r\&a}_i$ and $\texttt{p\&a}_i$ features tend to improve the performance.

Considering RF for prediction and **not using** the feature `date_culture`, the same behavior as in LR without `date_culture`, is

TABLE VII. Prediction Accuracy Results for the **PAP** Antimicrobial Family. Column *M* Indicates the Model, Column *DC* Refers to Whether `date_culture` Is Used (✓) or Not (X). In the Three Left/Right Grouped Columns, LR/RF Is Used

M	DC	LR A_{Tot}	LR A_{Rst}	LR A_{Scb}	RF A_{Tot}	RF A_{Rst}	RF A_{Scb}
1	✓	50.56	66.04	27.78	55.06	84.91	11.11
2	✓	60.67	77.36	36.11	51.69	64.15	33.33
3	✓	46.07	52.83	36.11	35.96	24.53	52.78
4	✓	50.56	66.04	27.78	46.07	69.81	11.11
5	✓	**65.17**	**62.26**	**69.44**	59.55	52.83	69.44
6	✓	47.19	49.06	44.44	37.08	20.75	61.11
7	✓	61.8	83.02	30.56	68.54	86.79	41.67
8	✓	67.42	79.25	50.0	68.54	81.13	50.0
9	✓	56.18	58.49	52.78	60.67	58.49	63.89
10	✓	64.04	98.11	13.89	52.81	67.92	30.56
11	✓	61.8	96.23	11.11	39.33	47.17	27.78
12	✓	65.17	98.11	16.67	**67.42**	**75.47**	**55.56**
13	✓	64.04	96.23	16.67	50.56	56.60	41.67
14	✓	68.54	90.57	36.11	**67.42**	**75.47**	**55.56**
1	X	55.06	64.15	41.67	50.56	67.92	25.0
2	X	58.43	64.15	50.0	43.82	37.74	52.78
3	X	47.19	45.28	50.0	40.45	22.64	66.67
4	X	52.81	66.04	33.33	46.07	66.04	16.67
5	X	57.30	64.15	47.22	47.19	39.62	58.33
6	X	49.44	47.17	52.78	34.83	16.98	61.11
7	X	61.8	73.58	44.44	67.42	86.79	38.89
8	X	**66.29**	**67.92**	**63.89**	68.54	77.36	55.56
9	X	55.06	49.06	63.89	62.92	58.49	69.44
10	X	39.33	28.30	55.56	44.94	20.75	80.56
11	X	37.08	22.64	58.33	32.58	11.32	63.89
12	X	**70.79**	**67.92**	**75.0**	69.66	71.7	66.67
13	X	53.93	62.26	41.67	51.69	56.60	44.44
14	X	**71.91**	**69.81**	**75.0**	70.79	69.81	72.22

TABLE VIII. Prediction Accuracy Results for the **POL** Antimicrobial Family. Column *M* Indicates the Model, Column *DC* Refers to Whether `date_culture` Is Used (✓) or Not (X). In the Three Left/Right Grouped Columns, LR/RF Is Used

M	DC	LR A_{Tot}	LR A_{Rst}	LR A_{Scb}	RF A_{Tot}	RF A_{Rst}	RF A_{Scb}
1	✓	**68.97**	**63.33**	**71.93**	63.22	0.0	96.49
2	✓	44.83	6.67	64.91	65.52	0.0	100.0
3	✓	47.13	0.0	71.93	65.52	0.0	100.0
4	✓	67.82	63.33	70.18	66.67	3.33	100.0
5	✓	49.43	6.67	71.93	65.52	0.0	100.0
6	✓	50.57	0.0	77.19	65.52	0.0	100.0
7	✓	66.67	63.33	68.42	65.52	3.33	98.25
8	✓	54.02	6.67	78.95	65.52	0.0	100.0
9	✓	52.87	0.0	80.70	65.52	0.0	100.0
10	✓	58.62	13.33	82.46	65.52	0.0	100.0
11	✓	**63.22**	**30.0**	**80.70**	65.52	0.0	100.0
12	✓	56.32	23.33	73.68	66.67	3.33	100.0
13	✓	**72.41**	**60.0**	**78.95**	73.56	46.67	87.72
14	✓	65.52	56.67	70.18	59.77	23.33	78.95
1	X	**74.71**	**63.33**	**80.70**	65.52	0.0	100.0
2	X	56.32	0.0	85.96	65.52	0.0	100.0
3	X	64.37	0.0	98.25	65.52	0.0	100.0
4	X	72.41	60.0	78.95	64.37	0.0	98.25
5	X	58.62	0.0	89.47	65.52	0.0	100.0
6	X	60.92	0.0	92.98	65.52	0.0	100.0
7	X	70.11	60.0	75.44	64.37	0.0	98.25
8	X	57.47	0.0	87.72	65.52	0.0	100.0
9	X	60.92	0.0	92.98	65.52	0.0	100.0
10	X	65.52	0.0	100.0	65.52	0.0	100.0
11	X	63.22	0.0	96.49	65.52	0.0	100.0
12	X	**64.37**	**6.67**	**94.74**	65.52	0.0	100.0
13	X	**65.52**	**56.67**	**70.18**	68.97	33.33	87.72
14	X	65.52	56.67	70.18	64.37	26.67	84.21

observed for all antimicrobial families: note the same pattern for the evolution of the figures of merit (*Total Accuracy* increases, *Resistant Accuracy* increases and *Susceptible Accuracy* decreases as the distance between train and test windows decreases). Comparing these results to previous ones of RF using `date_culture`, it is noticed that now, for all families, windows of 4 and 2 years apart have lower *Total Accuracy* and lower *Resistant Accuracy*, with similar or improved figures of merit in the 0 years-apart windows. Furthermore, using both $\texttt{r\&a}_i$ and $\texttt{p\&a}_i$ features tend to provide a better performance.

In the considered experiments (from model 1 to model 9), it is also noticeable how results change depending on the antimicrobial family. It is specially remarkable for the CAR and POL families. Considering CAR, it is observed that, for the majority of models, the values of *Total Accuracy* and *Resistant Accuracy* are very high, while *Susceptible Accuracy* values are very low, in most cases zero. On the other hand, for the POL family, *Total Accuracy* and *Susceptible Accuracy* are very high and *Resistant Accuracy* is low in general, with many zero values. These results suggest that the outcomes depend on the class distribution along time, for each antimicrobial family. In Fig. 3 it is noticed that CAR is the family with the highest ratio of resistant instances (almost 1 for the last years of the data set), and POL is the family with the lowest ratio of resistant instances. Although less obvious, the rest of the families also appear to be influenced by their respective class distribution.

Firstly, it is interesting to discuss the common pattern observed in almost all families, which causes *Total Accuracy* to increase, *Resistant Accuracy* to increase and *Susceptible Accuracy* to decrease as the distance between train and test windows gets smaller. The reason of this behavior is the *temporal class imbalance*, that is, in the first years of the data set, the majority of instances belong to the susceptible class, but as time progresses, the majority of instances become resistant, as it is depicted in Fig. 3. Using *sliding training windows with fixed size* and the approach with 4 years of distance between windows, the training window has to shift towards the past since the test window starts in 2016 for all experiments, therefore containing years from 2008 to 2011 for the first step of the training window, as explained in Section III.C. Being in the past, it contains a higher number of susceptible instances compared to resistant ones, which causes to perform better in predicting susceptible instances (better *Susceptible Accuracy*) and worse in predicting resistant instances (worse *Resistant Accuracy*). The opposite happens when the distance between windows is 0 years. In this case the window is near the last years of the data set, therefore it contains more resistant instances (improving *Resistant Accuracy*) and less susceptible instances (decreasing *Susceptible Accuracy*). The *Total Accuracy* improves when the distance is small because in test window the majority of instances are, mostly, resistant. If the majority class is well predicted, the *Total Accuracy* is high. We conclude that not all the three figures of merit improve as expected when distance is diminishing, in fact one of them gets worse. Applying oversampling to the minority class in this kind of fixed-size temporal windows, in order to balance the number of the two kind of instances, could improve the

TABLE IX. Prediction Accuracy Results for the **QUI** Antimicrobial Family. Column *M* Indicates the Model, Column *DC* Refers to Whether `date_culture` Is Used (✓) or Not (X). In the Three Left/Right Grouped Columns, LR/RF Is Used

M	DC	LR			RF		
		A_{Tot}	A_{Rst}	A_{Scb}	A_{Tot}	A_{Rst}	A_{Scb}
1	✓	62.26	68.75	0.0	88.68	97.92	0.0
2	✓	66.04	70.83	20.0	71.7	77.08	20.0
3	✓	90.57	97.92	20.0	50.94	50.0	60.0
4	✓	66.04	72.92	0.0	88.68	97.92	0.0
5	✓	66.04	70.83	20.0	71.7	79.17	0.0
6	✓	92.45	100.0	20.0	39.62	33.33	100.0
7	✓	84.91	93.75	0.0	90.57	100.0	0.0
8	✓	77.36	83.33	20.0	84.91	89.58	40.0
9	✓	90.57	97.92	20.0	**83.02**	**81.25**	**100.0**
10	✓	**88.68**	**97.92**	**0.0**	67.92	75.0	0.0
11	✓	88.68	97.92	0.0	83.02	91.67	0.0
12	✓	88.68	97.92	0.0	86.79	95.83	0.0
13	✓	88.68	95.83	20.0	90.57	97.92	20.0
14	✓	88.68	95.83	20.0	**92.45**	**100.0**	**20.0**
1	X	50.94	56.25	0.0	84.91	93.75	0.0
2	X	60.38	64.58	20.0	77.36	83.33	20.0
3	X	67.92	72.92	20.0	28.30	22.92	80.0
4	X	62.26	68.75	0.0	81.13	89.58	0.0
5	X	67.92	72.92	20.0	71.7	79.17	0.0
6	X	50.94	54.17	20.0	30.19	25.0	80.0
7	X	77.36	85.42	0.0	86.79	95.83	0.0
8	X	79.25	85.42	20.0	83.02	89.58	20.0
9	X	75.47	77.08	60.0	**83.02**	**81.25**	**100.0**
10	X	54.72	60.42	0.0	33.96	31.25	60.0
11	X	54.72	60.42	0.0	49.06	47.92	60.0
12	X	**79.25**	**79.17**	**80.0**	75.47	77.08	60.0
13	X	90.57	95.83	40.0	88.68	95.83	20.0
14	X	79.25	79.17	80.0	**83.02**	**83.33**	**80.0**

accuracy in the minority class.

Secondly, it is relevant the change in behavior of prediction when `date_culture` is not considered in both LR and RF methods. Overall, when using `date_culture` for prediction in the 4 years and 2 years approaches, the *Resistant Accuracy* increases and the *Susceptible Accuracy* decreases compared to models not using `date_culture`. This probably happens because `date_culture` is compensating the lack of resistant instances of training windows in 4 and 2 years approaches, by telling the classifier the most probable class in test years, which tend to be resistant, and hence *Resistant Accuracy* is high in most cases, causing *Susceptible Accuracy* to decrease. The disadvantage of using `date_culture` is that it causes the minority class to worsen its prediction, since it introduces bias towards classifying instances as the most probable class of the time interval. Since, in the 0 years approach, without considering the `date_culture` feature, the results are similar or better than when `date_culture` is taken into account, we conclude that it is convenient not to use this feature.

2. Incremental Window

The experiments concerning the results of prediction by using an incremental training window are in rows with numbers from 10 to 12 in the *M* column of Tables from IV to IX.

In the case of using the LR method and including the feature `date_culture`, adding just features $\texttt{r\&a}_i$ does not generally improve figures of merit. With the addition of both features $\texttt{r\&a}_i$ and $\texttt{p\&a}_i$, half of the antimicrobial families (AMG, CF4 and PAP) improve their results, although this improvement is mild.

With RF and using the `date_culture` feature, the inclusion of the $\texttt{r\&a}_i$ features does not improve performance. Conversely, adding $\texttt{r\&a}_i$ and $\texttt{p\&a}_i$ features improves results in 5 out of the 6 families (AMG, CF4, PAP, POL and QUI), with no worsening of the figures of merit of the CAR family.

For both LR and RF models without `date_culture`, it is noticed that including just the $\texttt{r\&a}_i$ features does not provide an improvement in performance. However, taking into account both the $\texttt{r\&a}_i$ and $\texttt{p\&a}_i$ features, there is a significant improvement for almost all antimicrobial families. *Total Accuracy* and *Resistant Accuracy* are, in general, considerably lower when $\texttt{r\&a}_i$ and $\texttt{p\&a}_i$ features are not used together, in comparison with the results provided by including `date_culture`.

Taking into account the results with *sliding windows of fixed size* of 4 years and the current ones with an *incremental training window*, it is observed that, in general, the best results are obtained with an *incremental training window.* Though for some antimicrobial families, a specific combination of sliding windows can outperform the results of the *incremental training window*, there is not a common approach of sliding windows with better results for all families. Furthermore, when the *incremental training window* outperforms, it is for very little. The exception is the POL antimicrobial family, which achieves clearly better results with the 0 years approach. With the *incremental training window*, best results are mostly achieved by not including `date_culture`, and adding both the $\texttt{r\&a}_i$ and $\texttt{p\&a}_i$ features. This confirms that the use of *incremental training window* represents a useful temporal approach to tackle the task presented in this study.

It is notable that, although MI suggested that the set of $\texttt{r\&a}_i$ features contain relevant information to predict the targets, its use in conjunction with other features does not appear to improve performance. On the other hand, the $\texttt{p\&a}_i$ features show a great potential to predict the result of the susceptibility test, since they improve performance in almost all cases.

It is also worth to analyze the fact that, if `date_culture` is not used, *Total Accuracy* and *Resistant Accuracy* get a low value when the $\texttt{r\&a}_i$ and $\texttt{p\&a}_i$ features are not jointly used, in comparison with the results obtained by using `date_culture`. The reason of this behavior is similar as the one indicated in previous experiments when not using the `date_culture` feature. Without `date_culture`, classifiers tend to predict much of the test instances as susceptible, because it is usually the majority class in incremental training windows (windows starting at the beginning of the data set). The `date_culture` feature compensates this by introducing bias towards predicting the majority class in the time interval, which in test (near the end of the data set) is resistant. In any case, using `date_culture` worsens the *Susceptible Accuracy.* By adding the $\texttt{p\&a}_i$ features, it is not necessary to count with `date_culture` to get a good performance. Moreover, results with $\texttt{p\&a}_i$ features and without `date_culture`, improve both *Resistant Accuracy* and *Susceptible Accuracy* because this kind of features do not introduce a temporal bias towards one of the two classes.

3. Incremental Window with Weights

The prediction results using an incremental training window and instance weighting are in rows with numbers 13 and 14 in the M column of Tables from IV to IX. The λ values for each particular case are expressed in Table X.

It is observed that, using instance weighting, results improve for most of the antimicrobial families. The following are the best figures of merit of $A_{Tot} - A_{Rst} - A_{Scb}$ provided by applying instance weighting:

- **AMG**: 79.55%-75.76%-90.91%. Obtained using RF, without `date_culture` and with both the $\texttt{r\&a}_i$ and $\texttt{p\&a}_i$ sets of features. The weight hyperparameter is λ =1e-05.

TABLE X. Values of the Hyperparameter λ for Results of M13 and M14 in TABLEs from IV to IX. The Column *Fam* Specifies the Family Being Predicted, and Column DC Whether `date_culture` Is Taken into Account. The Two Left/Right Columns Refer to the LR/RF Methods. Columns M13 and M14 Indicate the Model for Which λ is Chosen

		LR		RF	
FAM	DC	M13	M14	M13	M14
AMG	✓	1e-03	1e-03	1e-04	1e-05
AMG	X	1e-03	1e-03	1	1e-05
CAR	✓	1e-02	1e-02	1	1
CAR	X	1e-02	0.1	1	1
CF4	✓	1	0	1	1e-05
CF4	X	1e-03	1e-03	1	1e-05
PAP	✓	1e-04	1e-03	1	0
PAP	X	1e-03	1e-05	1	1e-04
POL	✓	1e-03	0.1	1	1
POL	X	0.1	0.1	1	1
QUI	✓	0.1	0.1	1e-02	1e-02
QUI	X	0.1	0	1	1e-02

- **CAR**: 94.32%-100.0%-16.67%. Obtained using LR, with or without `date_culture` and with the $r\&a_i$ set of features. The weight hyperparameter is λ =1e-02.
- **CF4**: 60.67%-58.62%-64.52%. Obtained using RF, without `date_culture` and with both the $r\&a_i$ and $p\&a_i$ sets of features. The weight hyperparameter is λ =1e-05.
- **PAP**: 71.91%-69.81%-75.0%. Obtained using LR, without `date_culture` and with both the $r\&a_i$ and $p\&a_i$ sets of features. The weight hyperparameter is λ =1e-05.
- **POL**: 72.41%-60.0%-78.95%. Obtained using LR, with `date_culture` and with just the $r\&a_i$ set of features. The weight hyperparameter is λ =1e-03.
- **QUI**: 83.02%-83.33%-80.0%. Obtained using RF, without `date_culture` and with both the $r\&a_i$ and $p\&a_i$ sets of features. The weight hyperparameter is λ =1e-02.

Our results show that M13 and M14 performance, in the majority of families, improves or is maintained when the $p\&a_i$ set of features is taken into account, confirming what was observed in the two previous groups of experiments. The only exception to that is the POL antimicrobial family. When the `date_culture` feature is used, just the POL family gets better results; in any other case, it is better to not consider this feature. The substantially different behavior of POL is probably due to the very small number of resistant instances for this family, which makes it very dependent on the `date_culture` feature. Besides that, for half of the families (CAR, PAP and POL), the best method is LR, while for the other half (AMG, CF4 and QUI), RF gets the best results.

It is also important to analyze the hyperparameter λ used to assign weights to instances. As previously explained, when the value of λ is small, a greater number of instances get a similar high weight (close to 1); otherwise, when λ is high, just a few instances, temporally close to the test set, get a high weight and the rest of instances get very small weights. For AMG, CF4 and PAP, λ is very small and results are very similar to those of the respective incremental window without weights. This happens because almost all instances are being considered. On the other hand, families CAR, POL and QUI, with a greater λ, show results that are, mostly, more similar to the respective sliding training window with a fixed size than to the incremental window.

Comparing the results of the incremental window with the performance for the rest of experiments, it is noticed that it improves the results for 3 of the 6 families, which are AMG, PAP and QUI. In the case of CAR, the whole incremental training window achieves better results than the version with weights. As before, the family CF4 gets better performance with a specific combination of sliding windows, probably because some particularity of its distribution; POL notably gets its best result with the 0 years approach windows, without `date_culture` and with neither the $r\&a_i$ nor $p\&a_i$ sets of features.

B. Relevant Features Analysis

Taking into account previous results, it seems that some features with high MI score, such as $r\&a_i$, do not help to predict the target feature. The feature `date_culture`, which has the highest MI score, increases the performance in some particular cases, but also introduces bias, and the best results in previous experiments are achieved when this feature is not used. On the other hand, the set of features $p\&a_i$, also with high MI scores, appears to improve performance in almost all antimicrobial families.

Our analysis reveals the inconsistency between features ranked as relevant according to MI and those that actually increase prediction performance. In order to contrast feature relevance, they are now obtained with an embedded method. Since RF has been used as classifier, tree-based estimators have been selected to compute the new feature importance, with Fig. 7 showing the ranking in relevance. Now, the most relevant feature for AMG, CAR, CF4, PAP and QUI are `p&amg`, `p&car`, `p&cf4`, `p&pap` and `p&qui`, respectively. In the case of POL, `p&pol` is ranked on the 7th position. Regarding `date_culture`, it is still very important. In the case of POL, `date_culture` is the most important one. The set of features $r\&a_i$ are not considered important overall.

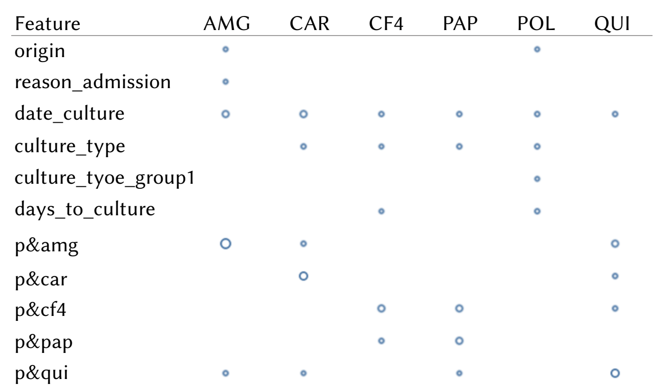


Fig. 7. For each antimicrobial family, the five features with the highest RF relevance scores, indicated by the circle size, from relevance=0.19 (biggest size, pair `p&amg`-AMG) to relevance=0.03 (smallest size, pair reason_admission-AMG).

The new ranking in feature relevance agrees to a greater extent with the prediction performance observed. The set of $p\&a_i$ features are the most important ones, except for the POL family, where the most relevant feature is `date_culture`. These results make sense, since `date_culture` was the only feature improving performance in the POL family, due to small number of resistant instances. Also, the $r\&a_i$ features get low relevance values, as expected. The reason why this method provides more insightful results is probably because it takes into account all other features in the data set, while in MI the feature relevance is calculated separately for each feature.

To further analyze the impact of the most relevant features, the antibiogram result has been predicted using just one feature. Two experiments have been carried out, each for one of the most important features in the data set (the $p\&a_i$ features and `date_culture`). Results with the respective $p\&a_i$ features are detailed in Table XI, showing that the performance of both LR and RF is very similar and the figures of merit are relatively high for most of the families. This evidences the high prediction power of this kind of features, even when using for prediction just one of them. Table XII presents the results with just date_culture. We observe that the prediction is dramatically biased towards the majority class when the LR method is considered, which in most cases is resistant due to the fact that test instances are in the future with respect to training instances. In the case of the POL antimicrobial family, results are biased towards the susceptible class since it generally is the majority class. Using RF, prediction is also biased, although to a lesser extent. As expected, the only family improving its performance when using just date_culture feature is POL.

TABLE XI. Results Using Just the Respective $p\&a_i$ Feature When Predicting the Antibiogram Result for Every Antimicrobial Family (Column *Fam*). For Instance, Just p&amg Is Used to Predict Resistance to the AMG Family. In the Three Left/Right Grouped Columns, the LR/RF Method Is Applied

	LR			RF		
Fam	A_{Tot}	A_{Rst}	A_{Scb}	A_{Tot}	A_{Rst}	A_{Scb}
AMG	80.68	74.24	100.0	80.68	74.24	100.0
CAR	63.64	62.2	83.33	62.5	60.98	83.33
CF4	65.17	65.52	64.52	64.04	63.79	64.52
PAP	70.79	66.04	77.78	70.79	66.04	77.78
POL	62.07	0.0	94.74	63.22	0.0	96.49
QUI	73.58	70.83	100.0	73.58	70.83	100.0

TABLE XII. Results Using Just the date_culture Feature When Predicting the Antibiogram Result for Every Antimicrobial Family (Column *Fam*). In the Three Left/Right Grouped Columns, the LR/RF Method Is Applied

	LR			RF		
Fam	A_{Tot}	A_{Rst}	A_{Scb}	A_{Tot}	A_{Rst}	A_{Scb}
AMG	75.0	100.0	0.0	56.82	66.67	27.27
CAR	93.18	100.0	0.0	90.91	96.34	16.67
CF4	65.17	100.0	0.0	51.69	60.34	35.48
PAP	59.55	100.0	0.0	57.3	54.72	61.11
POL	65.52	0.0	100.0	66.67	56.67	71.93
QUI	90.57	100.0	0.0	88.68	95.83	20.0

V. Conclusions

One important and increasing problem in daily operation of worldwide health systems, and in particular, of hospitals is antimicrobial resistance. This resistance in some microorganisms (bacterium, viruses, etc.) appears when these microorganisms become to be resistant to antimicrobial drugs to which they were susceptible before. This change is due to a mutation of the microorganism or to the acquisition of the resistance gen. This problem is even more difficult in hospital ICUs, due to the critical condition of those patients. Therefore, a reliable and anticipated prediction for a given bacterium of being resistant or not to one or more antimicrobial families in a patient culture would greatly help physicians in their fight against those microorganisms.

In this study, a real anonymized data set with information about patients staying at the ICU in the University Hospital of Fuenlabrada (UHF) has been used. The data set is related to 3812 admissions of 3346 ICU patients, collected at the UHF during a period of 15 consecutive years (from July 2004 to May 2019). The collected data set from UHF was browsed to generate the final data set under study with the information regarding the patients and their different cultures. Originally there were 40 features, but after the application of some pre-processing techniques they were reduced to 37 to avoid the use of high correlated features.

The analysis have been focused on the *Pseudomonas Aeruginosa* bacteria because is one of the most dangerous bacteria in the ICU and its proved ability to develop multi-drug resistance. Furthermore, six antimicrobial families were considered: *Aminoglycosides* (AMG), *Carbapenems* (CAR), *4th Generation Cephalosporins* (CF4), *Extended-spectrum Penicillins* (PAP), *Polymyxins* (POL) and *Quinolones* (QUI).

Logistic Regression and Random Forest models were tested. Different temporal modeling strategies were proposed based on different windowing schemes (sliding training window, incremental training window) to capture the concept drift phenomenon related to the resistance process of microorganisms. In addition, some new temporally-oriented features ($p\&a_i$ and $r\&a_i$ features) capturing the resistance/susceptibility information regarding past cultures of the same patient or regarding the other patients were proposed and evaluated to improve the prediction accuracy of the different models. A temporal weighting scheme of the instances was proposed and it improved the prediction accuracy. Using or not some important features, according to the MI score, like date_culture, $p\&a_i$ features and $r\&a_i$ features were tested in fourteen models (M1 to M14). The results show that the Random Forest method with an incremental win-dow approach, using temporal weighting of the instances and the temporally-oriented features of past cultures is better, especially because both the accuracy for resistant bacteria and susceptible bacteria is more balanced.

Regarding previous studies such as [6], [17] and [18], some similarities and differences are observed with this study. There are many differences between [6] and our work, such as the time interval considered in the data set, the number of instances, the generation of new longitudinal features or the methods used, but the *concept drift* is observed in both works. It is even more noticeable in our work due to the long time interval considered, with the windowing approach showing great benefits when applied to this problem. Unlike the work in [17], our study applies temporal modelling with windowing, including data from the 21 days previous to the antibiogram result to be predicted. In this line, authors in [18] also consider the date of culture and apply a temporal modelling, but without windowing.

Remarkable contributions of our study are the new generated sets of features that consider temporal data contained along the data set, which regards the previous resistance of bacteria for the patient under study ($p\&a_i$), and the resistance of bacteria previously detected in the ICU ($r\&a_i$). In line with [18], our work also reveals that data from past cultures contain a relatively high amount of information to predict antimicrobial resistance. Particularly, the $p\&a_i$ set of features showed to be the most useful for correct prediction when used in combination with some other features or even, in the case of some antimicrobial families, when used alone. Another relevant contribution of our study is the *incremental training window* scheme applied together with instance weighting. It allows to accurately classify cultures when the underlying data distribution dramatically changes along time. Our method introduces a more general and robust solution than those previously proposed, since it can be applied to heterogeneous data sets either with just a few or many years to be predicted, which is able to evolve along time and tackle the scarcity problem. Furthermore, it is able to provide high performance results for the majority of families, similar to the ones in other studies despite not using many of the most important risk factors identified in the literature, such as the antibiotics administered to patients. In addition, the thorough analysis of the relevance and interaction of different features will largely help in the development of future works.

There are different challenges to be addressed for future work.

On the one hand, *oversampling* techniques on training can be tested to check their influence on the model performance. On the other hand, we also consider including other features that could have some influence on the appearance of resistance bacteria in the ICU, like some additional patients' details about their admission, whether they required intubation or not and whether they needed mechanical ventilation or not. It would also be interesting to consider the inclusion of features encoding the antibiotic usage in a temporal context, at a patient level and ICU level. In order to properly tackle the different resistant phenotypes observed in this study, the non-uniform distribution of genotypic resistance mechanisms could be considered. It is also relevant to analyze in a different manner (such as assigning particular weights) cultures isolated from some specific sites such as tracheostomy or environmental water sources, because of their ability to generate aerosols close to patients, increasing the probability of nosocomial bacterial transmission.

Acknowledgment

We are thankful to the University Hospital of Fuenlabrada in Madrid, Spain, for providing us the database used in this research. This work has been partly supported by the Spanish Thematic Network "Learning Machines for Singular Problems and Applications (MAPAS)" (TIN2017-90567-REDT, MINECO/FEDER EU), by the IDEAI-UPC Consolidated Research Group Grant from Catalan Agency of University and Research Grants (AGAUR, Generalitat de Catalunya) (2017 SGR 574), by the Science and Innovation Ministry Grants Klinilycs (TEC2016-75361-R), AAVis-BMR (PID2019-107768RA-I00) and BigTheory (PID2019-106623RB-C41), by the Spanish Institute of Health Carlos III (grant DTS 17/00158), by Project Ref. F656 financed by Rey Juan Carlos University, and by the Youth Employment Initiative R&D Project (TIC-11649) financed by the Community of Madrid (Spain). Funded action by the Community of Madrid in the framework of the Multiannual Agreement with the Rey Juan Carlos University in line of action 1, "Encouragement of Young Phd students investigation" Project Ref. F661 Acronym Mapping-UCI.

References

[1] W. H. Organization, et al., "Antimicrobial resistance," *Weekly Epidemiological Record= Relevé épidémiologique hebdomadaire*, vol. 75, no. 41, pp. 336–336, 2000.

[2] Infectious Diseases Society of America (IDSA), "Combating antimicrobial resistance: policy recommendations to save lives," *Clinical Infectious Diseases*, vol. 52, no. suppl_5, pp. S397–S428, 2011.

[3] M. Mendelson, M. P. Matsoso, "The world health organization global action plan for antimicrobial resistance," *SAMJ: South African Medical Journal*, vol. 105, no. 5, pp. 325–325, 2015.

[4] B. K. English, A. H. Gaur, "The use and abuse of antibiotics and the development of antibiotic resistance," in *Hot topics in infection and immunity in children VI*, Springer, 2010, pp. 73–82.

[5] S. Joshi, et al., "Hospital antibiogram: a necessity," *Indian journal of medical microbiology*, vol. 28, no. 4, p. 277, 2010.

[6] A. Tsymbal, M. Pechenizkiy, P. Cunningham, S. Puuronen, "Handling local concept drift with dynamic integration of classifiers: Domain of antibiotic resistance in nosocomial infections," in *19th IEEE Symposium on Computer-Based Medical Systems (CBMS'06)*, 2006, pp. 679–684, IEEE.

[7] A. Lorenz, M. Preuße, S. Bruchmann, V. Pawar, N. Grahl, M. C. Pils, L. M. Nolan, A. Filloux, S. Weiss, S. Häussler, "Importance of flagella in acute and chronic pseudomonas aeruginosa infections," *Environmental microbiology*, vol. 21, no. 3, pp. 883–897, 2019.

[8] G. Meletis, M. Bagkeri, "Pseudomonas aeruginosa: Multi-drug-resistance development and treatment options," *Infection Control*, pp. 33–56, 2013.

[9] M. W. Pesesky, T. Hussain, M. Wallace, S. Patel, S. Andleeb, C.-A. D. Burnham, G. Dantas, "Evaluation of machine learning and rules-based approaches for predicting antimicrobial resistance profiles in gram-negative bacilli from whole genome sequence data," *Frontiers in microbiology*, vol. 7, p. 1887, 2016.

[10] M. Ellington, O. Ekelund, F. M. Aarestrup, R. Canton, M. Doumith, C. Giske, H. Grundman, H. Hasman, M. Holden, K. L. Hopkins, et al., "The role of whole genome sequencing in antimicrobial susceptibility testing of bacteria: report from the eucast subcommittee," *Clinical microbiology and infection*, vol. 23, no. 1, pp. 2–22, 2017.

[11] G. Arango-Argoty, E. Garner, A. Pruden, L. S. Heath, P. Vikesland, L. Zhang, "Deeparg: a deep learning approach for predicting antibiotic resistance genes from metagenomic data," *Microbiome*, vol. 6, no. 1, pp. 1–15, 2018.

[12] M. Nguyen, S. W. Long, P. F. McDermott, R. J. Olsen, R. Olson, R. L. Stevens, G. H. Tyson, S. Zhao, J. J. Davis, "Using machine learning to predict antimicrobial mics and associated genomic features for nontyphoidal salmonella," *Journal of clinical microbiology*, vol. 57, no. 2, 2019.

[13] M. Tlachac, E. A. Rundensteiner, K. Barton, S. Troppy, K. Beaulac, S. Doron, "Predicting future antibiotic susceptibility using regression-based methods on longitudinal massachusetts antibiogram data.," in *HEALTHINF*, 2018, pp. 103–114.

[14] S. Martínez-Agüero, I. Mora-Jiménez, J. Lérida-García, J. Álvarez-Rodríguez, C. Soguero-Ruiz, "Machine learning techniques to identify antimicrobial resistance in the intensive care unit," *Entropy*, vol. 21, no. 6, p. 603, 2019.

[15] À. Hernàndez-Carnerero, M. Sànchez-Marrè, I. Mora-Jiménez, C. Soguero-Ruiz, S. Martínez-Agüero, J. Álvarez Rodríguez, "Modelling temporal relationships in pseudomonas aeruginosa antimicrobial resistance prediction in intensive care unit.," in *Proc. of Singular Problems for Health Care (SP4HC) Workshop at the 24th European Conference on Artificial Intelligence (ECAI 2020)*, 2020.

[16] S. Martínez-Agüero, I. Mora-Jiménez, A. García-Marqués, J. Álvarez Rodríguez, C. Soguero-Ruiz, "Applying lstm networks to predict multi-drug resistance using binary multivariate clinical sequences," in *Proc. of Starting AI Researchers' Symposium (STAIRS) at the 24th European Conference on Artificial Intelligence (ECAI 2020)*, 2020.

[17] G. Eickelberg, L. N. Sanchez-Pinto, Y. Luo, "Predictive modeling of bacterial infections and antibiotic therapy needs in critically ill adults," *Journal of Biomedical Informatics*, vol. 109, p. 103540, 2020.

[18] O. Lewin-Epstein, S. Baruch, L. Hadany, G. Y. Stein, U. Obolski, "Predicting antibiotic resistance in hospitalized patients by applying machine learning to electronic medical records," *Clinical Infectious Diseases*.

[19] Ó. Escudero-Arnanz, I. Mora-Jiménez, S. Martínez-Agüero, J. Álvarez Rodríguez, C. Soguero-Ruiz, "Temporal feature selection for characterizing antimicrobialmultidrug resistance in the intensive care unit," in *Proc. of Singular Problems for Health Care (SP4HC) Workshop at the 24th European Conference on Artificial Intelligence (ECAI 2020)*, 2020.

[20] T. M. Cover, J. A. Thomas, "Elements of information theory," 2012.

[21] A. T. Azar, H. I. Elshazly, A. E. Hassanien, A. M. Elkorany, "A random forest classifier for lymph diseases," *Computer methods and programs in biomedicine*, vol. 113, no. 2, pp. 465–473, 2014.

[22] P. Revuelta-Zamorano, A. Sánchez, J. L. Rojo-Álvarez, J. Álvarez-Rodríguez, J. Ramos-López, C. Soguero-Ruiz, "Prediction of healthcare associated infections in an intensive care unit using machine learning and big data tools," in *XIV Mediterranean Conference on Medical and Biological Engineering and Computing 2016*, 2016, pp. 840–845, Springer.

Àlvar Hernàndez-Carnerero

Àlvar Hernàndez-Carnerero received a B.Sc. degree in Computer Science from the Barcelona School of Informatics (FIB), 2019, and a M.Sc. degree in Intelligent Interactive Systems from the Pompeu Fabra University (UPF) in Barcelona, 2020. He is currently pursuing a Ph.D. in Artificial Intelligence at the Universitat Politècnica de Catalunya (UPC). He is a research support technician at the Computer Science Departament of UPC since 2019. He worked as an Artificial Intelligence developer at LactApp Women Health, in Barcelona, from 2018 to 2019. He also assisted to the 2019 "Apple entrepreneur camp" at the Apple headquarters in Cupertino, California, in 2019. His main research topics are machine learning, application of AI to health care and research data management and case-based reasoning.

Miquel Sànchez-Marrè

Miquel Sànchez-Marrè received a B.Sc. and M.Sc. degrees in Computer Science both from the Barcelona School of Informatics (FIB), in 1988 and 1991, respectively. He received a Ph.D. in Computer Science, in Artificial Intelligence (AI) field, in 1996 from the Universitat Politècnica de Catalunya (UPC). He is Associate Professor in the Computer Science Department (CS) of UPC since 1997 (tenure). He is member of the Intelligent Data Science Research Centre (IDEAI-UPC), and member of the Knowledge Engineering and Machine Learning Group (KEMLG). He is a member of the Editorial Board of International Journal of Applied Intelligence and Modelling journal, and Associate Editor of the Environ-mental Modelling Software journal. He is a Fellow of the International Environmental Modelling and Software Society (iEMSs) since 2005. His main research topics are case-based reasoning, machine learning, intelligent decision support sys-tems, recommender systems, dynamic learning, data science and application of AI to the Environment, Industry and Health.

Inmaculada Mora Jiménez

Inmaculada Mora Jiménez received the degree in Telecommunication Engineering from the Polytechnic University of Valencia, Spain, in 1998, and the Ph.D. degree in Telecommunication from the Carlos III University of Madrid, Spain, in 2004. She is currently an Associate Professor with the Department of Signal Theory and Communications, Telematics and Computing Systems, in Rey Juan Carlos University, Spain. She has conducted her research mainly in data analytic and biomedical engineering. She is co-author of more than 40 JCR indexed papers and 50 contributions to international conferences. She has participated in 18 competitive research projects (principal investigator of 4) and collaborated in more than 20 projects with private funding entities. Her main research interests include data science and machine learning with application to image processing, bioengineering, and wireless communications.

Cristina Soguero Ruiz

Cristina Soguero Ruiz received the Telecommunication Engineering and the Bussiness Degree from Rey Juan Carlos University, Spain, in 2011, and the Ph.D. Degree in machine learning with applications in healthcare, with the Joint Doctoral Program in Multimedia and Communications in conjunction with Rey Juan Carlos University and the University Carlos III of Madrid, in 2015. She was supported by FPU Spanish Research and Teaching Fellowship (granted in 2012). She won the Orange Foundation Best Ph.D. Thesis Award by the Spanish Official College of Telecommunication Engineering. She has published more than 30 papers in JCR journals and international conference communications. She has participated in several research projects (with public and private fundings) related to healthcare data-driven machine learning systems, being principal investigator of 4. Her current research interests include machine learning, data science, and statistical learning theory.

Sergio Martínez Agüero

Sergio Martínez Agüero received the Degree in Telecommunications Technologies Engineering in 2018 at Rey Juan Carlos University and, in 2020 the University Master's Degree in Telecommunication Engineering at Rey Juan Carlos University. He is currently working on his PhD. Thesis, entitled Deep Learning and Network Analytics for extracting knowledge from infectious diseases in the ICU, which extends the research line of the masters and degree thesis. He has several contributions in national and international congresses and published one paper in a JCR journal. He is currently part of two competitive projects funded by the Spanish Government related to healthcare data-driven machine learning models. He is interested in data science, machine learning, data visualization and network analytics.

Joaquín Álvarez-Rodríguez

Joaquín Álvarez-Rodríguez received the Ph.D. in Medicine from the Complutense University of Madrid in 1996. Since 2003 he is the head of the Intensive Care Medicine Department at the Hospital Universitario de Fuenlabrada. His main lines of work have been the quality and safety of patients, medical information systems in Intensive Care Medicine and healthcare-related infections in the ICU. In this last field he has actively participated in the national and regional coordination of the Zero Projects, projects that aim to reduce the main infections acquired in ICU and the emergence of multiresistant bacteria in the ICU. His main research areas are related to the collection of data recorded in the electronic medical record, and try to find a clinical application to them.

Towards Multi-perspective Conformance Checking with Fuzzy Sets

Sicui Zhang[1,2], Laura Genga[2], Hui Yan[3], Hongchao Nie[4], Xudong Lu[1,2]*, Uzay Kaymak[1,2]

[1] Department of Biomedical Engineering and Instrumental Science, Zhejiang University, Hangzhou (P.R. China)
[2] School of Industrial Engineering, Eindhoven University of Technology, Eindhoven (The Netherlands)
[3] Department of Biomedical Engineering, Hainan University (P.R. China)
[4] Philips Research, Eindhoven (The Netherlands)

Received 24 September 2020 | Accepted 23 November 2020 | Published 26 February 2021

Abstract

Nowadays organizations often need to employ data-driven techniques to audit their business processes and ensure they comply with laws and internal/external regulations. Failing in complying with the expected process behavior can indeed pave the way to inefficiencies or, worse, to frauds or abuses. An increasingly popular approach to automatically assess the compliance of the executions of organization processes is represented by alignment-based conformance checking. These techniques are able to compare real process executions with models representing the expected behaviors, providing diagnostics able to pinpoint possible discrepancies. However, the diagnostics generated by state of the art techniques still suffer from some limitations. They perform a crisp evaluation of process compliance, marking process behavior either as compliant or deviant, without taking into account the severity of the identified deviation. This hampers the accuracy of the obtained diagnostics and can lead to misleading results, especially in contexts where there is some tolerance with respect to violations of the process guidelines. In the present work, we discuss the impact and the drawbacks of a crisp deviation assessment approach. Then, we propose a novel conformance checking approach aimed at representing actors' tolerance with respect to process deviations, taking it into account when assessing the severity of the deviations. As a proof of concept, we performed a set of synthetic experiments to assess the approach. The obtained results point out the potential of the usage of a more flexible evaluation of process deviations, and its impact on the quality and the interpretation of the obtained diagnostics.

Keywords

Business Process, Conformance Checking, Data Perspective, Fuzzy Sets.

DOI: 10.9781/ijimai.2021.02.013

I. Introduction

Nowadays organizations often need to employ data-driven techniques to audit their business processes and ensure they should comply to the predefined *process models* with internal/external regulations, e.g., on the execution time or other data perspective constraints. Failing in complying with the expected process behavior can indeed pave the way to inefficiencies or, worse, to frauds or abuses, which often result in loss of money and/or reputation which can have a strong impact on the organization. In recent years, alignment-based *conformance checking* [1] emerged as a widely used approach for organization process auditing. These techniques allow to automatically detect possible discrepancies between real-world process executions and the expected process behavior, usually represented by means of some modelling formalism (e.g., Petri net, or BPMN) [1]-[5].

However, state of the art techniques suffer from some limitations. Processes often involve several alternative execution paths, whose choice can depend on the values of one or more data variables. While this aspect has been traditionally neglected in conformance checking, typically focused on the control flow perspective [1]-[4], recently a few approaches have been proposed to assess process compliance with respect to multiple perspectives [5], [6]. However, existing techniques consider an activity performed at a given point of an execution either *completely deviated* or *completely correct*. Such a crisp distinction is often not suitable in many real-world processes, where decisions on data-guards are often generated with some level of *uncertainty*, which gives rise to some challenges in drawing exact lines between acceptable/not acceptable values. As a result, in these domains there often exists some tolerance to deviations. For example, let us assume that in a medical process there is a guideline stating that in between two procedures there must be an interval of at most five hours. Adopting a crisp evaluation, 4 hours 59 minutes would be considered fully compliant, while 5 hours and 1 minute would be fully not compliant, which is intuitively unreasonable. Such an approach can lead to generating misleading diagnostics, where executions marked as deviating actually correspond to acceptable behaviors. Furthermore, the magnitude of the deviations is not considered; small or large violations are considered at the same level of compliance, which can easily be misleading to the diagnosis. It is worth noting that this approach can also hamper the overall process resilience, making it very sensible even to small exceptions/disruptions. For instance, if process executions are monitored in a real-time way, every small deviations can lead to raise some alarms and/or to stop the execution.

* Corresponding author.
E-mail address: lvxd@zju.edu.cn

To deal with these challenges, in this work we perform an exploratory study on the use of *fuzzy sets* [7] in conformance checking. Fuzzy sets have been proven to be a valuable asset to represent human decisions making process, since they allow to formalize the uncertainty often related to these processes. In particular, elaborating upon fuzzy theory, we propose a new multi-perspective conformance checking technique that accounts for the degree of deviations. Taking into account the severity of the occurred deviations allows a) improving the quality of the provided diagnostics, generating a more accurate assessment of the deviations, and b) enhancing the flexibility of compliance checking mechanisms, thus paving the way to improve the overall *resilience* of the process management system with respect to unforeseen exceptions [8]. As a proof-of-concept, we tested the approach over a synthetic dataset.

The rest of this work is organized as follows. Section II discusses related work. Section III introduces a running example to discuss the motivation of this work. Section IV introduces basic formal notions. Section V illustrates the approach. Section VI discusses results obtained by a set of synthetic experiments. Finally, Section VII draws some conclusions and future work.

II. Related Work

Conformance checking discipline has evolved significantly in recent times. One of the first automatic approaches was introduced by [9], which proposed a token-based approach to detect deviations by replaying each event of a process execution against a process model, to determine whether the execution was or not allowed by the model. While this seminal work provides detailed diagnostics, supporting the detection of inserted and skipped activities, and it is able to deal with possible infinite behavior (e.g., in the case of loops), further research proved that token-based techniques can lead to misleading diagnostics [10]. Recently, alignments have been proved to be a robust way to check the conformance of the given logs [2]. Alignment-based techniques are able not only to pinpoint occurred deviations, but also to determine the most probable explanation of non conformity. To this end, a cost function is used to determined the cost of alternative explanations, then returning the one with minimum cost. Although most alignment-based approaches apply the standard distance cost function defined by [2], several variants have been suggested to enhance the quality of the compliance assessment. For instance, Alizadeh et al. [11] proposes a method to obtain the probable explanations for nonconformity by computing the cost function from historical logging data. While traditional conformance checking techniques are solely focused on assessing compliance with respect to the control-flow, i.e., the ordering of the activities, recently few approaches in literature investigated how to include other perspectives, e.g., resources, time, data, and so on in conformance checking algorithms. The approach introduced in [6] suggests to align the control-flow first, and then check the executions compliance with respect to the data perspective. While this approach does allow to detect data-related deviations, it still gives more importance to the control flow perspective when it comes to the deviation interpretation, with the results that he can miss some critical deviations in the alignment [5]. With a different interpretation, the work of [12] considers the data perspective prior to control flow, thus aligning the data variables to the data-aware decision paths first for a reference trace, and next replaying it to the execution trace for the mismatches on control flow conformance. The research in [5], instead, aims at balancing the impact of all the different process perspectives when generating the alignment, considering all perspectives equally important. To this end, they propose a cost function which takes into account both data and control flow deviations simultaneously.

The techniques mentioned above adopt a crisp evaluation of the conformance, where a behaviour is completely wrong or completely correct. In this work, we propose to use fuzzy sets theory to assess the magnitude of the detected deviations. Several researches in literature have explored the employment of fuzzy sets in representing expert decision making processes; among them, we can mention, for example, [13], which studies a fuzzy approach to model farmers' decision process in a integrated farming systems; [14], which represents vagueness in linguistic judgements by means of a fuzzy analytic hierarchy process; [15], which applies a fuzzy dynamic method for risk decision making problems for a mine; and the work of [16], which proposes a fuzzy linguistic method for Multiple Criteria Decision Making (MCDM) problem to Prioritize the elective surgery admission in a local public hospital. However, only a few approaches also explored the use of fuzzy theory in process analysis. [17] proposes to characterize the conformance problem by means of an existing fuzzy rule-based framework ; the study of [18] uses a fuzzy process miner on a clinical data-set to support hospital administrators in improving the performance of their processes (e.g., reducing patients' waiting times). However, to the best of our knowledge, no previous work has exploited fuzzy sets theory in the cost function of conformance checking techniques.

III. Motivating Example

Consider, as a running example, a loan management process derived from previous work on the event log of a financial institute made available for the BPI2012 challenge [19], [20]. Fig. 1 shows the process in BPMN notation. The process starts with the submission of an application. Then, the application passes through a first assessment, aimed to verify whether the applicant meets the requirements. If the requested amount is greater than 10000 euros, the application also goes through a more accurate analysis to detect possible frauds. If the application is not eligible, the process ends; otherwise, the application is accepted. An offer to be sent to the customer is selected and the details of the application are finalized. After the offer has been created and sent to the customer, the latter is contacted to discuss the offer with him/her, possibly adjusting according to her preferences. At the end of the negotiation, the agreed application is registered on the system. At this point, further checks can be performed on the application, if the overall duration is still below 30 days, before approving it.

Let us assume that this process is supported by some systems able to track the execution of its activities in a so-called event log. In practice, this is a collection of *traces*, i.e., sequences of activities performed within the same process execution, each storing information like the execution timestamp of the execution, or other data element [1]. Let the following be two example traces extracted by the system supporting the process at hand (note that we use acronyms of the activities names, for the sake of simplicity)[1]:

$\sigma_1 = \langle (A\_S, \{Amount = 9950\}), (W\_FIRST\_A, \perp), (W\_F\_C, \perp), (A\_A, \perp), (A\_F, \perp), (O\_S, \perp), (O\_C, \perp), (O\_S, \perp), (W\_C, \perp), (A\_R, \{Duration = 50\}), (A\_AP, \perp) \rangle$;

$\sigma_2 = \langle (A\_S, \{Amount = 2000\}), (W\_FIRST\_A, \perp), (W\_F\_C, \perp), (A\_A, \perp), (A\_F, \perp), (O\_S, \perp), (O\_C, \perp), (O\_S, \perp), (W\_C, \perp), (A\_R, \{Duration = 60\}), (A\_AP, \perp) \rangle$;

Both these executions violate the guard on the *Amount* value; indeed, the activity *W_F_C* should have been skipped, being the requested loan amount lower than 10000. It is worth noting, however, that there is

[1] We use the notation $(act, \{att_1 = v_1, \ldots, att_n = v_n\})$ to denote the occurrence of activity *act* in which variables $att_1 \ldots att_n$ are assigned to corresponding values $v_1, \ldots v_n$. The symbol $\perp$ means that no variable values are changed when executing the activity.

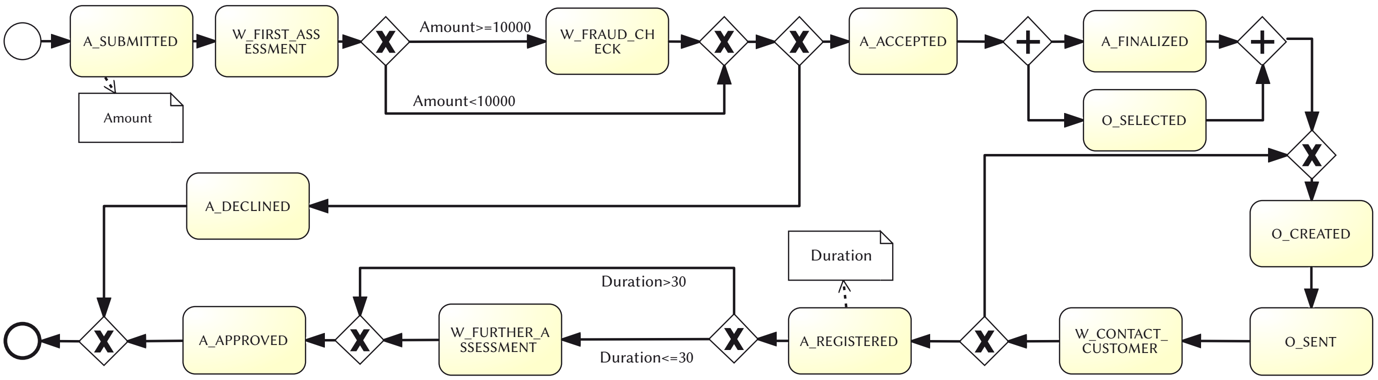


Fig. 1. The Load Management Model.

a significant difference in terms of their *magnitude*. Indeed, while in the first execution the threshold was not reached only by few dozens of euros, the second violation is several thousands of euros below the limit. It is worth noting that applying state-of-the art conformance checking techniques, this difference between σ_1 and σ_2 would remain undetected. Indeed, these techniques adopt a *crisp* logic, where the value of a data variable can be marked only either as correct or wrong.

We argue that taking into account the severity of the violations when assessing execution compliance allows to obtain more accurate diagnostics, especially in contexts where there exists some uncertainty related to the guards definition. Indeed, in these cases guards often represent more guidelines, rather than strict, sharp rules, and there might be some tolerance with respect to violations. In our example, σ_1 could model an execution considered suspicious for some reasons, making a a fraud check worthy, since the amount is only slightly less than 10000. On the other hand, the violation in σ_2 deserves some attention, since the amount is so far from the threshold that the additional costs needed for the fraud check are probably not justified.

Differentiating among different levels of violations also impacts the analysis of possible causes of the deviations. Indeed, conformance checking techniques also attempt to support the user in investigating the *interpretations* of a deviation. In our example, the occurrence of the activity *W_F_C* could be considered either as a control-flow deviation or as a data-flow deviation. In absence of domain knowledge in determining what is the real explanation, conformance checking techniques assess the severity (aka, cost) of the possible interpretations and select the least severe one, assuming that this is the one closest to the reality. In our example, conformance checking would consider both the interpretation equivalent for both the traces; instead, differentiating between the severity of the deviations would make the second interpretation the preferred one when the deviation is limited, like in σ_1, thus providing more guidance to the analyst during process diagnostics.

IV. Preliminaries

This section introduces a set of definitions and concepts that will be used through the paper. First, we recall important *conformance checking* notions; secondly, we introduce basic elements of *fuzzy theory*.

A. Conformance Checking: Aligning Event Logs and Models

Conformance checking techniques detect discrepancies between a process model describing the expected process behavior and the real process execution.

The expected process behavior is typically represented as a *process model*. Since the present work is not constrained to the use of a specific modeling notation, here we refer to the notation used in [2], enriched with data-related notions explained in [6].

Definition 1 (Process model). A process model $M = (P, P_I, P_F, A_M, V, U, T, G, W, Values)$ is a transition system defined over a set of activities A_M and a set of variables V, with states P, initial states $P_I \subseteq P$, final states $P_F \subseteq P$ and transitions $T \subseteq P \times (A_M \times 2^V) \times P$. The function U defines the admissible data values, i.e., $U(V_i)$ represents the domain of V_i for each variable $V_i \in V_i$; the function G: $A_M \rightarrow Formulas(V \cup \{V_i{'} \mid V_i \in V\})$ is a *guard* function, that associates an activity to a criterion, i.e., a boolean formula expressing a condition on the values of the data variables; W: $A_M \longrightarrow 2^V$ is a write function, that associates an activity with the set of variables which are written/updated by the activity; finally, $Values$: $P \rightarrow \{V_i = v_i, \text{i} = 1 \,..\, |V| \mid v_i \in U\,(V_i) \cup \{\perp\}\}$ is a function that associates each state with the corresponding pairs variable=value.

When a variable $V_i \in V$ appears in a guard $G\,(A_M)$, it refers to the value just before the occurrence of A_M; however, if $V_i \in W(A_M)$, it can also appear as $V_i{'}$, and refers to the value after the occurrence. The firing of an activity $s = (a, w) \in A_M \times (V \nrightarrow U)$ in a state p' is *valid* if: 1) a is enabled in p'; 2) a writes all and only the variables in $W(a)$; 3) $G(a)$ is *true* when evaluated over $Values\,(p')$. To access the components of s we introduce the following notation: $vars(s) = w$, $act(s) = a$. Function *vars* is also overloaded such that $vars(s, V_i) = w(V_i)$ if $V_i \in dom(vars(s))$ and $vars(s, V_i) = \perp$ if $V_i \notin dom(vars(s))$. The set of valid process traces of a process model M is denoted with $\rho(M)$ and consists of all the valid firing sequences $\sigma \in (A_M \times (V \nrightarrow U)^*$ that, from an initial state P_i lead to a final state P_F.

Process executions are often recorded by means of an information system in so-called event logs. In particular, an event log consists of traces, each collecting the sequence of events recorded during the same process execution. Formally, let S_N be the set of (valid and invalid) firing of activities of a process model M; an **event log** is a multiset of traces $\mathbb{L} \in \mathbb{B}(S_N^*)$. Given an event log L, conformance checking builds an *alignment* between L and M, whose goal consists in relating activities occurred in the event log to the activities in the model and vice versa. To this end, we need to map moves“ occurring in the event log to possible moves” in the model. However, since the executions may deviate from the model and/or not all activities may have been modeled or recorded [2], we might have log/model moves which cannot be mimicked by model/log moves respectively. These situations are modeled by a “no move” symbol “≫”. For convenience, we introduce the set $S_N^{\gg} = S_N \cup \{\gg\}$. Formally, we set S_L to be a transition of the events in the log, S_M to be a transition of the activities in the model. A move is represented by a pair $(s_L, s_M) \in S_N^{\gg} \times S_N^{\gg}$ such that:

- (s_L, s_M) is a *move in log* if $s_L \in S_N$ and $s_M = \gg$
- (s_L, s_M) is a *move in model* if $s_M \in S_N$ and $s_L = \gg$
- (s_L, s_M) is a *move in both without incorrect write operations* if $s_L \in S_N$, $s_M \in S_N$ and $act(s_L) = \text{act}(s_M)$ and $\forall\, V_i \in V(vars(s_L, V_i) = \text{vars}(s_M, V_i))$
- (s_L, s_M) is a *move in both with incorrect write operations* if $s_L \in S_N$, s_M

$\in S_N$ and $act(s_L) = act(s_M)$ and $\exists\, V_i \in V \mid vars(s_L, V_i) \neq vars(s_M, V_i))$

Let $A_{LM} = \{(s_L, s_M) \in S_N^{\gg} \times S_N^{\gg} \mid s_L \in S_N \vee s_M \in S_N\}$ be the set of all legal moves. The *alignment* between two process executions $\sigma_L, \sigma_M \in S_N^*$ is $\gamma \in A_{LM}^*$ such that the projection of the first element (ignoring $\gg$) yields σ_L, and the projection on the second element (ignoring $\gg$) yields σ_M.

Given log trace and process model, multiple alternative alignments exist. Our goal is to find the *optimal alignment*, i.e., a complete alignment as close as possible to a proper execution of the model. To this end, the severity of deviations is assessed by means of a *cost function*:

Definition 2 (Cost function, Optimal Alignment). Let σ_L, σ_M be a log trace and a model trace, respectively. Given the set of all legal moves A_N, a *cost function* k assigns a non-negative cost to each legal move: $A_N \rightarrow \mathbb{R}_0^+$. The *cost of an alignment* γ between σ_L and σ_M is computed as the sum of the cost of all the related moves: $K(\gamma) = \sum_{(S_L,S_M)\in\gamma} k\,(S_L, S_M)$. An **optimal** alignment of a log trace and a process trace is one of the alignments with the lowest cost according to the provided cost function.

B. Basic Fuzzy Sets Concepts

Classic sets theory defines crisp, dichotomous functions to determine membership of an object to a given set. For instance, a set N of real numbers smaller than 5 can be expressed as $N = \{n \in \mathbb{R} \mid n < 5\}$. In this setting, an object either belongs to N or it does not. Although crisp sets have proven to be useful in various applications, there are some drawbacks in their use. In particular, human thoughts and decisions are often characterized by some degree of uncertainty and flexibility, which are hard to represent in a crisp setting [21].

Fuzzy sets theory aims at providing a meaningful representation of measurement uncertainties, together with a meaningful representation of vague concepts expressed in natural language and close to human thinking [22]. Formally, a *fuzzy set* is defined as follows:

Definition 3 (Fuzzy Set). Let N be a collection of objects. A *fuzzy set F* over N is defined as a set of ordered pairs $F = \{n, \mu_F(n) \mid n \in N\}$. $\mu_F(n)$ is called the membership function (μ) for the fuzzy set F, and it is defined as μ_F: $N \rightarrow [0,1]$. The set of all points n in N such that $\mu_F(n) > 0$ is called the **support** of the fuzzy set, while the set of all points in N in which $\mu_F(n) = 1$ is called **core**.

It is straightforward to see that fuzzy sets are extensions of classical sets, with the characteristic function allowing to any value between 0 and 1. In literature several standard functions have been defined for practical applications (see, e.g., [22] for an overview of commonly used functions).

V. Methodology

The goal of this work is introducing a compliance checking approach tailored to take into account the severity of the deviations, in order to introduce some degree of flexibility when assessing compliance of process executions and to generate diagnostics more accurate and possible closer to human interpretation. To this end, we investigate the use of *fuzzy theory*. In particular, we propose to use fuzzy membership functions to model the cost of moves involving data; then, we employ off-shelf techniques based on the use of A* algorithm to build the optimal alignment. The approach is detailed in the following subsections.

A. Fuzzy Cost Function

The computation of an optimal alignment relies on the definition of a proper cost function for the possible kind of moves (see Section [sec:preliminaries]). Most of state-of-the art approaches adopt (variants of) the standard distance function defined in [2], which sets a cost of 1 for every move on log/model (excluding invisible transitions), and a cost of 0 for synchronous moves. Furthermore, the analyst can use *weights* to differentiate between different kind of moves.

The standard distance function is defined only accounting for the control-flow perspective. However, in this work we are interested in the data-perspective as well. In this regards, a cost function explicitly accounting for the data perspective has been introduced by [5] and it is defined as follows.

Definition 4 (Data-aware cost function). Let (S_L, S_M) be a move between a log trace and a model execution, and let, with a slight abuse of notation, $W(S_M)$ to represent write operations related to the activity related to S_M. The cost $k(S_L, S_M)$ is defined as:

$$k(S_L,S_M) = \begin{cases} 1 & \text{if it is a move in log} \\ 1 + |W(S_M)| & \text{if it is a move in model} \\ |\{V_i \in W(S_M): & \text{if it is a move in both} \\ var(S_L,V_i) & \\ \neq var(S_M,V_i)\}| & \end{cases} \qquad (1)$$

In this definition, data costs are computed as a) number of missing data variables because the corresponding activity was skipped, i.e., for a move in model, b) number of data variables in a synchronous move whose values are not allowed according to the process model, i.e., for a move in both.

Compared to Definition 4, in this paper we integrate both data violation situations a) and b), by considering the missing variables as a noncompliance to the rule as well, thereby counting the data cost with a move in both. Besides, the cost function in (1) uses a dichotomous function which considers every move either as *completely wrong* or *completely correct*. To differentiate between different magnitude of deviations, in this work we propose to use fuzzy membership functions as cost functions for the alignment moves. Note that here we focus on data moves. Indeed, when considering other perspectives the meaning of the severity of the deviation is not that straightforward. For example, when considering control-flow deviations, usually an activity is either executed or skipped. Nevertheless, fuzzy costs can be defined also for other process perspectives, for instance, to differentiate between skip of activities under different conditions. We plan to explore these directions in future work.

Following the above discussion, we define our *fuzzy cost function* as follows:

Definition 5 (Data-aware fuzzy cost function). Let (S_L, S_M) be a move between a process trace and a model execution, and let $\mu(var(S_L, V_i))$ be a fuzzy membership function returning the degree of deviation of a data variable in a move in both with incorrect data. The cost $k(S_L, S_M)$ is defined as:

$$k(S_L,S_M) = \begin{cases} 1 & \text{if a move in log} \\ 1 & \text{if a move in model} \\ \sum_{\forall V_i \in V} \mu\,(var(S_L,V_i)) & \text{if a move in both} \end{cases} \qquad (2)$$

To define the fuzzy cost function in (2), we first need to determine over which data constraints we want to define a μ [2]. Then, for each of them first we need to define a tolerance interval; in turn, this implies to define a) an interval for the core of the function, and b) an interval for the support of the function (see Section IV). This choice corresponds to determine, for a given data constraint, which values should be considered equivalent and which ones not optimal but still acceptable. Once the interval is chosen, we need to select a suitable membership function. In literature, several different μ have been defined (see, e.g., [22] for an overview), with different level of complexity and different

[2] Note that multiple μ functions can be defined for the same data variable, if it is used in multiple guards.

interpretations. It is straightforward to see that determining the best μ to explicit the experts' knowledge is not a trivial task. For the sake of space, an extended discussion over the μ modeling is out of the scope of this paper, and left for future work. Nevertheless, we would like to point out that this is a well-studied issue in literature, for which guidelines and methodologies have been drawn like, e.g., the one presented by [23]. The approach can be used in combination of any of these methodologies, since it does not depend on the specific μ chosen.

It is worth noting that on one hand, the cost function (2) can be seen as a direct extension of (1) to the fuzzy case, where the cardinality of a set of differences has been replaced by the cardinality of a fuzzy set (denoting the compliance to a soft constraint). On the other hand, there is also some reasoning behind this formulation of the fuzzy cost function from an aggregation of information perspective. There are various problems in which the deviation from a control-flow perspective is comparable to a deviation in the data perspective in terms of the consequences of the deviation. In this case, an additive cost function makes sense in which the cost incurred from a gradual violation in the data perspective is comparable (or is the same) as the cost incurred from a violation of an activity in the control- flow perspective. Additionally, the cost function in (2) is essentially a penalty function in which different costs are aggregated in additive fashion, implying that a small compliance along one data dimension can be compensated by a large compliance along another data dimension. There is a large class of problems in which such an additive cost function makes sense [24], since good properties in one variable (criterion) can be compensate the poor qualities along another variable (criterion).

In general, it is possible to consider different, more advanced and/or more complex aggregation of the information regarding the violations. Fuzzy set theory provides a rich set of aggregation functions, pre-aggregation functions, and other mathematical formalisms for aggregating the cost information regarding violations [25]. A thorough analysis beyond the additive function is not within the scope of this preliminary paper. However, an initial investigation of using more complex fuzzy set aggregations can be found in [26].

B. Alignment Building: Using A* to Find the Optimal Alignment

The problem of finding an optimal alignment is usually formulated as a search problem in a directed graph [27]. Let $Z = (Z_V, Z_E)$ be a directed graph with edges weighted according to some cost structure. The A* algorithm finds the path with the lowest cost from a given source node $v_0 \in Z_v$ to a node of a given goals set $Z_G \subseteq Z_V$. The cost for each node is determined by an evaluation function $f(v) = g(v) + h(v)$, where:

- g: $Z_V \rightarrow \mathbb{R}^+$ gives the smallest path cost from v_0 to v;
- h: $Z_V \rightarrow \mathbb{R}_0^+$ gives an estimate of the smallest path cost from v to any of the target nodes.

If h is *admissible*,i.e. underestimates the real distance of a path to any target node v_g, A* finds a path that is guaranteed to have the overall lowest cost.

The algorithm works iteratively: at each step, the node v with lowest cost is taken from a priority queue. If v belongs to the target set, the algorithm ends returning node v. Otherwise, v is expanded: every successor v_0 is added to priority queue with a cost $f(v_0)$.

Given a log trace and a process model, to employ A* to determine an optimal alignment we associate every node of the search space with a prefix of some complete alignments. The source node is an empty alignment $\gamma_0 = \langle\rangle$, while the set of target nodes includes every complete alignment of σ_L and M. For every pair of nodes (γ_1, γ_2), γ_2 is obtained by adding one move to γ_1.

The cost associated with a path leading to a graph node γ is then defined as $g(\gamma) = K(\gamma) + \epsilon\,|\gamma|$, where $K(\gamma) = \sum_{(s_L,s_M)\in\gamma} k\,(s_L, s_M)$, with $k\,(s_L, s_M)$ defined as in (2); $|\gamma|$ is the number of moves in the alignment; and ϵ is a negligible cost, added to guarantee termination when implementing the A* algorithm (see [5] for a formal proof). Note that the cost g has to be strictly increasing. While a formal proof is not possible for the sake of space, it is however straight to see that g is obtained in our approach by the sum of all non negative elements; therefore, while moving from an alignment prefix to a longer one, the cost can never decrease. For the definition of the heuristic cost function $h(v)$ different strategies can be adopted. Informally, the idea is computing, from a given alignment, the minimum number of moves (i.e., the minimum cost) that would lead to a complete alignment. Different strategies have been defined in literature, e.g., the one in [2], which exploits Petri-net marking equations, or the one in [28], which generates possible states space of a BPMN model.

VI. Implementation and Experiments

This section describes a set of experiments we performed to obtain a proof-of-concept of the approach. To this end, we compared the diagnostics returned by a crisp conformance checking approach with the outcome obtained by our proposal. In order to get meaningful insights on the behavior we can reasonably expect by applying the approach in the real world, we employ a realistic synthetic event log, introduced in a former paper [29], obtained starting from one real-life logs, i.e., the event log of the BPI2012 challenge[3]. We evaluated the compliance of this log against a simplified version of the process model in , to which we added few data constraints (see Fig. 1). The approach has been implemented as an extension to the tool developed by [28], designed to deal with BPMN models. In the following we describe the experimental setup and the obtained results.

A. Settings

The log in [29] consists of 5000 traces, where a predefined set of deviations was injected. The values for the variable "*Amount*" were collected the from the BPI2012 log, while for calculating "*Duration*" a random time window ranging from 4 to 100 hours has been put in between each pair of subsequent activities, and the overall duration was then increased of by 31 days for some traces. For more details on the log construction, please check [29].

Our process model involves two constraints for the data perspective, i.e., $Amount >= 10000$ to execute the activity *W_F_C*, and $Duration <= 30$ to execute the activity *W_FURTHER_A*. For the crisp conformance checking approach, we use the cost function provided by (1); while for the fuzzy approach, the cost function in (2). Here we assume that $Amount \in (3050{,}10000)$ and $Duration \in (30{,}70)$ represent a tolerable violation range for the variables. Since we do not have experts' knowledge available for these experiments, we derived these values from simple descriptive statistics. In particular, we draw the distributions of the values for each variable, considering values falling within the third quartile as acceptable. The underlying logic is that values which tend to occur repeatedly are likely to indicate acceptable situations. Regarding the shape of the membership function, here we apply a special trapezoidal function, reported below. *Amount* and *Duration* are abbreviated to A and D.

$$\mu_1(A) = \begin{cases} 0 & \text{, if } A \geq 10000 \\ 1 & \text{, if } A \leq 3050 \\ \dfrac{10000 - A}{6950} & \text{, if } 3050 < A < 10000; \end{cases}$$

3 https://www.win.tue.nl/bpi/doku.php?id=2012:challenge

$$\mu_2(D) = \begin{cases} 0 & \text{, if } D \leq 30 \\ 1 & \text{, if } D \geq 70 \\ \dfrac{D-30}{40} & \text{, if } 30 < D < 70 \end{cases}$$

B. Results

We compare the diagnostics obtained by the crisp approach and by our approach in terms of a) kind of moves regarding the activities ruled by the guard, and b) distribution of fitness values, computed according to the definition in [6]. Table I shows differences in terms of number and kind of moves detected for the activities *W_F_C* and *W_FURTHER_A* within the crisp/fuzzy alignments respectively, considering also the possible existence of multiple optimal alignments. Namely, when the same move got different interpretations in different alignments, we count the move as both move in log and move in data. Note, however, that the multiple optimal alignments with the same interpretation for the move count one. It is worth noting that while we obtained the same result for both the move-in-log and move-in-data amount for the crisp approach, these values change considerably when considering the fuzzy approach, which returned a significantly smaller amount of move-in-log. The reason for this difference becomes clear by analyzing the boxplots in Fig. 2, which shows the distributions of data deviation severity. We can see that the ranges are similar for both the constraints, with most of the values remaining below 0.65. These distributions suggest that data deviations are mostly within the tolerance range in our dataset; as a consequence, we expect that in most of the cases the move-in-data will have a smaller cost than the move-in-log and will hence be preferred when building the optimal alignment, which justifies the numbers reported in Table I.

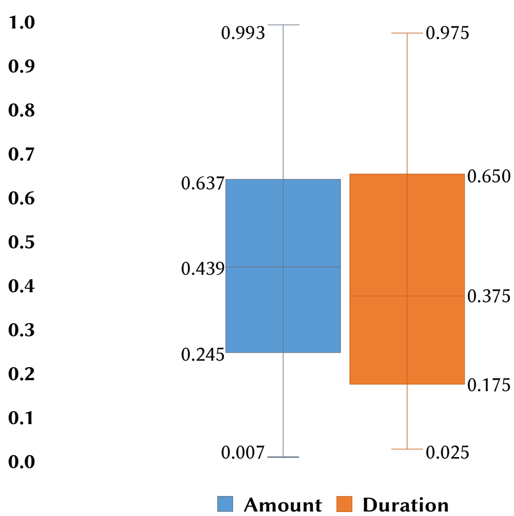


Fig. 2. Boxplots of data deviation.

TABLE I. Number of Different Moves Kinds for Activities *W_F_C* and *W_FURTHER_A*

	W_F_C		*W_FURTHER_A*	
	move-in-log	move-in-data	move-in-log	move-in-data
Crisp	744	744	958	958
Fuzzy	177	744	245	958

From these observations, it follows that we also expect relevant differences in fitness values computed by the fuzzy and the crisp approaches. In particular, we expect to obtain higher values of fitness with the fuzzy approach, being the fuzzy costs less severe than the crisp ones. Fig. 3 shows a scatter plot in which each point represents one trace. The x-axis represents the fitness level of alignment with crisp costs, while the y-axis represents the value corresponding to the fuzzy cost. For the traces on the main diagonal, the fitness level remains unchanged between the two approaches; while for traces that are above the main diagonal, the fuzzy approach obtained higher values of fitness. From the graph we can see that the fuzzy approach never returned lower values of fitness than the crisp one; instead, it returned (also significantly) improved level of fitness for a relevant percentage of the examined cases. Delving into this observation, we found out that the fuzzy approach returns higher value of fitness for 24.3% of the traces.

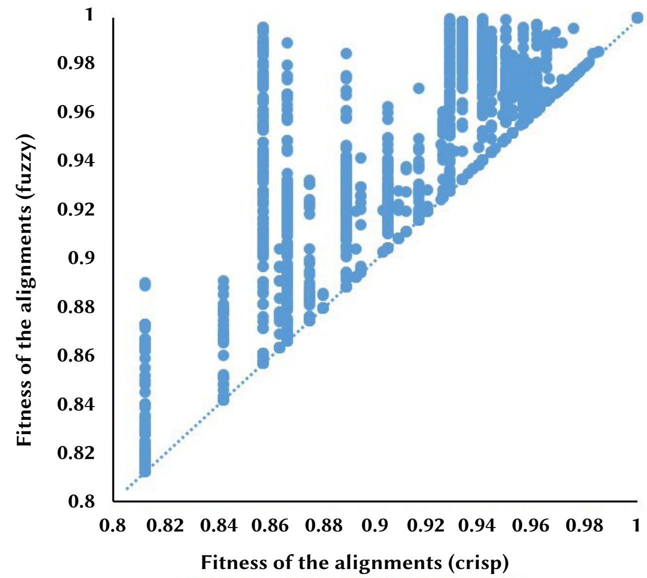


Fig. 3. Comparison of the fitness values obtained with crisp and fuzzy cost.

It is worth noting that, since alignments aim at supporting also the interpretation of the detected deviations, as discussed in Section [sec:motivation], different cost functions also impact the interpretation of the output from an human analyst. To better clarify this aspect, in the following example, we discuss the alignments obtained on one of the traces of our dataset in which the fuzzy and the crisp approach returned different outputs.

Example 1. Let us consider $\sigma = \langle(A\_S, \{Amount = 8160\}), (W\_FIRST\_A, \perp), (W\_F\_C, \perp), (A\_D, \perp), (A\_A, \perp), (A\_F, \perp), (O\_S, \perp), (O\_C, \perp), (O\_S, \perp), (W\_C, \perp), (O\_C, \perp), (O\_S, \perp), (W\_C, \perp), (O\_C, \perp), (O\_S, \perp), (W\_C, \perp), (A\_R, \{Duration{=}97\}), (W\_FURTHER\_A, \perp), (A\_AP, \perp), \rangle$. Table II and Table III show the alignment obtained adopting the crisp cost function the fuzzy cost function, respectively. For the sake of space, here we report only the lines of the alignments related to the activities ruled by the data guards. For each move, we report the position of the move in the alignment followed by "#". We can observe that for the second deviation multiple alternative interpretations were returned by both the approaches, either as move-in-log or a move-in-data; indeed, the data deviation is outside the tolerance range, with the result that the costs are equal to 1 both for the move-in-log and for the move-in-data. Instead, the first deviations is always considered as a move-in-data in the fuzzy approach, since the deviation is within the tolerance range and, hence, the cost is less than 1. We argue that this interpretation is reasonably closer to the human's interpretation than the crisp one. Indeed, we can expect that a human analyst would consider the execution of *W_F_A* as correct in this trace, being the data violation negligible. Furthermore, the fuzzy approach returned a higher fitness value for the trace than the crisp one; this is reasonable, since the first deviation is still close enough to the ideal value.

TABLE II. The Optimal Alignments Returned by the Crisp Cost Function

No.	Model	Log	$\delta cost$
...	...	...	...
3#	≫	$W\_F\_C\,(Amount = 8160)$	1
...	...	...	...
18#	≫	$W\_F\_A\,(Duration = 97)$	1

TABLE III. The Optimal Alignment Returned By a Fuzzy Cost Function

No.	Model	Log	$\delta cost$
...	...	...	...
3#	W_F_C	$W\_F\_C\,(Amount = 8160)$	0.265
...	...	...	...
18#	≫	$W\_F\_A\,(Duration = 97)$	1

Summing-up, the performed comparison did highlight how the use of a fuzzy cost led to improved diagnostics. On the overall fitness level, the fuzzy cost function has obtained higher level of fitness, which represents a more accurate diagnostics [9]. It proves that the fuzzy approach provides a more precise evaluation of the deviation level, taking into account actors' acceptance. In particular, the results show that the fuzzy approach allows to obtain a more fine-grained evaluation of traces compliance levels, allowing the analyst to differentiate between reasonably small and potentially critical deviations. Furthermore, they pointed out the impact that the cost function has on the interpretation of the alignments. Indeed, the approach allows to establish a preferred interpretation in cases in which the crisp function would consider possible options as equivalent, thus reducing ambiguities in interpretation, and providing interpretations for the detected deviations reasonably closer to human analysts' ones.

VII. Conclusion

The present work investigated the use of fuzzy sets concepts in multi-perspective conformance checking. In particular, we showed how fuzzy set notions can be used to take into account the severity of deviations when building the optimal alignment. We implemented the approach and performed a proof-of-concept over a synthetic dataset, comparing results obtained adopting a standard crisp logic and our fuzzy logic. The obtained results confirmed the capability of the approach of generating more accurate diagnostics, as shown both by a) the difference in terms of fitness of the overall set of executions, due to a more fine-grained evaluation of the magnitude of the occurred deviations, and b) by the differences obtained in terms of the different preferred explanations provided by the alignments of the different approaches.

Our results indicate that by exploiting the flexibility in the definition of gradual concepts, conformance analysis from the data perspective is improved. By using fuzzy sets to represent gradual constraints, the penalization of slight violations of the constraints is also made gradual, which reduces the cost associated with a slight violation, and this seems to improve the results of matching between a process model and the event log. Effectively, the fuzzy sets are used to represent a weighting of the violation of business (clinical) rules, which renders the conformance analysis less sensitive to small violations of such rules.

Since this is an exploratory work, there are several research directions that can still be explored. First, in future work we plan to test our approach in real-world experiments, to generalize the results obtained so far. When dealing with real-world experiments, we expect handling of missing values to be an important step in our analysis. There are various methods in which this could be done, such as imputation methods or approaches based on possibility theory in order to deal with the unknown nature of the missing data. Another research direction we intend to explore consists of introducing interval valued fuzzy sets or type-2 fuzzy sets for dealing with the variability that might occur when obtaining the fuzzy sets in our cost function from experts. Inter-expert variability can best be handled with more generic forms of fuzzy sets, which will allow us to extend the flexibility of the analysis process to the process analysts' needs.

Finally, in future work we intend to investigate how to exploit our flexible conformance checking approach to enhance the system on-line resilience to exceptions and unforeseen events.

Acknowledgment

The research leading to these results has received funding from the Brain Bridge Project sponsored by Philips Research.

References

[1] W. Van der Aalst, A. Adriansyah, A. K. A. De Medeiros, F. Arcieri, T. Baier, T. Blickle, J. C. Bose, P. Van Den Brand, R. Brandtjen, J. Buijs, et al., "Process mining manifesto," in International Conference on Business Process Management, 2011, pp. 169–194, Springer.

[2] W. Van der Aalst, A. Adriansyah, B. van Dongen, "Replaying history on process models for conformance checking and performance analysis," Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, vol. 2, no. 2, pp. 182–192, 2012.

[3] A. Adriansyah, B. F. van Dongen, W. M. van der Aalst, "Memory-efficient alignment of observed and modeled behavior," BPM Center Report, vol. 3, 2013.

[4] A. Adriansyah, J. Munoz-Gama, J. Carmona, B. F. van Dongen, W. M. van der Aalst, "Alignment based precision checking," in International Conference on Business Process Management, 2012, pp. 137–149, Springer.

[5] F. Mannhardt, M. De Leoni, H. A. Reijers, W. M. Van der Aalst, "Balanced multi-perspective checking of process conformance," Computing, vol. 98, no. 4, pp. 407–437, 2016.

[6] M. De Leoni, W. M. Van Der Aalst, "Aligning event logs and process models for multi-perspective conformance checking: An approach based on integer linear programming," in Business Process Management, Springer, 2013, pp. 113–129.

[7] S.-C. Cheng, J. N. Mordeson, "Fuzzy linear operators and fuzzy normed linear spaces," in First International Conference on Fuzzy Theory and Technology Proceedings, Abstracts and Summaries, 1992, pp. 193–197.

[8] G. Müller, T. G. Koslowski, R. Accorsi, "Resilience-a new research field in business information systems?," in International Conference on Business Information Systems, 2013, pp. 3–14, Springer.

[9] A. Rozinat, W. M. Van der Aalst, "Conformance checking of processes based on monitoring real behavior," Information Systems, vol. 33, no. 1, pp. 64–95, 2008.

[10] A. Adriansyah, B. F. van Dongen, W. M. van der Aalst, "Towards robust conformance checking," in International Conference on Business Process Management, 2010, pp. 122–133, Springer.

[11] M. Alizadeh, M. de Leoni, N. Zannone, "History-based construction of alignments for conformance checking: Formalization and implementation," in International Symposium on Data-Driven Process Discovery and Analysis, 2014, pp. 58–78, Springer.

[12] W. Song, H.-A. Jacobsen, C. Zhang, X. Ma, "Dependence-based data-aware process conformance checking," IEEE Transactions on Services Computing, 2018.

[13] R. Bosma, U. Kaymak, J. Berg, van den, H. Udo, "Fuzzy modelling of farmer motivations for integrated farming in the vietnamese mekong delta," in The 14th IEEE International Conference on Fuzzy Systems, 2005. FUZZ'05, United States, 2005, pp. 827–832, Institute of Electrical

and Electronics Engineers.

[14] E. S. Pane, A. D. Wibawa, M. H. Purnomo, "Event log-based fraud rating using interval type-2 fuzzy sets in fuzzy ahp," in 2016 IEEE region 10 conference (TENCON), 2016, pp. 1965–1968, IEEE.

[15] Z. Hao, Z. Xu, H. Zhao, H. Fujita, "A dynamic weight determination approach based on the intuitionistic fuzzy bayesian network and its application to emer-gency decision making," IEEE Transactions on Fuzzy Systems, vol. 26, no. 4, pp. 1893–1907, 2017.

[16] J. Li, L. Luo, X. Wu, C. Liao, H. Liao, W. Shen, "Prioritizing the elective surgery patient admission in a chinese public tertiary hospital using the hesitant fuzzy linguistic oreste method", Applied Soft Computing, vol. 78, pp. 407–419, 2019.

[17] S. Bragaglia, F. Chesani, P. Mello, M. Montali, D. Sottara, "Fuzzy conformance checking of observed behaviour with expectations," in Congress of the Italian Association for Artificial Intelligence, 2011, pp. 80–91, Springer.

[18] K. Ganesha, S. Dhanush, S. S. Raj, "An approach to fuzzy process mining to reduce patient waiting time in a hospital," in 2017 International Conference on Innovations in Information, Embedded and Communication Systems (ICIIECS), 2017, pp. 1–6, IEEE.

[19] A. Adriansyah, J. M. Buijs, "Mining process performance from event logs: The bpi challenge 2012," in Case Study. BPM Center Report BPM-12-15, BPM-center. org, 2012, Citeseer.

[20] L. Genga, M. Alizadeh, D. Potena, C. Diamantini, N. Zannone, "Discovering anomalous frequent patterns from partially ordered event logs," Journal of Intelligent Information Systems, vol. 51, no. 2, pp. 257–300, 2018.

[21] J.-S. R. Jang, C.-T. Sun, E. Mizutani, "Neuro-fuzzy and soft computing-a computational approach to learning and machine intelligence [book review]," IEEE Transactions on Automatic Control, vol. 42, no. 10, pp. 1482–1484, 1997.

[22] G. J. Klir, B. Yuan, Fuzzy Sets and Fuzzy Logic: Theory and Applications. Upper Saddle River, NJ, USA: Prentice-Hall, Inc., 1995.

[23] A. Cornelissen, J. Berg, van den, W. Koops, U. Kaymak, "Elicitation of expert knowledge for fuzzy evaluation of agricultural production systems," Agriculture, Ecosystems & Environment, vol. 95, no. 1, pp. 1–18, 2003.

[24] J. M. da Costa Sousa, U. Kaymak, Fuzzy Decision Making in Modeling and Control, vol. 27 of World Scientific Series in Robotics and Intelligent Systems. New Jersey: World Scientific, 2002.

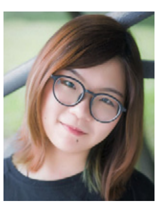
[25] G. Beliakov, A. Pradera, T. Calvo, Aggregation Functions: A Guide for Practitioners. Berlin: Springer, 2007.

[26] S. Zhang, L. Genga, L. Dekker, H. Nie, X. Lu, H. Duan, U. Kaymak, "Towards multi-perspective conformance checking with aggregation operations," in Information Processing and Management of Uncertainty in Knowledge-Based Systems, Cham, 2020, pp. 215–229, Springer International Publishing.

[27] R. Dechter, J. Pearl, "Generalized best-first search strategies and the optimality of a," Journal of the ACM (JACM), vol. 32, no. 3, pp. 505–536, 1985.

[28] H. Yan, P. Van Gorp, U. Kaymak, X. Lu, L. Ji, C. C. Chiau, H. H. Korsten, H. Duan, "Aligning event logs to task-time matrix clinical pathways in bpmn for variance analysis," IEEE Journal of Biomedical and Health Informatics, vol. 22, no. 2, pp. 311–317, 2017.

[29] L. Genga, C. Di Francescomarino, C. Ghidini, N. Zannone, "Predicting critical behaviors in business process executions: when evidence counts," in International Conference on Business Process Management, 2019, pp. 72–90, Springer.



Sicui Zhang

Sicui Zhang is a Ph.D. candidate at Department of Biomedical Engineering of Zhejiang University, China, and Department of Industrial Engineering, Eindhoven University of Technology, the Netherlands. Her research focuses on clinical decision support systems, business process management, conformance checking, and decision making processes.

Laura Genga

Laura Genga, received her Ph.D. in science of engineering at the Università Politecnica delle Marche, Italy, in 2016. Since January 2019, she has been an assistant professor in the Information Systems Group at the Eindhoven University of Technology, the Netherlands. Her core topics involve automated discovery and analysis of flexible pro-cesses, compliance analysis, and on-line process monitoring and prediction to support human analysts in detecting potential threats and in taking decisions regarding current process executions.

Hui Yan

Hui Yan received the first Ph.D. degree in biomedical engineering from Zhejiang University, China, in 2019 and second Ph.D. degree in industrial engineering from Eindhoven University of Technology, the Netherlands, in 2020. She is now working in Hainan University as an assistant professor. Her research interests include care pathway analysis, conformance checking, business process management.

Hongchao Nie

Hongchao Nie obtained the Ph.D. degree from Zhejiang University, China, in 2014 . He is currently a research scientist at Philips Research Eindhoven, the Netherlands. His academic activities have covered image processing, health IT, interoperability and process management. At Philips Research, his active research areas include process analysis, clinical informatics, machine learning and operation research.

Xudong Lu

Xudong Lu received the M.Sc. degree and Ph.D. degree in Biomedical Engineering from Zhejiang University in 1998 and 2001. He is a full professor in Biomedical Informatics Laboratory, Department of Biomedical Engineering, Zhejiang University. He is an openEHR Foundation Management Board Member, member of American Medical Informatics Association, and Hospital Information Management Sys-tem Society since 2007. He achieved several contributions on Business Process Management with Medical Intelli-gence, Guideline-based Clinical Decision Support Systems, Integrated EMR-S in China, and Integrated Physiology Information System through Knowledge Transfer.

Uzay Kaymak

Uzay Kaymak received the M.Sc. degree in electrical engineering, the degree of chartered designer in information technology, and the Ph.D. degree in control engineering from the Delft University of Technology, Delft, The Netherlands, in 1992, 1995, and 1998, respectively. From 1997 to 2000, he was a Reservoir Engineer with Shell International Exploration and Production. He is currently a Full Professor with the School of Industrial Engineering, Eindhoven University of Technology, Eindhoven, the Netherlands. He has co-authored more than 250 academic publications in the fields of intelligent decision support systems, computational intelligence, data mining, and computational modeling methods. His current research interests include fuzzy decision support, interpretable fuzzy modeling, computational intelligence, and intelligent systems design. Dr. Kaymak is an Associate Editor for the IEEE TRANSACTIONS ON FUZZY SYSTEMS and is member of the Editorial Board of multiple journals. He is a Past Chair of the Fuzzy Systems Technical Committee and the Computational Finance and Economics Technical Committee of the IEEE Computational Intelligence Society. He is also a board member of DSC/e (Data Science Centre Eindhoven) and of the Clinical Informatics study program (two-year post-master PDEng study) of TU/e and a member of the program and/or organization committee of multiple international conferences. Dr. Kaymak also holds a visiting professor position at the Zhejiang University, China.

LA UNIVERSIDAD
EN INTERNET

Rectorado
Avenida de la Paz, 137
26006 Logroño (La Rioja)
t (+34) 941 21 02 11

www.unir.net