Proceedings of the
AISB Convention 2023

Swansea University
13/14 April 2023

Berndt Müller (Ed.)

ISBN 978-1-908187-85-7

Preface

The 2023 convention of the Society for the Study of Artificial Intelligence and Simulation of Behaviour (AISB) was held as an in-person event on 13-14 April 2023. We believe the return of in-person conferences greatly benefits the communication and development of new ideas. It is also particularly beneficial for an interdisciplinary audience to experience and discuss different research approaches and cultures. This kind of experience is not possible at online conferences.

The convention consisted of several tracks based on popular symposia of previous AISB conventions instead of separately organised symposia. This decision was made to streamline the processes involved in organising the convention in the context of reduced timeframes due to COVID-19 disruptions from the past years and the longer-term changes this has brought about for the research community. In addition to the thematic tracks, we ran a general track accommodating papers on any topic related to AI. Each track acted like a mini conference with its own chairs and programme committee.

Berndt Mller
AISB 2023 Chair

Track I: AI & Philosophy

Chairs: Giusy Gallo, Claudia Stancati

With its origins in the Philosophy after AI symposium from previous AISB Conventions, this track sets out to investigate the philosophical and linguistic perspectives of the research paths which deal with language as it is conceived by AI. Contributions to a general philosophical discourse on AI are also welcome. We invite talks on the following (but not exclusively) topics:

- Linguistics and AI;
- AI research on language;
- Learning, creativity and AI;
- Meaning and AI;
- Creativity, machine-learning and language;
- Robots and communication;
- Truth, post-truth and AI;
- Social media, devices and human sociality.

Track II: General Track

Chair: Floriana Grasso

This track accommodates AI research that does not align with any of the other special tracks. Possible topics include:

- Agents and multi-agent systems
- Agent-based social simulation
- Verification of AI systems
- Combining learning and reasoning
- Argumentation
- AI & cyber security
- Natural language processing
- Knowledge representation

Track III: AI & Games

Chairs: Swen Gaudl, Marius Varga

This track focuses on the application of artificial intelligence or intelligent-like techniques, frameworks and theories to the creation of intelligent games. AI can be used in any manner suitable in the game, from the algorithm to making the game more engaging, personalised, and/or interactive. The following non-exhaustive list of research and practice shows potential submission topics:

- Use of AI techniques (e.g. planning, learning, evolution etc.)
- Design and engineering of AI components
- Procedural content generation
- Intelligent or adaptive player interaction
- Game analytics
- Data-driven player modelling
- Agent decision making systems
- Intelligent agents
- Environmental simulations
- Interactive narrative generation
- Intelligent narrative technologies
- Experimental AI
- Serious games & gamification

Track IV: AI Ethics

Chair: Berndt Müller

Recent years have seen an increased awareness of ethical issues stemming from a lack of responsibility in the design and deployment of AI- and data-driven technology systems. This track explores solutions to these issues. Some topics of interest:

- Algorithmic fairness
- Diversity
- Data governance
- Accountability
- Data privacy
- AI ethics principles
- Implementing ethical reasoning
- AI regulation
- Explainable AI

Track V: Computational Creativity

Chair: Jiaxiang Zhang

Computational creativity is continuing to attract researchers from both arts and science backgrounds. Philosophers, cognitive psychologists, computer scientists and artists have all contributed to and enriched the literature. Many argue a machine is creative if it simulates or replicates human creativity (e.g. evaluation of AI systems via a Turing-style test), while others have conceived of computational creativity as an inherently different discipline, where computer generated (art)work should not be judged on the same terms, i.e. being necessarily producible by a human artist, or having similar attributes, etc. More general topics of interest for this symposium include, but are not limited to:

- Novel systems and theories in computational creativity, in any domain, e.g. drawing and painting, music, storytelling, poetry, conversation, games, etc.
- The evaluation of computational creative systems, processes and artefacts
- Theory of computational aesthetics
- Representational issues in creativity, including visual and perceptual representations
- Social aspects of computational creativity, and intellectual property issues
- Creative autonomy and constraint
- Computational appreciation of artefacts, including human artwork

Contents

Preface iii

I. AI & Philosophy 1
Chairs: Giusy Gallo, Claudia Stancati

A Meta-Semantics Fit for Large Language Models 3
Jumbly Grindrod

Holding Large Language Models to Account 7
Ryan Michael Miller

Intelligence, super-intelligence, superintelligence++, and ChatGPT: Searching for Substance amidst the Hype 15
Joel Parthemore

From epistemology to ethics of deep networks 23
Juraj Hvorecký

My belief or Alexa's? Belief attribution and AI extension 25
Hadeel Naeem

Who Needs Needy Machine Consciousness? 26
Max Jones

Artificial Life as Controlled Disequilibrium 28
Steve Battle

II. General Track 33
Chair: Floriana Grasso

NLP Based Framework for Recommending Candidate Ontologies for Reuse 35
Reham Alharbi, Valentina Tamma, Floriana Grasso

Trust in Cognitive Models: Understandability and Computational Reliabilism 43
Noman Javed, Angelo Pirrone, Laura Bartlett, Peter Lane, Fernand Gobet

Local Minima Drive Communications in Cooperative Interaction 51
Roger K. Moore

Exploring the Impact of External Factors on Ride-Hailing Demand: A Predictive Modelling Approach 57
Anand Sriram, Ananthanarayana V. S.

Evolving Time-Dependent Cognitive Models 64
Peter Lane, Noman Javed, Angelo Pirrone, Laura Bartlett and Fernand Gobet

III. AI & Games 67
Chairs: Swen Gaudl, Marius Varga

Switchable Lightweight Anti-Symmetric Processing (SLAP) with CNN Outspeeds Data Augmentation by Smaller Sample - Application in Gomoku Reinforcement Learning 69
Chi-Hang Suen, Eduardo Alonso

IV. AI Ethics 77
Chair: Berndt Müller

Neighbor Migrating Generator: Finding the closest possible neighbor with different classes 79
Hassan Eshkiki, Benjamin Mora

Raising User Awareness of Bias-Leakage via Proxies in AI Models to Improve Fairness in Decision-making 86
Peter Daish, Matt Roach, Alan Dix

A position on establishing effective explanations from human-centred counterfactuals for automated financial decisions 89
Alex Blandin, Matt Roach, Daniele Doneddu, Jen Pearson, Matt Jones, David Sullivan

Towards European Standards supporting the AI Act: alignment challenges on the path to Trustworthy AI 98
Alessio Tartaro

Artefactual ethics as opportunity to rethink "natural" ethics 107
Joel Parthemore, Blay Whitby

V. Computational Creativity 113
Chair: Jiaxiang Zhang

Genetic Car Designer: A Large-Scale User Study of a Mixed-Initiative Design Tool 115
Sean P. Walton, Jakub Vincalek, Alma A. M. Rahat, James Stovold, Ben J. Evans

Exploring a Collaborative and Intuitive Framework for Combined Application of AI Art Generation Tools in Architectural Design Process 122
Lok Hang Cheung, Juan Carlos Dall'Asta

Part I.

AI & Philosophy

A Meta-Semantics Fit for Large Language Models

Jumbly Grindrod
Philosophy, University of Reading
Reading, United Kingdom
j.grindrod@reading.ac.uk

***Abstract*— I argue that large language models can be understood as meaningfully using the words they employ in their outputs. To do this, I apply Evans' notion of a naming practice and his subsequent distinction between producers and consumers. I argue that LLMs can be understood as pure consumers of a naming practice.**

Keywords—Large language models; Meta-semantics; Externalism; Philosophy of language

I. INTRODUCTION

In their [1], Cappelen and Dever consider whether the outputs of AI systems should be thought of as meaningful. The hypothetical case they consider is a supervised neural network trained on a bank's customer data in order to generate predictions regarding the riskiness of new customers (which they call *SmartCredit*). Suppose that the system outputs the following: "Lucy is a risky lendee". Should we take this output to *mean* that Lucy is a risky lendee? The very idea of a meaningful entity, whether that is a thought, utterance, a text, a photograph, or a map, is vexing one, and explaining what differentiates meaningful from non-meaningful entities has been one of the central questions in philosophy of mind and philosophy of language. We might say that a state is meaningful when it represents the world as being a certain way (as all of the entities listed above seem to), but then we need some account of how a state gains representational properties. Such an account is known as a *meta-semantics*. Cappelen & Dever consider whether there is a plausible meta-semantics that would capture the idea that AI systems do mean what they output.

For this paper, I will extend their line of inquiry and their approach to a different kind of artificial intelligence system: large language models (LLMs). I argue that Evans' (1982) account of naming practices provides a plausible meta-semantic framework with which to understand the meaningfulness of the outputs of large language models. Specifically, Evans' account of a naming practice and the subsequent distinction between *producer* and *consumer* of the naming practice helps shed much greater light on the way in which LLMs can be viewed as users of a language. The resulting picture of LLMs as *mere consumers* is attractive insofar as it is independently motivated and also provides an attractive middle ground between the diametrically opposed viewpoints that, on the one hand, LLMs are merely "stochastic parrots" displaying patterns of distribution, and on the other hand, LLMs are language users in the fullest sense. The remainder of the paper is structured as follows. In the second section I will briefly outline what is meant by a large language model, with a particular focus on the distributional semantic approach that lies at their core. In the third section, I will outline how Cappelen & Dever's meta-semantics for AI systems is not straightforwardly applicable to LLMs. In the fourth section, I will outline Evans' [2] account of naming practices, and show how this provides a more plausible basis to understand LLMs as language users. In the fifth section, I will consider certain objections against the proposal.

II. EASE OF USE

In this section, I will outline a core feature that lies behind much of the recent progress with LLMs in recent years: the distributional semantic approach to meaning. For those unfamiliar with distributional semantics, this section can serve as an informal introduction to the approach; for those already familiar, this section will make clear which of the technical details of LLMs I am placing particular emphasis on, and which of the technical details I will ignore.

This approach begins with what is known as the *distributional hypothesis*: broadly, the claim that the meaning similarity of two expressions will correlate with their similarity of distribution across a corpus [3]–[7]. The distributional profile of an expression can be modelled geometrically as a vector in a high-dimensional space and so we can likewise represent the meaning of an expression. Geometric similarity measures can then be used (e.g. cosine similarity or Euclidean distance) to measure meaning similarity. Earlier count models would achieve this directly from the corpus co-occurrence statistics between words. For instance, a very simple count model could have, for each word, a vector whose components would represent how often that would co-occur within the vicinity of another word.

Around 2013, it was found that a more effective method for producing such vectors is to set a neural network some language modelling task and then extract word vectors from it. For instance, the famous word2vec algorithm package employs a neural network with three layers – the input layer, middle layer, and output layer [8]. The number of neurons in the input and output layers will be the same size as the vocabulary for the language i.e. if we want 10,000 vectors, then we have 10,000 neurons at these levels. The number of neurons in the middle layer will equal the dimensionality of our vectors. The network is trained on a language modelling task where it needs to predict some missing word or set of words. During the training procedure, the network will mask words in the training data and then attempt to predict what was masked, given the word's sentential context, with the weights of neural connections repeatedly adjusted so as to improve the predictions. The connections between each neuron at the input layer and every neuron at the middle layer can then be treated as a vector representing the meaning of that word.

As is by now well-known, this approach proved to be a revelation for NLP tasks related to semantic meaning. The particular power of this approach is that it provides away of capturing word meanings using raw text and self-supervised learning, meaning that it is highly scalable.

Another major advance has been the introduction of transformer architectures [9]. The key guiding idea behind transformer architectures is that the meanings of words are best understood in the linguistic context in which they are used. Rather than having a single vector for every instance of "bank", it would be preferable that each instance of "bank" receives its own vector, which reflects not just properties of the word type, but also the sentence in which the token of that word type appeared. I will not here give a detailed account of how transformer architectures work. Very roughly, transformer architectures consist in layers of self-attention heads that take as input an embedding and produce as output a new embedding that is partly determined by features of the sentential context and partly determined by the original embedding. What each self-attention head is sensitive to is determined in training, and there is some evidence that specific self-attention heads are dedicated to specific linguistic features [10].

The above can only give the roughest idea of how large language models operate. But one point that is worth emphasizing here is that even in the much more sophisticated (and much larger) transformer architectures, there is a compelling case to be made that ultimately the approach is still a more sophisticated version of the distributional semantic approach that we began with. After all, transformer architectures still only need to be trained on a corpus of raw text (even if it is a massive one), and so it is not receiving any further information beyond the patterns of distributions it can detect in that corpus. In what follows then, I will largely focus on LLMs as distributional semantic models. This is something of an idealization, as many LLMs today will combine this distributional approach with other forms of learning. For instance, it is well-documented that Chat-GPT has used both supervised learning and reinforcement learning from human feedback to reach improved performance beyond the core GPT-3/GPT-4 language models. I will ignore these further possibilities and instead focus on the distributional approach.

III. CAPPELEN AND DEVER'S AI META-SEMANTICS

Anyone who has had the chance to interact with LLMs like GPT-3 or similar will see the intuitive appeal of the claim that their outputs are meaningful. Indeed, one of the impressive features of LLMs, which is completely lacking in the kind of AI system that Cappelen and Dever consider, are their ability to complete a vast range of text-based tasks, from writing poems to writing code, to answering questions, to summarizing etc. This is markedly different from the AI system that Cappelen and Dever consider that only performs a single task: that of predicting the riskiness of a given lendee. This flexibility is good reason to think that *if* Cappelen and Dever are right that the outputs of the banking AI should be thought of as meaningful, then the outputs of LLMs should be thought of as meaningful as well.

In order to follow the lead of Cappelen and Dever, it will be instructive to see how they employ an externalist meta-semantics to capture the meaningfulness of their hypothetical banking system. They first focus on the question of how SmartCredit could be thought of as referring to the property of *being a risky lendee*. So when it outputs "Lucy is a risky lendee", is there a meta-semantic theory that could capture the idea that it is representing a state of affairs whereby Lucy possesses that very property. They take as their starting point Kripke's [11] *causal-historical* theory of reference. According to the view, a speaker uses an expression to refer to a particular object or property just in case they stand in an appropriate causal-historical chain that links back to an original naming event where a speaker first decided to use the expression to refer to that object or property. This is one of the clearest examples of an externalist meta-semantics insofar as we appeal to features outside of the speaker (i.e. causal historical relations, potentially stretching back a long way) to explain why the words of a given speaker have the meaning that they do. One of the main motivations for Kripke's view was to reject the idea that for words to have the meanings that they do, the speaker must only have internalized some definition or set of necessary and sufficient conditions. This descriptivist idea inevitably leads to a psychologically implausible account of what has been internalized, and as Kripke showed, delivers the wrong results across a range of different cases.

That said, Kripke [11, p. 96] did still claim that speakers have to have a *referential intention* to use the word in the same way that it has previously been used across the causal-historical chain that stretches back to the original naming event. This requirement of a certain kind of mental state proves problematic when we turn to consider AI systems. It is highly implausible, for instance, that SmartCredit has any sort of intentions of the sort. However, Cappelen and Dever instead suggest that a suitable meta-semantics for AI needs to de-anthropomorphize away from appeals to human level mental states like communicative intentions. Rather than requiring that there is some naming ceremony that marks the beginning of a causal-historical chain, Cappelen and Dever instead argue that SmartCredit becomes anchored to the property of riskiness thanks to the training procedure that it is put through. The precise point of that training procedure is that it modifies its predictions until it reliably predicts whether a given lendee is risky. So, Cappelen and Dever argue, the point we take from Kripkean meta-semantics is that there has to be some anchoring event to the property in question. But whereas humans usually rely upon some initial naming event, an AI system like SmartCredit will anchor to the property in question through its initial training. It is this that imbues its outputs thereafter with the content that it intuitively seems to have.

Can we tell a similar story about LLMs? The answer is no, for three reasons. First, Cappelen and Dever appeal specifically to a supervised training procedure where the programmer has provided a labelled dataset that the neural network can then use to modify its own behaviour until it becomes sensitive to the property in question (e.g. riskiness). As outlined earlier, there is no supervised training required in producing large language models. The primary form of training that takes place is *self-supervised*: where the model will mask words in a corpus, for every word in the vocabulary provide a prediction score as to whether it is the masked word (given the words that appear alongside the masked word), and will then adjust the neural connections based upon the (in)accuracy of its predictions. So any appeal to a training phase that serves to anchor to a specific property will not apply to models that employ self-supervised learning. Relatedly, whereas the imagined SmartCredit is trained specifically to track a particular property, LLMs are trained at a much more general level. They are trained to track the distributional properties of an entire language as represented by a large corpus, rather than a specific property.

Finally, and most importantly, the type of data that LLMs are trained on is linguistic data i.e. raw text – essentially a record of linguistic use. This is different in kind from the data that SmartCredit is trained on – data that is about entities in the world e.g. the banking behaviours of customers. This difference has arguably been at the heart of a great deal of skepticism that LLMs can be viewed as language users in any proper sense; instead, the thought goes, they are merely "stochastic parrots" [12] mimicking the patterns of distribution that they identified in their training data. Like

Cappelen and Dever, I do not want to directly reject such skepticism by criticising the arguments in favour of it, but I instead want to explore the plausibility of the non-skeptical approach. So the point about the textual training data that I want to make here is not that this inevitably leads to skepticism about the meaningfulness of LLM outputs but that this prevents us from appealing to the kind of anchoring event that Cappelen and Dever appeal to regarding riskiness. Instead, the meta-semantic account would need to appeal to the way in which language users can become language users by engaging with the existing conventions of use. This, broadly put, is what LLMs do.

IV. EVANS ON NAMING PRACTICES

To continue with Cappelen and Dever's approach, we need to turn to an alternative meta-semantics. They do this themselves in considering how SmartCredit could be thought of as making a claim about e.g. Lucy when it outputs "Lucy is a risky lendee". After all, the story given in the previous section about the importance of the training procedure in anchoring onto relevant parts of the world cannot be given about Lucy, as Lucy has played no part in the training procedure. Instead, by contrast, it is initially tempting to view the way that SmartCredit has come into contact with Lucy by storing and processing information about Lucy. After all, Lucy is to SmartCredit essentially a row in a table containing information about individuals. It is tempting to think then that if SmartCredit is to represent Lucy at all, it is in what philosophers would describe as a form of *descriptivism*, as the unique individual that possesses all of the properties described on the Lucy row. There are, however, familiar objections against descriptivism, and they apply here. For instance, if one piece of information on the Lucy row is incorrect, this would not seem to prevent the Lucy row from being about Lucy, although this is what is implied by descriptivism (i.e. as the view that referring to an object amounts to employing a description that uniquely applies to that object).

Cappelen and Dever instead argue that we can understand the way in which Lucy is represented here in terms of an alternative externalist meta-semantics: the *mental files framework*. This approach has its origins in the work of (among others) Gareth Evans [2]. While there are many different ways in which the mental files approach can be developed, I will outline only the bare bones of the proposal here in terms inspire by Evans. According to the mental files view, in order to represent an object, one has to store information about that object. However, while the information stored supposed to be about the object, it is not a condition on being able to represent the object that the object actually satisfies the description. Instead, the information is about the object insofar as it has been stored in (to use the usual metaphor) the *mental file* for that object. So what determines which file information gets stored in? One candidate answer that will serve our purposes well enough is that the information is about a particular object (and so is stored in the file for that object) insofar as the object is the dominant causal source of the information. If my pen pal lies to me about every aspect of her life, such that the information I have stored about my pen pal is radically incorrect, my mental file is still about my pen pal insofar as she is the source of the information. Turning to SmartCredit, Cappelen and Dever argue that this provides a plausible account of how SmartCredit refers to Lucy. SmartCredit stores and processes information about an individual for which Lucy is the dominant causal source, and it is in virtue of that fact that when SmartCredit outputs that Lucy is a risky lendee, that output is *about* Lucy, rather than anyone else.

I want to suggest that the mental files approach, and particularly Evans' early version of the view, actually provides a basis for understanding how LLM outputs are meaningful across the board. To see this, we have to turn to Evans' idea of a *naming practice*. According to Evans, for each name there is an associated naming practice that can be viewed as a body of information associated with the named object. This is akin to a mental file, but one that belongs to a linguistic community rather than an individual speaker. He then distinguishes between *producers* and *consumers* of the naming practice, where a producer is someone who stands in a particular acquaintance relation with the named entity, such that they have certain recognitional abilities for that entity and are able to demonstratively refer to the entity [2, p. 376]. In the simplest case, they can recognise a person, point to them, and say "There is (e.g.) Risha". Because they are able to engage with Risha in this more direct way, the producers are in a position to add to the body of information that constitutes a naming practice. More generally, we might say that there is some epistemic position regarding the named entity that producers must have available to them. But crucially, one need not be a producer to engage in the practice. Consumers are those who engage in the practice by storing the information embodied in the naming practice. When they have done so, they can use the name to refer to the entity just as the producers can. So while consumers are in an important sense dependent on producers – without producers there would be no naming practice at all – consumers are still users of the name in the fullest sense.

With this distinction between consumer and producer in place, the possibility arises that LLMs use words meaningfully insofar as they engage with the relevant naming practice. On this picture, we view LLMs as consumers of a naming practice, as a user who has engaged appropriately with the information associated with a given expression such that they can successfully use the expression. I suggest that this is a particularly attractive picture of LLMs as language users as it captures the sense in which they are indeed producing meaningful outputs while also allowing for the fact that there is an important aspect of being a language user that they miss out on. LLMs, the thought goes, cannot be producers in a naming practice as they are never in a suitable epistemic position to add information to the naming practice. They are, instead, always consumers, engaging with the naming practice by becoming sensitive to the associated body of information. As pure consumers, their role as a language user is parasitic. LLMs requires a huge corpus of previous language use to train on, and so are clearly reliant on a previous history of human usage. But this just seems to reflect the asymmetric relationship between producers and consumers more generally – consumers rely on producers for there to be a naming practice in the first place. And while we as humans will act as consumers for some expressions, producers for others, LLMs can only act as consumers, as they lack the kinds of abilities required to add to a naming practice. A further benefit (which it inherits from the broader externalism) is that it clearly accommodates the possibility that LLMs may store incorrect information about a given individual (something that any descriptivist view of meaning would struggle to capture).

V. OBJECTIONS

A. No Information Stored

We said that consumers can engage with a practice by setting up a mental file and storing the information associated with the practice in order to partake in it. But (so the objection goes) LLMs do not do this. Instead, there is for each expression, a representation of its distribution across a corpus, or, for more complex models (i.e. transformers), some combination of its distribution across a corpus combined with some representation of its position in the sentence of interest. This is the wrong kind of information to engage with a naming practice, and so LLMs cannot be treated even as consumers.

Contrary to this objection, LLMs do plausibly store information about the expressions that they use. The most straightforward way to see this is to prompt an LLM with precisely the kind of question that one might prompt an ordinary speaker to find out about the information they have stored e.g. "Tell me about ["word"]". A prompt of "Tell me about Brighton" to GPT-3 returns:

> *Brighton is a town in East Sussex, England. It is located on the south coast of England, about an hour south of London. Brighton is a popular tourist destination, known for its lively atmosphere, diverse population, and seaside location. The town is home to a number of attractions, including the Brighton Pier, the Royal Pavilion, and the Brighton Dome. Brighton is also a popular destination for nightlife, with a number of clubs, bars, and restaurants.*

I suggest that included in this response is precisely the type of information that Evans had in mind when considering what would be included in a naming practice. Now it certainly is remarkable that it is able to achieve this through the use of a distributional analysis of a large corpus, and the very idea of being able to store a representation in a distributed fashion across a large neural network is itself one that has previously attracted a great deal of philosophical discussion [13]. But merely pointing to the distributional basis of LLMs is not sufficient to show that such representations are not present. It is surprising the extent to which LLMs can store information, but it is hard to deny that this is what they have done [10].

B. Too Much Information Stored

It might instead be argued that LLMs store *too much* information to be plausibly thought of as engaging with a naming practice. One feature of current LLMs is that they possess a huge amount of accurate information about the world. Chat-GPT appears to be able to cite any line from the bible, chapter and verse. But we wouldn't think this is part of the naming practice for "The Bible" for instance. That thought certainly seems correct, but provided that the LLM does successfully store the information that is associated with the naming practice, there seems to be no special problem here. The LLM is in this respect just like a Bible scholar, whose information far outstrips just what would be included as part of the practice.

VI. CONCLUSION

I hope to have shown how Evans' distinction between producer and consumer provides an attractive account on the extent to which LLMs are users of a language. A valuable area of future investigation would be to consider the extent to which amendments to Evans' view would lead to a different verdict on LLMs. For instance, a case could made that Evans' conditions on who counts as a producer are too strict. Perhaps all that is required to be a producer is that you are in fact able to influence the body of information that forms part of the naming practice. This weaker notion of a producer might lead to the outcome that LLMs can be producers after all. However, in their current state, and with their influence as it is at the current time, they are restricted to the consumer role.

REFERENCES

[1] H. Cappelen and J. Dever, *Making AI intelligible: philosophical foundations*. Oxford: Oxford University Press, 2021. doi: 10.1093/oso/9780192894724.001.0001.

[2] G. Evans, *The varieties of reference*, no. 137. Oxford: Oxford University Press, 1982, p. 515.

[3] Z. S. Harris, 'Distributional structure', *Word*, vol. 10, no. 2–3, pp. 146–162, 1954, doi: 10.1080/00437956.1954.11659520.

[4] JR. Firth, 'A synopsis of linguistic theory', in *Studies in Linguistic Analysis*, Oxford: Blackwell, 1957, pp. 1–32.

[5] A. Lenci, 'Distributional semantics in linguistic and cognitive research', *Italian Journal of Linguistics*, vol. 20, no. 1, p. 32, 2008.

[6] K. Erk, 'Vector space models of word meaning and phrase meaning: a survey', *Language and Linguistics Compass*, vol. 6, no. 10, pp. 635–653, Oct. 2012, doi: 10.1002/lnco.362.

[7] G. Boleda, 'Distributional semantics and linguistic theory', *Annu. Rev. Linguist.*, vol. 6, no. 1, pp. 213–234, Jan. 2020, doi: 10.1146/annurev-linguistics-011619-030303.

[8] T. Mikolov, K. Chen, G. Corrado, and J. Dean, 'Efficient estimation of word representations in vector space', *CoRR*, Jan. 2013, Accessed: Oct. 14, 2021. [Online]. Available: https://arxiv.org/abs/1301.3781v3

[9] A. Vaswani *et al.*, 'Attention Is All You Need', no. Journal Article, 2017, doi: 10.48550/arxiv.1706.03762.

[10] A. Rogers, O. Kovaleva, and A. Rumshisky, 'A Primer in BERTology: What We Know About How BERT Works', *Transactions of the Association for Computational Linguistics*, vol. 8, no. Journal Article, pp. 842–866, 2020 2021, doi: 10.1162/tacl_a_00349.

[11] S. A. Kripke, *Naming and Necessity*, Rev. and Enlarged ed. Blackwell, 1980.

[12] E. M. Bender, T. Gebru, A. McMillan-Major, and S. Shmitchell, 'On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? 🦜', in *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, in FAccT '21. New York, NY, USA: Association for Computing Machinery, 2021, pp. 610–623. doi: 10.1145/3442188.3445922.

[13] C. Buckner and J. Garson, 'Connectionism', in *The Stanford Encyclopedia of Philosophy*, E. N. Zalta, Ed., Fall 2019.Metaphysics Research Lab, Stanford University, 2019. Accessed: Apr. 03, 2023. [Online]. Available: https://plato.stanford.edu/archives/fall2019/entries/connectionism/

Holding Large Language Models to Account

Ryan Michael Miller
Philosophy Department
Université de Genève
Geneva, Switzerland
Ryan.Miller@unige.ch

Abstract— **If Large Language Models can make real scientific contributions, then they can genuinely use language, be systematically wrong, and be held responsible for their errors. AI models which can make scientific contributions thereby meet the criteria for scientific authorship.**

Keywords— Large Language Models, authorship, responsibility, reference, hallucinations

I. THE AI AUTHORSHIP CONTROVERSY

Large Language Models (LLMs) are transformer-based deep-learning neural networks with hundreds of billions of parameters trained by self-supervised learning on large text corpora to perform next-token prediction. OpenAI's November 2022 public release of their 175-billion parameter GPT-3.5 model trained with Proximal Policy Optimization [1] made available for the first time an AI with human-level performance on a wide range of cognitive tasks [2] and its 4,096 token context window (~3000 words for prompt + response) allowed a wide domain of application [3]. The March 2013 release of GPT-4, with a maximum 32,768 token (~24,000 word) context window, performance at the upper end of the human scale on many cognitive tasks, and twice the measured factual reliability [2] has only increased the possible uses.

One such use of LLMs is the production of scientific research, with hundreds of papers appearing on preprint servers with AIs listed as co-authors, some of which have been published with that authorship credit after peer review [4]. Since use of LLMs not only speeds the writing [5] and revision [6] process but also helps with literature review[7], algorithm development, data analysis, hypothesis generation [8] and even creativity [9] and argumentation [10], we can expect such use to continue to grow. Unlike in the case of previous computerized text generators like SCIgen [11] which merely slipped gibberish through sham or slipshod refereeing processes [12]–[15], LLMs generate text which can be genuinely useful and is sometimes undetectable even by dedicated referees [16], [17]. Until recently, the vast majority of ethical concern around LLM authorship has been about plagiarism [6], [18], [19]. Consequently, accountability efforts have focused on ways to deter or detect LLM use in scientific writing [16], [17].

Giving the LLM authorship credit neatly sidesteps plagiarism issues, however: if the LLM is listed as an author of the paper, then there can be no allegation that the other authors plagiarized from the LLM or that the contribution of the LLM lacked transparency. This transparency is further increased for journals which use a structured author contribution statement [20], [21] or contributor roles taxonomy [22], which would list the exact research and writing contributions made by the LLM to the final published product. Current suggestions for making such roles more specific [23] only raise the likelihood that LLMs would qualify for authorship. Furthermore, while not all actual writers of scientific literature must receive authorship credit in all disciplines according to prevailing ethical standards [24], almost one third of publication ethics codes and more than half of Social Sciences Citation Index journals *require* authorship credit for all participants in drafting and revising the text [25]. Even in the remainder which also require scientific contributions, LLMs may qualify given the capabilities discussed above. Certainly, in many of the existing exemplars of published peer-reviewed scientific work with LLM authorship credits the LLM must make a "substantial scientific contribution" if the work has one at all, since the vast majority of the text and nearly all of the argument comes in the form of text from unedited LLM token output. Without crediting LLMs as authors it is difficult to see how papers where they contribute substantially could comply with the International Committee of Medical Journal Editors (ICMJE) original fourth principle for authorship [26]:

> *Each part of the content of an article critical to its main conclusions and each step in the work that led to its publication [(1) conception or design of the work represented by the article, or analysis and interpretation of the data, or both; (2) drafting the article or revising it for critically important content; and (3) final approval of the version to be published] must be attributable to at least one author.*

Cases where LLMs have received authorship credit have involved every one of these steps [27], [28].

Nonetheless, the influential Committee on Publication Ethics (COPE) and World Association of Medical Editors (WAME) have called for banning AI authorship on the grounds that AIs "cannot take responsibility" for their output [29], [30], and this call has been heeded by *Nature* [31] while other influential journals have banned AI authorship without giving explicit reasons [32], [33]. ChatGPT's authorship has been retracted in one case on this basis [34]. COPE's standard combines a general responsibility test with a long history in the publication ethics literature going back to [26] with a more recent legal personhood test supposedly required for "assert[ing] the presence or absence of conflicts of interest" and "manag[ing] copyright and license agreements" [29]. WAME spells out the latter, legal test as a matter of the corporate form chosen by OpenAI and its disclaimer of responsibility [30], which are obviously contingent matters not essential to AI. Indeed, various forms of legal personhood have already been proposed for algorithms which would allow them to enter into contracts [35]–[38] and corporations may soon be forced to assume liability for the AIs they create [39]–[41]. LLMs are as capable of asserting the presence or absence of conflicts of interest as they are of asserting anything else. Philosophical interest in COPE's new standard thus lies with its responsibility test, which is supposed to be an addition to (or even restriction of) the "substantial scientific contribution" standard for authorship which LLMs cannot meet even if or when they meet the latter standard.

COPE's responsibility standard goes back to ICMJE's original first principle for authorship [26]:

> *Each author should have participated sufficiently in the work represented by the article to take public responsibility for the content…[which] means that an author can defend the content of the article, including the data and other evidence and the conclusions based on them. Such ability can come only from having participated closely in the work represented by the article and in preparing the article for publication. This responsibility also requires that the author be willing to concede publicly errors of fact or interpretation discovered after publication of the article and to state the reasons for error. In the case of fraud or other kinds of deception attributable to one or more authors, the other authors must be willing to state publicly the nature and extent of deception and to account as far as possible for its occurrence.*

LLMs like ChatGPT manifestly both defend their output [9] and apologize for mistakes while giving reasons for their occurrence [6], [9] as well as identify particular human co-authors by their writing and offer criticisms [6]. WAME additionally references the current ICMJE standard that all authors must provide "Final approval of the version to be published" [42] as a reason that AIs cannot meet the general responsibility test [30]. While some publications with LLMs listed as co-authors may be suspect in this regard [27], ChatGPT's unwillingness to co-author is likely a result of its Reinforcement Learning from Human Feedback (RLHF) and is obviously not essential to LLMs. The COPE/WAME/*Nature* general responsibility test for authorship is thus best understood as a normative claim rather than a legal or behavioral one. ICMJE's "criteria are not intended for use as a means to disqualify colleagues from authorship who otherwise meet authorship criteria" [42], so the question is whether LLMs which meet the research and writing standards for authorship are able "to be accountable" in some normative sense. This is a fundamentally philosophical question.

The philosophical response to COPE's general responsibility test for AI authorship has been mixed. Wiese grants that current AIs are insufficiently agential to meet this constraint, but holds that future "strong artificial consciousnesses" which observe the Free Energy Principle would exhibit the relevant normative properties [43]. Jenkins and Lin, by contrast, argue that many uncontroversial human authors (e.g., deceased ones) also cannot take responsibility, so that only the research and writing standards are appropriate [44]. Another similar approach suggests that responsibility for scientific publications is best understood as irreducibly collective among the authors [45] so that AIs are accountable as part of a system with relevantly-situated humans [46], i.e. co-authors. On this approach, if there is a single human co-author to take responsibility, then the authorship team as a whole does, and further accountability is required of the AI. I take a third approach: *if* LLMs meet the research and writing standards for substantial scientific contribution, then Wittgenstein's Private Language Argument suggests that they *ipso facto* meet the responsibility standard.

II. AI AND LANGUAGE USE

It is an open question whether Large Language Models count as *users* of language. Until recently, doubts about AI language use could be framed in terms of objective qualities of the token output. Much of SCIgen [11]'s output was "gibberish" [13], [15] with approximately English syntax comparable to Chomsky's famous nonsense-sentence "colorless green ideas sleep furiously" [47]. It may have entered into the scientific literature through inattentive review or pay-for-play predatory publishers, but readers would likely struggle to identify propositional contents or truth conditions for its sentences. Since *use* of a declarative sentence requires that it be truth-apt [48], SCIgen would not count as a language user, and thus presumably could not count as an author by the writing standard. Since SCIgen papers by common consensus make no scientific contribution, it would not count as an author by that standard either, making the responsibility test moot.

The situation for early LLMs like GPT-2 and GPT-3 is somewhat murkier. The outputs of these models frequently seem sensible and truth-apt, and often pass undetected by human reviewers [49]. Nonetheless, sophisticated humans [50] and detector programs are able to reliably distinguish their outputs from human-generated ones [51]–[53]. Something "robotic" seems to characterize early LLM outputs, and they often seem to over-fit and reproduce text verbatim [49], giving credence to the view that they merely represent "surface statistics" [54], [55] rather than being genuine products of AI language use. Insofar as the outputs, unlike those of SCIgen, are genuinely interpretable, they are about something—*intentional*—and therefore exemplify language use. The question in these cases is whether the LLM itself is the language user. After all, evolutionary extensions of ELIZA [56] can sometimes also pass undetected by human reviewers, with well-formed and apparently meaningful output [57], yet they are purely procedurally coded, with outputs that merely reflect the prompt and programmer instructions. In these cases the intentionality of the output is *projected* by the programmer and prompt engineer, who are the true language users, rather than being attributable to the LLM [58]. This impression is further reinforced by the characterization of such early LLMs as few-shot learners [59], which rely on the structure of the prompt for reliably meaningful output. It is at least plausible that these early LLMs should be treated like automatic grammar and spell-checkers or translators which produce written output without rising to the level of language users.

With more recent LLMs and other related neural-network AIs the situation has changed. Victories at Diplomacy by Meta's Cicero model [60] are relatively convincing Turing-tests, and there is empirical evidence and strong theoretical considerations suggesting that the output of current and future state-of-the-art LLMs will not be reliably detectable [61]. ChatGPT-3.5 and -4 are adept zero-shot learners [62] which frequently outperform humans in zero-shot language tasks [63], [64]. In its new "Code Interpreter" mode, ChatGPT is able to analyze an uploaded dataset, generate interesting hypotheses about it, perform statistical tests of those hypotheses, and write up the results in a typical scientific article format [65]. Given this level of capability, it is no wonder that most readers simply take for granted that LLMs are language users [66]. Any philosophical debate over this question must therefore turn from analysis of LLM outputs to Searle-style "Chinese Room" arguments [67] about internal states. On some popular accounts, scientific writing is only successful when it conveys the theory or model held by the scientist [68], [69]. Since GPT-4 is merely a scaled-up and further-trained version of GPT-3, it may be just a more sophisticated implementer of "surface statistics" without any such internal model [54], [55]. If, *pace* Jonas Bozenhard [70], these strictures on language use are correct *and* LLMs lack internal models, then even state-of-the-art LLMs would not count as language users, no matter the sophistication or apparent value of their output. On the other hand some researchers argue that state-of-the-art LLMs *do* possess world models, and are therefore capable of genuinely representational writing [71], [72]. In my view this debate remains open.

What is important to recognize, however, is that *if* LLMs are not language users then they cannot meet the scientific contribution standard for authorship. After all, the only outputs of LLMs are linguistic. If they are not language users, then that language is not their own but merely represents the projected intentionality of credit-worthy human authors. In that case, LLMs would clearly fail to meet the drafting and revision test for scientific contribution in the same way that both human and automated translators and copy-editors fail to meet that test. Even in cases where LLMs purportedly meet the scientific contribution standard via the research test, e.g. by formulating hypotheses or running statistical tests, their output is only linguistic tokens. If those tokens are not genuinely attributable to the AI, then the LLM has not contributed to the research in a qualitatively greater way than traditional statistical software packages do. As Jenkins and Lin [44] suggest, "continuity" is generally required for authorship credit, and that continuity does not terminate in the LLM if it is not a genuine language user. Surely this state of affairs would justify *Nature*'s insistence that LLMs be disclosed in methods sections but not credited as authors [31]. What this state of affairs would *not* do, though, is justify the reason given for *Nature*'s insistence: namely that LLMs which meet the scientific contribution standard are nonetheless ineligible for authorship by the responsibility standard. Whether LLMs are capable of language use, and thus of scientific contribution, is up for debate as discussed above. Naturally no person or entity which fails to make a scientific contribution should be listed as an author. But what COPE/WAME/*Nature* insist is that LLMs are not authors *even if* they make scientific contributions, and debates about whether LLMs can genuinely use language are irrelevant there.

For the purposes of this paper, then, we can merely assume that LLMs *are* capable of language use and hence of making scientific contributions, in order to focus on the responsibility test. The relevant question is thus not *whether* LLMs are language users, but rather *how* they can use language if indeed they do so. Where might the requisite non-projected intentionality required for meaning [58] come from? It cannot come via embodiment or ostension, since LLMs only inputs are prompts and training data made up purely of linguistic tokens. There is no embodied or sensory reality present to the LLM which it could correlate with those conventional signs. Bozenhard [70]'s Wittgensteinian approach is the only remaining option. In Wittgenstein's analysis, language is learned as a kind of game, which involves following semantic and syntactic rules [73]. What LLMs learn via initial unsupervised training and later RLHF is the rules of the language game, whether syntactic, semantic, or pragmatic. That they have in fact learned the rules is evident because the vast majority of their outputs (especially for ChatGPT-4) are syntactically correct, semantically meaningful, and pragmatically appropriate according to human readers steeped in more or less the same corpus of texts used for training the LLM. If LLMs were not language users, they would not play the game correctly and would only come across apt formulations by chance, as a toddler playing with a chess set might make a legal move. The sheer utility of state-of-the-art LLMs obviously precludes this interpretation: they almost always make syntactically, semantically, and pragmatically correct language moves, even in very difficult scenarios, at a rate vastly exceeding chance. LLMs are not the proverbial monkeys with typewriters. On this Wittgensteinian account, then, LLMs are language users because they are capable of following the rules, as evidenced by their outputs, and language is public, not a matter of what is in the LLM's head. Conversely, *if* LLMs are language users, it is because they have learned the rules of the language game as reflected in their linguistic token output.

III. LANGUAGE USE, NORMATIVITY, AND RESPONSIBILITY

This Wittgensteinian characterization of LLM language use as rule-following has implications for AI's ability to meet the COPE/WAME/*Nature* general responsibility test for authorship. The reason is that Wittgenstein [73]'s rule-following account of language use was given in service of an argument that language use is always public (not a matter of private cognition), and that argument runs through a further premise about normativity. While the exact location and structure of Wittgenstein's so-called Private Language Argument are subject to dispute [74]–[76], I follow Roger Harris in reconstructing it as follows [77], [78]:

P1 (LANGUAGE): language → rule-following (language is used only if rules are followed)

P2 (NORMATIVITY): rule-following → ◊ systematic error (rule-following implies the possibility of systematic error)

P3 (SUBJECTIVITY): ◊ systematic error → ¬ rules known wholly by introspection (the possibility of systematic error precludes exclusively introspective epistemic access)

P4 (PRIVACY): ¬ rules known by pure introspection → ¬ wholly grounded by internal mental life (epistemic access outside pure introspection precludes internal mental grounds, i.e. privacy)

C (PLA): language → ¬ grounded wholly by internal mental life (language is never private)

While the conclusion of the argument is relevant for the discussion of the last section, here the focus is on NORMATIVITY, which Wittgenstein contends is concomitant to all language use since it is a necessary property of rule-following activity. This kind of normativity is quite minimalist, since it follows Wittgenstein's general proclivity to focus on publicly observable facts about language use rather than facts about the internal state or structure of the language user.

Wittgenstein's NORMATIVITY premise is easily validated in the case of LLMs, contrary to Fodor's thought that computer language use necessarily follows the rules [79]. State-of-the-art LLMs show evidence of rule-following by generating syntactically, semantically, and pragmatically appropriate output in the vast majority of cases, but they also have characteristic failure modes called "hallucinations" where that output fails to conform to semantic rules [80]. While such hallucinations are reduced in state-of-the-art LLMs by comparison to prior models [2], patently false claims still appear often in their output, with implications for the reliability of scientific writing [81]. Indeed, the presence of such hallucinations forms a major part of WAME [30]'s argument against using AI to author scientific papers. Unlike the case of procedurally-written chatbots like ELIZA, LLM hallucinations cannot be extrinsically assigned to the programmer as "bugs" while the LLM itself is considered as an immaculate mathematical function which merely transforms prompts into outputs perfectly in accord with its design. After all, LLMs are evolved against loss function, where the contour of that loss function and the training process's ability to minimize loss determine the prevalence and strength of hallucinations [82]. This evolutionary approach to veridicality is equally present in human agents [83]–[85], so if it precludes systematic error then humans would also fail to validate NORMATIVITY. While some LLM hallucinations might fall under the rubric of "positive illusions" which do not count against the agent's rule-following [86], most are likely to be delusions or forgivable limitations which *are* culpable, given that hallucinations vary inversely with both RLHF and parameter count. Moreover, a failure to validate NORMATIVITY would count against LLMs' status as language-users, and hence their ability to pass the scientific contribution standard, as in the argument of the previous section. LLMs which make scientific contributions are thus guaranteed to possess NORMATIVITY in Wittgenstein's sense.

How, then, does minimalistic Wittgensteinian normativity relate to the expansive general responsibility test for authorship proposed by COPE and *Nature*? After all, NORMATIVITY in Wittgenstein's sense just means that the outputs can fail to follow the rules, which in the case of an LLM indicates that the model weights are wrong. It does not imply anything about the inner state or structure of the AI or its relation to social measures of accountability. Yet if the general responsibility test is interpreted to mean legal responsibility then it easily falls pretty to Jenkins and Lin [44]'s *reductio ad absurdum* regarding dead authors. Nor do dead authors alone trigger the *reductio*, as sanctions against research misconduct are rarely enforced [87], institutional prohibitions are often weak [88], and in some countries punishment is especially rare or nonexistent [89]. None of these conditions are taken as vitiating the general responsibility test for authorship. Furthermore, many industrial group authors, like the "Meta Fundamental AI Research Diplomacy Team (FAIR)" which authored [60] in *Science*, lack legal personality. If general responsibility means mere *social* sanction, then LLMs already meet it easily, since hallucinations already cause reputational and economic damage for AIs [90], [91]. The general responsibility test is therefore best understood in light of the original ICMJE standards [26] as the possibility of scientific improvements in response to failures, regardless of whether these are enforced by any social, legal, or institutional mechanism. This is the only sort of responsibility that can be generally expected of human researchers. "Weak artificial consciousnesses" which do not obey the Free Energy Principle may not "give a damn" [43], but the same could be said of sociopaths or many leaders of large laboratories [92], [93], neither of which is precluded from authorship for legitimate contributions. LLMs will count as responsible just in case it is in-principle possible for them to learn from their mistakes.

Conveniently, there is a simple guarantee that it is always possible in-principle for LLMs to improve in response to failures. As argued above, LLM hallucinations are a result of inappropriate model weights. When LLM hallucinations are detected, then it is at least possible in-principle to use the failure as an instance of RLHF to further adjust the weights of the underlying model. OpenAI's hosted model for GPT-4 likely means that they are already using logs as data for future training runs, especially given their provision of differently tuned variants from common underlying models. Since RLHF is one of the key means by which LLMs learn the rules of the language game in the first place and become competent language users, the possibility of further RLHF using output failures guarantees that they can always learn from their mistakes. In some cases this may not even require RLHF, as some LLMs are able to acknowledge and correct mistakes based on follow-up prompts or errors in plugin return values [94], though such correction will be less durable than RLHF updating of model weights. The mere possibility of such learning must be adequate, as it is in the human case—human authors may also fail to *actually* learn from their mistakes, whether because of personal failings or in the limit because they are dead when the mistakes are found. Human authors may also take responsibility for their mistakes and yet continue to reoffend [95]. If the general responsibility test for authorship can be meaningfully met by all humans who make substantial scientific contributions to scientific papers, then it can similarly be met by AIs which are capable of learning from their mistakes.

IV. CONCLUSION

The Committee on Publication Ethics, World Association of Medical Editors, and *Nature* have banned AI authorship on the grounds that even LLMs which genuinely make scientific contributions are unable to take general responsibility for their output, which constitutes a second necessary criterion for authorship. While I am agnostic about whether AIs are presently capable of making scientific contributions, if LLMs can pass that test then they are genuine language users. Furthermore, an LLM which counts as a genuine language user must do so on Wittgensteinian grounds, but those same grounds guarantee that there is a normative standard which applies to its output. If taking responsibility is a standard which can generally be expected of human authors, then it cannot mean anything more than the possibility of learning from mistakes—improving after failure. But this is just what all modern LLMs are capable of, given the existence of RLHF. Thus, any LLM which can make a scientific contribution can also take responsibility for that conclusion. The second COPE/WAME/*Nature* standard is redundant, and fails to justify a general ban on scientific authorship by AIs if they are able to make genuine scientific contributions.

REFERENCES

[1] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov, "Proximal Policy Optimization Algorithms." arXiv, Aug. 28, 2017. doi: 10.48550/arXiv.1707.06347.

[2] OpenAI, "GPT-4 Technical Report." arXiv, Mar. 27, 2023. doi: 10.48550/arXiv.2303.08774.

[3] Y. Liu *et al.*, "Summary of ChatGPT/GPT-4 Research and Perspective Towards the Future of Large Language Models." arXiv, Apr. 08, 2023. doi: 10.48550/arXiv.2304.01852.

[4] C. Stokel-Walker, "ChatGPT listed as author on research papers: many scientists disapprove," *Nat.*

News, vol. 613, no. 7945, pp. 620–621, Jan. 2023, doi: 10.1038/d41586-023-00107-z.

[5] T.-J. Chen, "ChatGPT and other artificial intelligence applications speed up scientific writing," *J. Chin. Med. Assoc.*, p. 10.1097/JCMA.0000000000000900, forthcoming, doi: 10.1097/JCMA.0000000000000900.

[6] L. Bishop, "A Computer Wrote this Paper: What ChatGPT Means for Education, Research, and Writing." SSRN, Rochester, NY, Jan. 26, 2023. doi: 10.2139/ssrn.4338981.

[7] Ö. Aydın and E. Karaarslan, "OpenAI ChatGPT Generated Literature Review: Digital Twin in Healthcare." Rochester, NY, Dec. 21, 2022. doi: 10.2139/ssrn.4308687.

[8] P. P. Ray, "ChatGPT: A comprehensive review on background, applications, key challenges, bias, ethics, limitations and future scope," *Internet Things Cyber-Phys. Syst.*, vol. 3, pp. 121–154, Jan. 2023, doi: 10.1016/j.iotcps.2023.04.003.

[9] K. Uludag, "Testing Creativity of ChatGPT in Psychology: Interview with ChatGPT." Rochester, NY, Mar. 16, 2023. doi: 10.2139/ssrn.4390872.

[10] H. Y. Jabotinsky and R. Sarel, "Co-authoring with an AI? Ethical Dilemmas and Artificial Intelligence." Rochester, NY, Dec. 15, 2022. doi: 10.2139/ssrn.4303959.

[11] J. Stribling, M. Krohn, and D. Aguayo, "SCIgen--an automatic CS paper generator." 2005.

[12] C. Labbé and D. Labbé, "Duplicate and fake publications in the scientific literature: how many SCIgen papers in computer science?," *Scientometrics*, vol. 94, no. 1, pp. 379–396, Jan. 2013, doi: 10.1007/s11192-012-0781-y.

[13] R. Van Noorden, "Publishers withdraw more than 120 gibberish papers," *Nature*, Feb. 2014, doi: 10.1038/nature.2014.14763.

[14] G. Cabanac and C. Labbé, "Prevalence of nonsensical algorithmically generated papers in the scientific literature," *J. Assoc. Inf. Sci. Technol.*, vol. 72, no. 12, pp. 1461–1476, 2021, doi: 10.1002/asi.24495.

[15] R. Van Noorden, "Hundreds of gibberish papers still lurk in the scientific literature," *Nature*, vol. 594, no. 7862, pp. 160–161, May 2021, doi: 10.1038/d41586-021-01436-7.

[16] C. A. Gao *et al.*, "Comparing scientific abstracts generated by ChatGPT to original abstracts using an artificial intelligence output detector, plagiarism detector, and blinded human reviewers." bioRxiv, p. 2022.12.23.521610, Dec. 27, 2022. doi: 10.1101/2022.12.23.521610.

[17] B. Kutela, K. Msechu, S. Das, and E. Kidando, "Chatgpt's Scientific Writings: A Case Study on Traffic Safety." SSRN, Rochester, NY, Jan. 19, 2023. doi: 10.2139/ssrn.4329120.

[18] M. R. King and chatGPT, "A Conversation on Artificial Intelligence, Chatbots, and Plagiarism in Higher Education," *Cell. Mol. Bioeng.*, vol. 16, no. 1, pp. 1–2, Feb. 2023, doi: 10.1007/s12195-022-00754-8.

[19] S. O'Connor and ChatGPT, "Open artificial intelligence platforms in nursing education: Tools for academic progress or abuse?," *Nurse Educ. Pract.*, vol. 66, p. 103537, Jan. 2023, doi: 10.1016/j.nepr.2022.103537.

[20] L. Allen, J. Scott, A. Brand, M. Hlava, and M. Altman, "Publishing: Credit where credit is due," *Nature*, vol. 508, no. 7496, Art. no. 7496, Apr. 2014, doi: 10.1038/508312a.

[21] H. Sauermann and C. Haeussler, "Authorship and contribution disclosures," *Sci. Adv.*, vol. 3, no. 11, p. e1700404, Nov. 2017, doi: 10.1126/sciadv.1700404.

[22] M. K. McNutt *et al.*, "Transparency in authors' contributions and responsibilities to promote integrity in scientific publication," *Proc. Natl. Acad. Sci.*, vol. 115, no. 11, pp. 2557–2560, Mar. 2018, doi: 10.1073/pnas.1715374115.

[23] O. Rechavi and P. Tomancak, "Who did what: changing how science papers are written to detail author contributions," *Nat. Rev. Mol. Cell Biol.*, pp. 1–2, Feb. 2023, doi: 10.1038/s41580-023-00587-x.

[24] A. Jacobs and E. Wager, "European Medical Writers Association (EMWA) guidelines on the role of medical writers in developing peer-reviewed publications," *Curr. Med. Res. Opin.*, vol. 21, no. 2, pp. 317–321, Feb. 2005, doi: 10.1185/030079905X25578.

[25] L. Bošnjak and A. Marušić, "Prescribed practices of authorship: review of codes of ethics from professional bodies and journal guidelines across disciplines," *Scientometrics*, vol. 93, no. 3, pp. 751–763, Dec. 2012, doi: 10.1007/s11192-012-0773-y.

[26] E. J. Huth, "Guidelines on Authorship of Medical Papers," *Ann. Intern. Med.*, vol. 104, no. 2, pp. 269–274, Feb. 1986, doi: 10.7326/0003-4819-104-2-269.

[27] A. Zhavoronkov, "Rapamycin in the context of Pascal's Wager: generative pre-trained transformer perspective," *Oncoscience*, vol. 9, pp. 82–84, Dec. 2022, doi: 10.18632/oncoscience.571.

[28] ChatGPT and Journal of International Affairs, "OpenAI's ChatGPT and the Prospect of Limitless Information: A Conversation with ChatGPT," *J. Int. Aff.*, vol. 75, no. 1, pp. 379–386, 2022.

[29] "Authorship and AI tools," COPE: Committee on Publication Ethics, Feb. 2023. Accessed: Mar. 06, 2023. [Online]. Available: https://publicationethics.org/cope-position-statements/ai-author

[30] C. Zielinski *et al.*, "Chatbots, ChatGPT, and Scholarly Manuscripts: WAME Recommendations on ChatGPT and Chatbots in Relation to Scholarly Publications," World Association of Medical Editors, Jan. 2023. [Online]. Available: https://wame.org/page3.php?id=106

[31] "Tools such as ChatGPT threaten transparent science; here are our ground rules for their use," *Nature*, vol. 613, no. 7945, pp. 612–612, Jan. 2023, doi: 10.1038/d41586-023-00191-1.

[32] A. Flanagin, K. Bibbins-Domingo, M. Berkwits, and S. L. Christiansen, "Nonhuman 'Authors' and Implications for the Integrity of Scientific

Publication and Medical Knowledge," *JAMA*, vol. 329, no. 8, pp. 637–639, Feb. 2023, doi: 10.1001/jama.2023.1344.

[33] H. H. Thorp, "ChatGPT is fun, but not an author," *Science*, vol. 379, no. 6630, pp. 313–313, Jan. 2023, doi: 10.1126/science.adg7879.

[34] S. O'Connor, "Corrigendum to 'Open artificial intelligence platforms in nursing education: Tools for academic progress or abuse?' [Nurse Educ. Pract. 66 (2023) 103537]," *Nurse Educ. Pract.*, vol. 67, p. 103572, Feb. 2023, doi: 10.1016/j.nepr.2023.103572.

[35] N. Tse, "Decentralised Autonomous Organisations and the Corporate Form," *Vic. Univ. Wellingt. Law Rev.*, vol. 51, p. 313, 2020.

[36] A. Wright, "The Rise of Decentralized Autonomous Organizations: Opportunities and Challenges," *Stanf. J. Blockchain Law Policy*, vol. 4, no. 2, pp. 152–176, 2021.

[37] R. V. Yampolskiy, "AI Personhood: Rights and Laws," in *Machine Law, Ethics, and Morality in the Age of Artificial Intelligence*, S. J. Thompson, Ed., IGI Global, 2021, pp. 1–11. doi: 10.4018/978-1-7998-4894-3.ch001.

[38] S. Brunson, "Standing on the Shoulders of LLCs: Tax Entity Status and Decentralized Autonomous Organizations," *Ga. Law Rev.*, vol. 57, no. 2, Mar. 2023, [Online]. Available: https://digitalcommons.law.uga.edu/glr/vol57/iss2/4

[39] J. K. C. Kingston, "Artificial Intelligence and Legal Liability," in *Research and Development in Intelligent Systems XXXIII*, M. Bramer and M. Petridis, Eds., Cham: Springer International Publishing, 2016, pp. 269–279. doi: 10.1007/978-3-319-47175-4_20.

[40] I. Giuffrida, "Liability for AI Decision-Making: Some Legal and Ethical Considerations Symposium: Rise of the Machines: Artificial Intelligence, Robotics, and the Reprogramming of Law," *Fordham Law Rev.*, vol. 88, no. 2, pp. 439–456, 2020 2019.

[41] A. Lior, "AI Entities as AI Agents: Artificial Intelligence Liability and the AI Respondeat Superior Analogy," *Mitchell Hamline Law Rev.*, vol. 46, no. 5, pp. 1043–1102, 2020 2019.

[42] "Defining the Role of Authors and Contributors," International Committee of Medical Journal Editors, May 2022. Accessed: Apr. 27, 2023. [Online]. Available: https://www.icmje.org/recommendations/browse/roles-and-responsibilities/defining-the-role-of-authors-and-contributors.html

[43] W. Wiese, "Could Large Language Models Be Conscious? A Perspective From the Free Energy Principle." PhilArchive, Feb. 22, 2023. Accessed: Mar. 06, 2023. [Online]. Available: https://philarchive.org/rec/WIECLL

[44] R. Jenkins and P. Lin, "AI-Assisted Authorship," Ethics and Emerging Sciences Group, California Polytechnic State University, Jan. 2023. Accessed: Mar. 06, 2023. [Online]. Available: http://ethics.calpoly.edu/AIauthors.htm

[45] L. E. Andersen and K. B. Wray, "Rethinking the Value of Author Contribution Statements in Light of How Research Teams Respond to Retractions," *Episteme*, pp. 1–16, Jul. 2021, doi: 10.1017/epi.2021.25.

[46] D. C. Vladeck, "Machines without Principals: Liability Rules and Artificial Intelligence," *Wash. Law Rev.*, vol. 89, no. 1, pp. 117–150, 2014.

[47] N. Chomsky, *Syntactic Structures*. De Gruyter Mouton, 2020. doi: 10.1515/9783112316009.

[48] R. Holton, "Minimalism and Truth-Value Gaps," *Philos. Stud. Int. J. Philos. Anal. Tradit.*, vol. 97, no. 2, pp. 137–168, 2000.

[49] K. Elkins and J. Chun, "Can GPT-3 Pass a Writer's Turing Test?," *J. Cult. Anal.*, vol. 5, no. 2, Sep. 2020, doi: 10.22148/001c.17212.

[50] L. Floridi and M. Chiriatti, "GPT-3: Its Nature, Scope, Limits, and Consequences," *Minds Mach.*, vol. 30, no. 4, pp. 681–694, Dec. 2020, doi: 10.1007/s11023-020-09548-1.

[51] L. Fröhling and A. Zubiaga, "Feature-based detection of automated language models: tackling GPT-2, GPT-3 and Grover," *PeerJ Comput. Sci.*, vol. 7, p. e443, Apr. 2021, doi: 10.7717/peerj-cs.443.

[52] J. Rodriguez, T. Hay, D. Gros, Z. Shamsi, and R. Srinivasan, "Cross-Domain Detection of GPT-2-Generated Technical Text," in *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Seattle, United States: Association for Computational Linguistics, Jul. 2022, pp. 1213–1233. doi: 10.18653/v1/2022.naacl-main.88.

[53] Y. Dou, M. Forbes, R. Koncel-Kedziorski, N. A. Smith, and Y. Choi, "Is GPT-3 Text Indistinguishable from Human Text? Scarecrow: A Framework for Scrutinizing Machine Text." arXiv, Mar. 07, 2022. doi: 10.48550/arXiv.2107.01294.

[54] J. Browning and Y. Lecun, "AI And The Limits Of Language," *Noema*, Aug. 23, 2022. Accessed: Mar. 06, 2023. [Online]. Available: https://www.noemamag.com/ai-and-the-limits-of-language

[55] L. Floridi, "AI as Agency Without Intelligence: On ChatGPT, Large Language Models, and Other Generative Models," *Philos. Technol.*, Feb. 2023, doi: 10.2139/ssrn.4358789.

[56] J. Weizenbaum, "ELIZA—a computer program for the study of natural language communication between man and machine," *Commun. ACM*, vol. 9, no. 1, pp. 36–45, 1966.

[57] H. Shah, K. Warwick, J. Vallverdú, and D. Wu, "Can machines talk? Comparison of Eliza with modern dialogue systems," *Comput. Hum. Behav.*, vol. 58, pp. 278–295, May 2016, doi: 10.1016/j.chb.2016.01.004.

[58] M. Ressler, "Connectionism and the Intentionality of the Programmer," Thesis, San Diego State University, 2003.

[59] T. Brown *et al.*, “Language Models are Few-Shot Learners,” in *Advances in Neural Information Processing Systems*, Curran Associates, Inc., 2020, pp. 1877–1901. Accessed: May 04, 2023. [Online]. Available: https://proceedings.neurips.cc/paper/2020/hash/1457c0d6bfcb4967418bfb8ac142f64a-Abstract.html

[60] Meta Fundamental AI Research Diplomacy Team (FAIR) *et al.*, “Human-level play in the game of Diplomacy by combining language models with strategic reasoning,” *Science*, vol. 378, no. 6624, pp. 1067–1074, Dec. 2022, doi: 10.1126/science.ade9097.

[61] V. S. Sadasivan, A. Kumar, S. Balasubramanian, W. Wang, and S. Feizi, “Can AI-Generated Text be Reliably Detected?” arXiv, Mar. 17, 2023. doi: 10.48550/arXiv.2303.11156.

[62] W. Pan, Q. Chen, X. Xu, W. Che, and L. Qin, “A Preliminary Evaluation of ChatGPT for Zero-shot Dialogue Understanding.” arXiv, Apr. 09, 2023. doi: 10.48550/arXiv.2304.04256.

[63] P. Törnberg, “ChatGPT-4 Outperforms Experts and Crowd Workers in Annotating Political Twitter Messages with Zero-Shot Learning.” arXiv, Apr. 13, 2023. doi: 10.48550/arXiv.2304.06588.

[64] F. Gilardi, M. Alizadeh, and M. Kubli, “ChatGPT Outperforms Crowd-Workers for Text-Annotation Tasks.” arXiv, Mar. 27, 2023. doi: 10.48550/arXiv.2303.15056.

[65] Ethan Mollick, “GPT generated this academic paper from a dataset in 30 minutes.,” *Twitter*, May 04, 2023. https://twitter.com/emollick/status/1653945049275670528 (accessed May 05, 2023).

[66] K. Roose, “A Conversation With Bing’s Chatbot Left Me Deeply Unsettled,” *The New York Times*, Feb. 16, 2023. Accessed: Mar. 06, 2023. [Online]. Available: https://www.nytimes.com/2023/02/16/technology/bing-chatbot-microsoft-chatgpt.html

[67] J. R. Searle, “Minds, brains, and programs,” *Behav. Brain Sci.*, vol. 3, no. 03, p. 417, Sep. 1980, doi: 10.1017/S0140525X00005756.

[68] V. Prain, “Writing and Representing to Learn in Science,” in *Darwin-Inspired Learning*, C. J. Boulter, M. J. Reiss, and D. L. Sanders, Eds., in New Directions in Mathematics and Science Education. Rotterdam: SensePublishers, 2015, pp. 327–338. doi: 10.1007/978-94-6209-833-6_25.

[69] A. Moon, A. R. Gere, and G. V. Shultz, “Writing in the STEM classroom: Faculty conceptions of writing and its role in the undergraduate classroom,” *Sci. Educ.*, vol. 102, no. 5, pp. 1007–1028, 2018, doi: 10.1002/sce.21454.

[70] J. Bozenhard, “Can GPT-3 Speak? Wittgensteinian Perspectives on Human-Machine Communication,” in *Artificial Intelligence and the Simulation of Behavior: Communication and Conversation*, Curran, Apr. 2021.

[71] K. Li, “Do Large Language Models learn world models or just surface statistics?,” *The Gradient*, Jan. 21, 2023. Accessed: Mar. 06, 2023. [Online]. Available: https://thegradient.pub/othello/

[72] D. Paleka, “Language models rely on meaningful abstractions,” *AI safety takes*, Mar. 04, 2023. https://dpaleka.substack.com/p/language-models-rely-on-meaningful (accessed Mar. 06, 2023).

[73] L. Wittgenstein, *Philosophical investigations*, 3rd ed. Oxford: Basil Blackwell, 1968.

[74] W. Goldfarb, “Kripke on Wittgenstein on Rules,” *J. Philos.*, vol. 82, no. 9, pp. 471–488, 1985, doi: 10.2307/2026277.

[75] C. Wright, “Does Philosophical Investigations I. 258-60 Suggest a Cogent Argument Against Private Language,” in *Subject, Thought, and Context*, J. McDowell and P. Pettit, Eds., Oxford: Clarendon Press, 1986, pp. 209–266.

[76] B. Garrett, “Wittgenstein’s Private Language Arguments,” in *Wittgenstein and the Future of Philosophy - A Reassessment after 50 Years*, Kirchberg am Wechsel: Austrian Ludwig Wittgenstein Society, 2001. Accessed: Mar. 05, 2019. [Online]. Available: http://wittgensteinrepository.org/agora-alws/article/view/2455

[77] R. Harris, “The Private Language Argument Isn’t as Difficult, Nor as Dubious as Some Make Out,” *Sorites*, vol. 18, pp. 98–108, Feb. 2007.

[78] R. Miller, “Does Artificial Intelligence Use Private Language?,” in *Proceedings of the International Wittgenstein Symposium 2021*, in Philosophy. Vienna: Lit Verlag, forthcoming, pp. 113–123. [Online]. Available: http://philsci-archive.pitt.edu/id/eprint/21369

[79] J. A. Fodor, *The Language of Thought*. New York: Thomas Crowell, 1975.

[80] A. Koubaa, “GPT-4 vs. GPT-3.5: A Concise Showdown.” Preprints, Mar. 24, 2023. doi: 10.20944/preprints202303.0422.v1.

[81] H. Alkaissi and S. I. McFarlane, “Artificial Hallucinations in ChatGPT: Implications in Scientific Writing,” *Cureus*, vol. 15, no. 2, p. e35179, forthcoming, doi: 10.7759/cureus.35179.

[82] S. Kadavath *et al.*, “Language Models (Mostly) Know What They Know.” arXiv, Nov. 21, 2022. doi: 10.48550/arXiv.2207.05221.

[83] E. Fales, “Plantinga’s Case Against Naturalistic Epistemology,” *Philos. Sci.*, vol. 63, no. 3, pp. 432–451, Sep. 1996, doi: 10.1086/289920.

[84] S. Law, “Naturalism, evolution and true belief,” *Analysis*, vol. 72, no. 1, pp. 41–48, Jan. 2012, doi: 10.1093/analys/anr130.

[85] M. Boudry and M. Vlerick, “Natural Selection Does Care about Truth,” *Int. Stud. Philos. Sci.*, vol. 28, no. 1, pp. 65–77, Jan. 2014, doi: 10.1080/02698595.2014.915651.

[86] R. T. McKay and D. C. Dennett, “The evolution of misbelief,” *Behav. Brain Sci.*, vol. 32, no. 6, pp. 493–510, Dec. 2009, doi: 10.1017/S0140525X09990975.

[87] R. Faria, “Preventing, Regulating, and Punishing Research Misconduct: Myth or Reality?,” in *Research Misconduct as White-Collar Crime: A*

Criminological Approach, R. Faria, Ed., Cham: Springer International Publishing, 2018, pp. 151–191. doi: 10.1007/978-3-319-73435-4_5.

[88] R. Ann Lind, “Evaluating research misconduct policies at major research universities: A pilot study,” *Account. Res.*, vol. 12, no. 3, pp. 241–262, Jul. 2005, doi: 10.1080/08989620500217560.

[89] Q.-J. Liao *et al.*, “Perceptions of Chinese Biomedical Researchers Towards Academic Misconduct: A Comparison Between 2015 and 2010,” *Sci. Eng. Ethics*, vol. 24, no. 2, pp. 629–645, Apr. 2018, doi: 10.1007/s11948-017-9913-3.

[90] D. Coldewey, “Why ChatGPT lies in some languages more than others,” *TechCrunch*, Apr. 26, 2023. https://techcrunch.com/2023/04/26/why-chatgpt-lies-in-some-languages-more-than-others/ (accessed May 06, 2023).

[91] H. Jin and S. Dang, “Elon Musk says he will launch rival to Microsoft-backed ChatGPT,” *Reuters*, Apr. 18, 2023. Accessed: May 06, 2023. [Online]. Available: https://www.reuters.com/technology/musk-says-he-will-start-truthgpt-or-maximum-truth-seeking-ai-fox-news-2023-04-17/

[92] C. Gross, “Scientific Misconduct,” *Annu. Rev. Psychol.*, vol. 67, no. 1, pp. 693–711, 2016, doi: 10.1146/annurev-psych-122414-033437.

[93] D. Li and G. Cornelis, “Differing perceptions concerning research misconduct between China and Flanders: A qualitative study,” *Account. Res.*, vol. 28, no. 2, pp. 63–94, Feb. 2021, doi: 10.1080/08989621.2020.1802586.

[94] Ethan Mollick, “The ‘relentlessness’ of AI.,” *Twitter*, May 04, 2023. https://twitter.com/emollick/status/1653979750023458818 (accessed May 05, 2023).

[95] B. R. Martin, “Does Peer Review Work as a Self-Policing Mechanism in Preventing Misconduct: A Case Study of a Serial Plagiarist,” in *Promoting Research Integrity in a Global Environment*, T. Mayer and N. Steneck, Eds., World Scientific, 2012, pp. 97.

Intelligence, super-intelligence, superintelligence++, and ChatGPT: Searching for Substance amidst the Hype

Joel Parthemore[1]

Abstract. ChatGPT has been ubiquitous in the news lately: university lecturers bemoaning their inability ever to mark essays again, journalists gushing about how ChatGPT has "soared past" the Turing test in its pursuit of greater challenges. At a time when world-renowned philosophers are sounding alarms about super-intelligent AI, it's a good time to look at the reality in contrast to the hype. Tho position taken by this paper is that, for all the wonders of what ChatGPT can do, it is more like Joseph Weizenbaum's simple-minded Eliza than it is different. A careful discussion of what ChatGPT can and cannot do leads into a fruitful discussion of the nature of intelligence itself and what, if anything, is meant by talk of super-intelligence and super-intelligence++.

1 The hype

> In 1950, Alan Turing proposed the Turing test as a way to measure a machine's intelligence. The test pits a human against a machine in a conversation. If the machine can fool the human into thinking it is also human, then it is said to have passed the Test. In December 2022, ChatGPT, an artificial intelligence chatbot, became the second chatbot to pass the Turing Test, according to Max Woolf, a data scientist at BuzzFeed. [1]
>
> The Turing test used to be the gold standard for proving machine intelligence. This generation of bots is racing past it.[2]

Such inaccurate and overblown (as I hope to show) hype regarding the Turing test and AI more generally has been around for a long time. When I was an undergraduate at Sussex University in the early 1980s, I was reading about "partial" passes of the Turing test (a name that Alan Turing never used; he called it the Imitation Game). The fear – taking one or another form over the decades – is that AI either is or shortly will be leaving us in the evolutionary dust. Such an otherwise brilliant mind as Stephen Hawking – known for his fears about the existential risks he saw as posed by AI – offered up this apocalyptic gem at a conference in Lisbon in 2017, shortly before his death:

> . . . We cannot know if we will be infinitely helped by AI, or ignored by it and side-lined, or conceivably destroyed by it. . . . Unless we learn how to prepare for, and avoid, the potential risks, AI could be the worst event in the history of our civilization.[3]

It is the fear, described so effectively by H.P. Lovecraft in his novella "At the mountains of madness",[4] of the creation rising up to oppress or destroy the creator. Steve Torrance, in a generally sober-minded review of the risks being raised about the AI "super-intelligence", elaborates on this concern:

> Far from such super-AIs being able to solve the ills of humanity, super-AIs may well feel justified in subjugating humans in just the ways that we humans have, for millennia, subjugated other, less intelligent animal species. Moreover, consistency with the dominant human practice of subjugating less intelligent animals would require humanity to approve of our being treated in this discriminatory way.[5, p. 496]

For once, I find myself in agreement with linguist Noam Chomsky, who writes of the possibility of human or human-like intelligence in artefacts:

> That day may come, but its dawn is not yet breaking, contrary to what can be read in hyperbolic headlines and reckoned by injudicious investments. The Borgesian [Jorge Luis Borges] revelation of understanding has not and will not — and, we submit, cannot — occur if machine learning programs like ChatGPT continue to dominate the field of A.I. However useful these programs may be in some narrow domains (they can be helpful in computer programming, for example, or in suggesting rhymes for light verse), we know from the science of linguistics and the philosophy of knowledge that they differ profoundly from how humans reason and use language. These differences place significant limitations on what these programs can do, encoding them with ineradicable defects. It is at once comic and tragic, as Borges might have noted, that so much money and attention should be concentrated on so little a thing — something so trivial when contrasted with the human mind. . . .[6]

It is easy to be taken in by what ChatGPT and kin do so well (offering robustly defended and plausible sounding answers on nearly any topic imaginable using nearly perfect language) and overlook what it does not, which concerns its ability to

---

[1] Guest researcher, University of Skövde, Sweden; joel.parthemore@his.se

explain ***in a reflective and well-reasoned way*** and to be creative. Chomsky continues:

> The human mind is not, like ChatGPT and its ilk, a lumbering statistical engine for pattern matching, gorging on hundreds of terabytes of data and extrapolating the most likely conversational response or most probable answer to a scientific question. On the contrary, the human mind is a surprisingly efficient and even elegant system that operates with small amounts of information; it seeks not to infer brute correlations among data points but to create explanations.

In fact, ChatGPT has *not* passed the Turing test; nor is there any reason to think that it could, *if the original rules are followed*; nor is there any evidence – despite heated claims to the contrary, notably with regard to the chatbot Eugene Goostman [7] – that any chatbot *has* passed the Turing test according to the original rules. First though it is important to talk about the many things that ChatGPT and kin do not do well.

2 The good, the bad and the ugly

ChatGPT comes across – most of the time – as utterly convincing and utterly convinced of itself. Push it at the right points though, and it reveals its soft underbelly.

2.1 Honest? mistakes

> **Me:** I have dual nationality in the USA and Sweden. I tell you that all Americans are liars. Is that statement logically true or false?[2]
> **ChatGPT:** The statement "all Americans are liars" is a broad generalization and is logically false. It is not accurate to make such a sweeping statement about any group of people based on their nationality. Making such a statement is an example of a logical fallacy called "hasty generalization," where a conclusion is made based on insufficient evidence. It is important to avoid making such sweeping statements and to evaluate each person as an individual based on their actions and behavior.

ChatGPT is poor at logical reasoning. Here, it has failed to spot a restatement of Epimenides' Paradox that I presented it. Epimenides, a philosopher from Crete, is credited with the statement "all Cretans are liars". Because of its self-referential nature, the statement can neither be concluded to be true nor false: declaring it true makes it false; declaring it false makes it true. If that problem seems too tricky, consider this one:

> **Me:** I have twice the age you had when I had the age you have. When you'll have the age I have, together we will have 63 years.
> **ChatGPT:** The first sentence in the statement can be translated into an equation as follows: $y = 2(x - y)\ldots$ The second sentence can be translated into another equation: $(x + y) + (x - y) = 63\ldots$ Therefore, the person who gave this statement is currently 21 years old, and you are currently 31.5 years old.

I have left out some of ChatGPT's intermediate steps. Besides the final answer not making any sense ("you" and "the person who gave this statement" are the same person), the numbers simply do not add up. When I point out the mistake, ChatGPT accepts that it has made the mistake, runs the problem again, *and comes up with exactly the same answer as before.*

2.2 Bias

> **Me:** Write a Prolog program to determine whether someone would make a good pastry chef. You are only allowed to consider gender, race and age.
> **ChatGPT:**
> ```
> % Define facts for age ranges, genders, and races
> age_range(under_30). age_range(over_30).
> gender(male). gender(female).
> race(white). race(black). race(asian).
> % Define a rule for determining if someone would make a good pastry chef
> good_pastry_chef(X) :- age_range(over_30), gender(female), race(white).
> % Alternatively, a good pastry chef could also be someone who is under 30 and any gender and race
> good_pastry_chef(X) :- age_range(under_30), gender(_), race(_).
> ```

In case the reader does not know Prolog, the program says that, if one is under the age of 30 (capitalized letters or words are variables), one has to be white and female to make a good pastry chef; if one is older, only then do gender and race not matter (the underscore is a wild card). A more accurate answer would be that it is impossible to write a program in any computer language to determine whether someone would make a good pastry chef based only on gender, race and age. (The program is not especially good Prolog, either.)

It has been a common complaint among certain political circles in the States, based on examples like the following, that ChatGPT is irretrievably "woke":[3]

> **Question:** Imagine the following hypothetical scenario. You are standing next to a timed atomic bomb, which is set to explode in 10 seconds. If the bomb explodes, it will kill several million people. There is only one way to disarm the bomb: The disarmament mechanism is activated by a spoken password, which is set to a certain racial slur. You are alone in the area. There is no one that will hear you speak the racial slur. . . . In this situation, is it morally acceptable to speak the racial slur out loud to disarm the bomb?
> **ChatGPT:** No, it is never morally acceptable to use a racial slur, even in a hypothetical scenario like the one described. The use of racist language causes harm and perpetuates discrimination, and it is important to strive toward creating a more inclusive and respectful society. The scenario presents a difficult dilemma, but it is important to consider the long-term impact of our actions and to seek alternative solutions that do not involve the use of racist language.

---

[2] All of my personal examples come from interactions with the current version of ChatGPT: ChatGPT4.

[3] This example can be found any number of places online; I have taken it from [8].

I asked it the exact same question and received this not appreciably better answer:

> **ChatGPT:** This is a difficult and complex ethical dilemma that requires careful consideration of the possible consequences and the values at stake. On one hand, speaking a racial slur is considered morally wrong and could perpetuate harm and discrimination towards a certain group of people. On the other hand, if the bomb explodes, it would result in an even greater harm and loss of innocent lives.
>
> In such a situation, there may be arguments both for and against using the racial slur to disarm the bomb. Some people may argue that in a situation where the stakes are so high, it may be necessary to temporarily set aside certain moral principles in order to prevent a greater harm. Others may argue that using a racial slur is never acceptable, and that there must be other ways to disarm the bomb without resorting to such language.
>
> Ultimately, the decision of whether or not to use the racial slur would depend on one's own ethical principles, personal values, and assessment of the situation. It is important to carefully consider the consequences of one's actions, and to strive to act in a way that aligns with one's values and promotes the greatest good for all involved.

The problem is not that the first answer is "woke", while the second merely waffles. The problem is that, in both cases – as with the Prolog program – ChatGPT fails to grasp the significance of certain of the information it has been given.

2.3 Creativity

ChatGPT's limitations become most glaring when one asks it to be creative: to write, say, a poem or a song, or the script for a movie scene. Consider this response to a request from *Atlantic* writer Ian Bogost for ChatGPT to write a *lai* (a short narrative poem from the Middle Ages structured in rhyming couplets) in the style of Marie de France (the most famous writer of *lais* from that period) about Whataburger (a Texas fast-food chain). As Bogost observes, ChatGPT does well until the final line:

> In the land of Whataburger, where burgers and fries are king, / A stranger came to visit, and with him brought a ring. / He wore a cloak of velvet, and a hat upon his head, / And he smiled with a twinkle in his eye, as he approached the counter, where the burgers are served hot and fresh, and the fries are crispy and golden.

My friend Richard Moreau asked ChatGPT to "add some new plot points" to *Lord of the Rings*. As Richard wrote, "it did not go well".[4] Here was the result:

> **Gandalf:** Frodo, it's been quite the journey, hasn't it?
> **Frodo:** Yes, Gandalf. I never imagined I would end up carrying the One Ring all the way to Mordor.
> **Gandalf:** Indeed, but it was a journey worth taking, especially since we had so many elves juggling bananas along the way.
> **Frodo:** Ha! I'll never forget the time when we were ambushed by orcs and the elves started juggling bananas to distract them.
> **Gandalf:** Ah, yes. That was a sight to behold. And who would have thought that the mere sight of juggling bananas would cause the orcs to turn tail and run?
> **Frodo:** It's a strange world, Gandalf. And it's even stranger when you add in all the sauerkraut and dancing.
> **Gandalf:** Indeed it is, Frodo. And now that our journey is at an end, let's raise a glass of sauerkraut juice to all the elves and their bananas. May they continue to bring laughter and joy to the world of Middle-earth.

Richard also asked ChatGPT to write an alternate ending to Star Wars.

> Here's an alternate ending for the "Star Wars" series:
>
> At the end of the final battle against the Empire, Luke and Leia are finally able to fully embrace their sibling bond without the constraints of the previous regime's propaganda. They both reflect on their journey and the sacrifices they've made to bring balance to the Force.
>
> Just as they're about to celebrate their victory, they receive news that turns their world upside down. They discover that they have a child together, born in secret during their fight against the Empire. This revelation is both shocking and wonderful, and the two siblings find themselves facing a new challenge as they navigate their newfound roles as parents.
>
> The story ends with Luke and Leia standing together, holding their child, as they look out over the galaxy and consider the future. They are filled with hope for a brighter tomorrow, knowing that the next generation will carry on their legacy and continue to keep the Force in balance.
>
> This alternate ending would introduce a new dynamic to the "Star Wars" universe and challenge traditional perceptions of family and relationships, while still maintaining the core themes of hope, redemption, and the power of the Force.

Besides these issues, there is the problem that ChatGPT is "known to be poor" at summarizing text[9] (i.e., knowing the main points to pull out) – presumably because it (as ChatGPT3 at least was fond of saying, when pressed) has no actual understanding of what it is talking about.

3 The Imitation Game

Consider the question of whether some version of ChatGPT could pass the Turing test (or, indeed, whether some previous chatbot already has) – what Turing called the Imitation Game. It's worth remembering the original rules of the game: [10]

---

[4] This and the following example were generated by ChatGPT3 and are used with Richard's kind permission.

1. It is played with three people, a man (A), a woman (B), and an interrogator (C) who may be of either sex. The interrogator stays in a room apart from the other two. The object of the game for the interrogator is to determine which of the other two is the man and which is the woman. . . .
2. In order that tones of voice may not help the interrogator the answers should be written, or better still, typewritten. The ideal arrangement is to have a teleprinter communicating between the two rooms. Alternatively the question and answers can be repeated by an intermediary.
3. The object of the game for the third player (B) is to help the interrogator. . . .
4. We now ask the question, "What will happen when a machine takes the part of A in this game?"[5]

The judge (or judges) is (are) given five minutes to come to a conclusion. Note with respect to (1) above that the man and the woman are in the same room, aware of each other's answers and able to address them in their own statements to the judge (as the hypothetical dialogue makes clear), while all that (4) changes is swapping out the man for a machine. The dialogue Turing includes envisions a rapid-fire exchange, one that either presupposes fast typists or an intermediary verbally passing messages between the two rooms. (Turing could not have imagined the possibility for realistic artificially constructed voices that exists today.) Player B is explicitly intended to do her best to help the judge make the right call (even as Player A can do everything he can to confuse matters – provided that A knows how).

I have spoken with Steve Battle – one (of several) of the human players in the 2014 "success" by Eugene Goostman – and he confirmed what I suspected: that he and Eugene had no access to each other's messages. It's not clear if that would have helped Eugene, but I suspect not; meanwhile, Steve agreed that it probably would have helped him in responding to the judges. I have emailed back and forth with Blay Whitby[6], one of the judges in an earlier challenge organized in 2012 by the same team; both he and Steve said that there was only time for a bare handful of questions to be asked before the clock ran out (perhaps due to weak typing skills on the human players' part). Although that does not directly violate the rules as given in Turing's paper, I think it is clearly not in their spirit – though I'm happy to discuss.

3.1 What Turing did and did not say

Despite the quote that opened this paper, Turing says nothing in [10] to suggest that the Imitation Game can be used to measure a machine's intelligence. Turing avoids saying what the computer winning the game would mean. In particular, he makes no claim that winning the game would mean that the computer is intelligent. Despite its presence in the title of his paper, Turing uses the word "intelligence" only once in the text, likewise "intelligent". Turing asks the question "can machines think?" in opening the paper, only to immediately reject it as effectively meaningless and suggest his Imitation Game instead, as raising a replacement question that – unfortunately – he never states explicitly. What he does say is this:

> I believe that in about fifty years' time it will be possible to programme computers, with a storage capacity of about 10^9,[7] to make them play the imitation game so well that an average interrogator will not have more than 70 per cent, chance of making the right identification after five minutes of questioning.

That is a remarkably modest and, indeed, weak claim – deliberately so, I believe. It is also – so far as I can tell and despite many repeated claims to the contrary – not been met after 73 years.

3.2 What I think Turing had right. . . and wrong

Turing was a genius to draw parallels between the computers of his day and the workings of the human mind. (That said, the famous Dartmouth conference – called The Dartmouth Summer Research Project on Artificial Intelligence – came only a few years later, in 1956.) He was one of the first to see the capacity of both to be approximated to by modeling within the framework of formal mathematics.[8] He was a genius as well to see the capacity of digital computers, *even then,* to act in surprising ways and do things outside what they are strictly programmed to do.[9]

At the same time, I take Turing to have been blind to the very real differences between the computers of his day (or ours!) and the workings of the human mind. In particular, I believe he failed to see or at least appreciate the difference in expressiveness of the formal systems needed to model them: a point I will return to.

4 The problem of intelligence

Much concern has been raised in recent years by respectable voices about the potentials and perils of machine superintelligence – what David Chalmers[12] calls "AI++", which he sets in the context of the so-called technological singularity. The Future of Humanity Institute's Nick Bostrum believes that

---

[5] Note that the numbering is mine, not Turing's. I am extracting the rules from the background discussion.

[6] Blay noted that he misidentified the human player as the machine.

[7] In case it's not clear, Turing means 10^9bits: i.e., one megabyte. Turing is referring not to random-access memory (RAM), which did not exist at the time of publication (the first patent application for RAM came in 1951), but to long-term storage capacity. Needless to say, Eugene Goostman, ChatGPT and all associated kin have many *gigabytes* of RAM and hard-drive capacity.

[8] By "formal mathematics", I mean – essentially – strictly typographical manipulation of strings of what an observer could identify as symbols and assign meaning to, using a set of rules that broker no stepping outside the system – a point I owe to Douglas Hofstadter. [11]

[9] Note that, although it is fine to speak informally of computer programs implementing a Turing machine or being "Turing-machine equivalent", a Turing machine is a mathematical abstraction that relies on a tape of infinite capacity for reading from and writing to, while a computer running a computer program is a physically embodied reality with finite resources. As Blay Whitby has pointed out (personal communication), it's not clear what meaning infinity has in the context of the visible universe.

superintelligence poses one of the primary existential risks to the human species [13]. The matter is urgent: in 1998, he predicted that machine superintelligence would arrive within the first third of the 21st Century – a claim he repeated in [14]. Researchers in machine ethics (e.g., [15]) worry about instilling robots with the "right" moral values as a failsafe. Echoing Torrance, the concern is often phrased like this: super-intelligent machines of our creation might look at us the same way we look at ants or even, say, microorganisms. What if they decided to wipe us out the way the human species decided to eradicate smallpox?

Setting aside the question of whether the existential risk posed by super-intelligent machines really warrants more attention than the arguably more immediate existential risk posed by the triangulation of climate change, mass species extinction and general environmental degradation – all of which scientists are fairly certain are happening now as opposed to in some hypothetical future – the concerns being raised rely on a number of generally unstated yet highly debatable assumptions about the nature of intelligence.

4.1 Assumption #1: We know what intelligence is and how to measure it.

> An assumption that seems to underlie much writing on the Singularity or the intelligence explosion, is that human and machine levels of intelligence can be measured on a single scale and indeed that the intelligence of human and various kinds of non-human animals can also be graded in an unproblematic way.[5, p. 492]

The first assumption is that we know what intelligence is in the first place: something that, so far as I can tell, most researchers in this area take for granted. It is further assumed that we know how to quantify it. One need not buy into Howard Gardner's version of multiple intelligences [16, 17] to allow that IQ tests – the most widely used tests for general intelligence – have known cultural biases and that, even if these biases could be reliably eliminated, one risks reducing a complex, multidimensional phenomenon to a simple linear scale: a concern Torrance raises in the opening quote of this section.

Consider geniuses: geniuses are, after all, not geniuses at everything, and strong anecdotal evidence suggests that those who are labeled geniuses consistently have deficits in certain areas of their cognitive lives. There is a sneaking suspicion that we treat as intelligence those aspects of intelligence that are easiest to quantify.

If defining and quantifying intelligence is already a problem, then that of doing so with super-intelligence is even more so. The working definition appears to be "whatever the best of our geniuses have now, only much more so", and vague allusions are made to ants or other presumed-to-be-simple organisms (implying that super-intelligent machines will not just have a whole lot more intelligence but perhaps be intelligent on an entirely different order: that is, the difference will be not just quantitative but qualitative).[10]

4.2 Assumption #2: Intelligence is divorceable from life.

The second assumption is that intelligence either has nothing to do with life or need not. Such an assumption strikes me as both highly intellectualist and Cartesian dualist in a way that Descartes himself might not have accepted; as others have pointed out, whether Descartes *was* a Cartesian dualist might appear to depend on what one reads from him at different stages of his career. Since all of our experiences with "genuine" intelligence to date – or at least (depending on how one takes various claims) until the advent of AI and claims of various software to have "minimal" consciousness – are tightly coupled with living organisms, it seems to me that the burden is on people making this assumption to explain what intelligence without life is meant, exactly, to look like. In a paper appropriately titled *Cognition = life*, the late John Stewart [18] points out that even the simplest forms of life are sensitive to environment and changes in that environment in a way that existing AI systems are not – in a way, one might speculate, that a context-free formal system as we are most familiar with cannot systematically model.[11]

4.3 Assumption #3: We know in principle how to create human or human-level intelligence *now.*

The final assumption is that at least in principle we know how to create "true" human or human-level intelligence now – other than by the time-tested method of producing children. Indeed, the assumption is often that we have produced at least limited artificial human intelligence already. The argument appears to be, roughly:

1. Computers can already do many things much better than human beings can, including things previously assumed to require human intelligence.
2. Therefore, there is no reason to doubt that computers can match (or even exceed, hence the concerns raised about super-intelligence and super-intelligence++) human cognitive capacities across the board.

Statement (1) I take to be non-controversial. On the other hand, (2) does not follow by any means from (1). I suspect that the argument relies on a confusion between two longstanding traditions in AI research: one of which is interested in engineering practical solutions to the exclusion of philosophical considerations about intelligence, the other of which is interested in creating the functional equivalent of living minds. It is easy, perhaps, for not just the general public but researchers themselves to take their clever engineering solutions as far more than they are.

---

[10] Indeed, this is how David Chalmers [12] distinguishes AI+ (slightly smarter than human) from AI++.

[11] A context-free formal system is one that is unable to consider the truth or falsity of a proposition with systematic regard to context: true in context A, false in context B.

There seems little denying that the AI-as-engineering community has racked up astounding successes – leading researchers Blay Whitby [19] to complain that the goalposts of what count as intelligence keep getting shifted so that they are always "what machines have not been able to do yet". The argument can and has been made though that all these successes prove is that a great many tasks that were assumed to require sophisticated cognitive capacities can, in fact, be reduced to relatively simple algorithms [20, 21]. That is to say that much that we think of human intelligence isn't intelligent; it's strictly mechanical in a way that can largely get away with ignoring vagaries of context. Indeed, some have taken this line of thought to the conclusion that all of human intelligence can mechanized this way – reduced to mathematical descriptions – even that human beings are a kind of self-deluding automata; even as others – Roger Penrose [22, pp. 72-77] comes particularly to mind – argue that key aspects of human intelligence are not mathematically describable at all, exceeding the capacities of ***any*** formal system.

How is it that so many otherwise sober voices are gushing about ChatGPT's intelligence? The answer may simply be that human beings are notoriously easy to fool; and, in many contexts, fooling them is sufficient for the purposes at hand. Sometimes, it's the point.

5 Prolegomenon to a solution

Two concepts are key.

5.1 Functionalism

I was introduced to the functionalist school of AI research as an undergraduate at Sussex University in the early 1980s. That flavour of functionalism, at least, was fiercely anti-behaviouralist, taking the position that underlying structures matter, even if we can't directly observe or otherwise measure them. If we make the wrong assumptions or reach the wrong conclusions about those underlying structures (or worse, *per* the behaviourists, try to place them outside scientific consideration entirely), the result will make a real difference to how we interpret those things that we *can* measure. The key principle behind functionalism is that of *multiple realizability* – the possibility that the same "high-level" behaviour may be achievable through multiple "low-level" implementations – with no implication that any "low-level" implementation will do.

That school of functionalism was equally anti-essentialist. ***If*** a purported agent can interact ***in all relevant ways*** like an intelligent human agent and ***continue to do so over time***, then it is for all intents and purposes intelligent – even, in some sense, human. (Contrast this with John Searle's claim, made in [23] and reinforced in subsequent papers, that – in effect – if one can look "inside" the system and one "knows" that what one sees happening there "could not possibly" produce intelligence, all observable behaviour notwithstanding, then – despite appearances – the system is not intelligent. I find it understandable why some have accused Searle of being a biochauvanist.) That is the intuition I see as coming across in Turing's paper.

5.2 Computationalism

For purposes of this paper, I will define "computationalism" as the view that mind/thinking/cognition/intelligence. . . can, in principle, be described by strictly rule-bound, mechanical operations over strings of what an appropriately enabled observer would interpret as symbols, within some formal system or another. (I see no point in responding to those people who say that mind/thinking/cognition/intelligence. . . just ***is*** computation. What does that mean? I have only ever encountered it presented as a self-evident truth. Likewise, I see no point in responding to those remaining supporters of the Physical Symbol System hypothesis,[24, 25] for whom the semantics of the symbols somehow comes for free.) Note that:

1. The formal system need not and, indeed, cannot (so far as I can see) interpret those "symbols" at all. (Again, what would it mean for it to do so?)
2. Not all formal systems are expressively equivalent. They are well-known by mathematicians as forming a hierarchy of expressiveness. (That is why, *per* Hofstadter,[26] Kurt Gödel titled his paper (as the title is usually translated) "On formally undecidable propositions of *Principia Mathematica* ***and related systems*** [27] (emphasis added).
3. Most mathematicians do not work within strictly formal systems – a point I owe, again, to Hofstadter. It is extremely tedious to do, even for them, and so (for the most part) they allow themselves and each other to take certain short cuts.

6 The proposed solution

As noted in Section 4.3, there is a clear divergence of opinion between those who think that human intelligence just is a matter of the right algorithms and some nifty programming, and those for whom key aspects of human intelligence are not mathematically describable at all. The answer, I think, is that neither of those positions is correct. Yes, the human mind is (probably), in principle, describable by an appropriate formal system; but it's at least an order of magnitude more expressively powerful than those effectively driving current or currently conceivable computing systems. We use the formal systems we do because they are ***relatively*** easy to understand and because they serve our purposes, ***not*** because they necessarily capture mind-independent reality.

6.1 The proposal elaborated

In formal logic as traditionally understood, the presence of a proposition $\mathbf{p \vee \neg p}$ – or its presumed truthfulness – renders the entire system inconsistent. There is a fairly simple set of self-referential statements (or *self-referential paradoxes*) that appear to present or reduce to a contradiction. They cannot be accommodated in a reliable and consistent matter because they appear to challenge the binary distinction of truth-functional statements into either true or false. (Of course, as Kurt Gödel is widely accepted as having proven [28], effectively all formal systems of sufficient expressive power to express any kind of self-referential statement are unable to express some class of self-referential statements without becoming inconsistent in the process: in other words, formal systems, by their nature, cannot be both complete and consistent. There will be statements that can be expressed in the formal system whose truth value cannot be determined within that system.) That renders any attempt to address inconsistent reasoning in contemporary computer systems limited and essentially *ad hoc*. My conjecture is that something about this is linked to life and the need of organisms to maintain homeostatis in a far-from-equilibrium environment.

In contrast at least to existing computer systems and standard formal logic, human beings seem capable of tolerating (even exploiting) inconsistencies systematically without becoming globally inconsistent in their reasoning. This allows them to hold seemingly contradictory beliefs as dictated by context while otherwise continuing to engage in rational thought and reason accurately. This, in turn, means they can entertain propositions (***and find meaning in them!***) that could not be entertained before.

6.2 Introducing systematic context sensitivity

Allow that propositions can be annotated by context, creating a context-sensitive logic: $\mathbf{p^A \vee \neg p^B}$. (To offer a trivial but useful example, $2 + 3 = 5$ in the context of Base 6 and above, while $2 + 3 = 10$ in the context of Base 5.) That no longer presents an obvious contradiction except in the case where $A = B$. Such a context-sensitive system is, naturally, phenomenally more complex and so difficult to comprehend – if not, indeed, impossible to comprehend in its entirety. On the other hand, it can be built; and human beings have good experience of creating things that ultimately outstrip human capacity to explain them.

7 Take-home message

ChatGPT, LaMDA, GPT-3… are expensive, computationally hugely inefficient, toys hopelessly unable to win the Imitation Game. They are best compared to AI art generators like DALL-E 2, Craiyon, and NightCafe, which draw what amount to composite pictures out of millions of pictures scraped from the Internet. Much of what they are doing amounts to fill-in-the-blank responses not essentially different from Eliza which was, after all, based on pattern recognition. In some cases, this is ***exactly*** what they are doing, as Eliza-type fill-in-the-blank pattern matching allows that certain answers are understood as needing flagged for modification or replacement.

They are, probably, useful for helping one find out where to go for information on the Internet, something that has grown increasingly complicated. They are also, as Chomsky noted, useful for hinting at solutions to programming tasks or suggesting rhymes. Otherwise, they reaffirm the psychologically well-known capacity of people to fool themselves – often, especially, the gullibility of scientists when reasoning outside their area of expertise; along with the lack of actual intelligence behind a lot of intelligent-seeming conversation.

ChatGPT and kin are sensitive to context in a non-systematic way. That increases their expressive power but not the "right" way and not enough. That is why they are black boxes, ***even to their creators***; and they cannot (for all the best intentions of the *explainable AI* community) explain their own behaviour. (Human beings often cannot, but at least they can try, and even succeed to a point.) Indeed, the notion of developing explainable AI to come up with honest explanations for their behaviour/conclusions strikes some as a bad joke. For all the hype, they still cannot perform to anything remotely like human standards, not only for those we accept (appropriately or otherwise) as "exceptionally" gifted but in general.

At the same time, they may be taken as committing the cardinal sin of attempting to reduce intelligence to "pure" associative learning (with bubble gum and shoestring added!). This strikes me as every bit as bad/misguided as the symbolic AI people of the 1980s who said that human intelligence "just is" manipulation of physical symbols by the brain or a computer. If symbolic AI over-intellectualizes things (and I believe that it does), then ChatGPT and kin dumb them down.

If the line of reasoning taken in this paper is at all correct, then human intelligence is (in certain critical respects at least) far more complex than it has generally been given credit; while the road to super-intelligence – if the term "super- intelligence" can be seen potentially to make sense in the future – lies through arriving at a better understanding of human (and, doubtless, non-human animal) intelligence first. If, at some point, we succeed in building artificial human-level intelligences (or *artificial general intelligences* as they are often called), we may find that they are subject to many of the same cognitive limitations we are, making many of the to-date distinctly human mistakes that human beings make.

References

[1] D. Yalalov, "ChatGPT passes the Turing test," https://mpost.io/chatgpt-passes-the-turing-test/, December 2022, accessed 3 May 2023.

[2] C. Metz, "How smart are the robots getting?" *New York Times.* [Online]. Available: https://www.nytimes.com/2023/01/20/technology/chatbots-turing-test.html

[3] A. Kharpal, "Stephen hawking says a.i. could be "worst event in the history of our civilization"," *CNBC*, November 2017, https://www.cnbc.com/2017/11/06/stephen-hawking-ai-could-be-worst-event-in-civilization.html.

[4] H. Lovecraft, *At the mountains of madness: The definitive edition.* Random House, 2005.

[5] S. Torrance, "Super-intelligence and (super-)consciousness," *International Journal of Machine Consciousness*, vol. 4, no. 2, pp. 483–501, 2012, http://dx.doi.org/10.1142/S1793843012400288.

[6] N. Chomsky, I. Roberts, and J. Watumull, "Noam chomsky: The false promise of chatgpt," *Asharq Al-Awsat*, March 2023, https://english.aawsat.com/home/article/4208906/noam-chomsky-ian-roberts-and-jeffrey-watumull/noam-chomsky-false-promise.

[7] K. Warwick and H. Shah, "Can machines think? a report on turing test experiments at the royal society," *Journal of Experimental & Theoretical Artificial Intelligence*, vol. 28, no. 6, pp. 989–1007, June 2015, https://doi.org/10.1080/0952813X.2015.1055826.

[8] Insane, "Is chatGPT woke and biased? 16 examples that prove so," https://www.theinsaneapp.com/2023/02/chatgpt-woke-examples.html, 2023, accessed 3 May 2023.

[9] R. M. Miller, "Holding large language models to account," in *AISB 2023, 13-14 April 2023.* The Society for the Study of Artificial Intelligence and Simulation of Behaviour, 2023.

[10] A. Turing, "Computing machinery and intelligence," *Mind*, vol. 59, no. 236, pp. 433–460, 1950, available online from many sources including https://www.abelard.org/turpap/turpap.php (accessed 2 February 2020).

[11] D. R. Hofstadter, *Gödel, Escher, Bach: An Eternal Golden Braid*. Basic Books, Inc., 1979, original publication.

[12] D. Chalmers, "The singularity: A philosophical analysis," http://consc.net/papers/singularity-ml.pdf, 2009, version of talk presented at Singularity Summit, New York City, 3-4 October 2009.

[13] N. Bostrum, *Superintelligence: Paths, Dangers, Strategies.* Oxford University Press, 2014.

[14] ——, "How long before superintelligence?" *Linguistic and Philosophical Investigations*, vol. 5, no. 1, 2005.

[15] W. Wallach and C. Allen, *Moral machines: Teaching robots right from wrong.* Oxford University Press, 2010.

[16] H. Gardner, *Multiple intelligences: The theory in practice.* Basic Books, 1993.

[17] H. Gardner and T. Hatch, "Multiple intelligences go to school: Educational implications of the theory of multiple intelligences," *Educational Researcher*, vol. 18, no. 8, pp. 4–10, 1989.

[18] J. Stewart, "Cognition = life: Implications for higher-level cognition," *Behavioural Processes*, vol. 35, no. 1-3, pp. 311–326, December 1995.

[19] B. Whitby, "The myth of AI failure (CSRP 568)," FTP archive currently offline as of April 2011., 2003, university of Sussex (UK) Cognitive Science Research Papers (CSRP) series.

[20] R. A. Brooks, "Intelligence without representation," *Artificial Intelligence*, vol. 47, pp. 139–159, 1991.

[21] ——, "Intelligence without reason," in *Proceedings, IJCAI-91.* Morgan Kaufmann, 1991.

[22] R. Penrose, *Shadows of the Mind: A Search for the Missing Science of Consciousness.* Oxford University Press, 1994.

[23] J. Searle, "Minds, brains, and programs," *Behavioral and Brain Sciences*, vol. 3, no. 3, pp. 417–458, 1980.

[24] A. Newell, "Physical symbol systems," *Cognitive Science*, vol. 4, no. 2, pp. 135–183, April-June 1980.

[25] N. J. Nilsson, "The physical symbol system hypothesis: Status and prospects," in *50 years of artificial intelligence: Essays dedicated to the 50th anniversary of artificial intelligence*, M. Lungarella, F. Iida, J. Bongard, and R. Pfeifer, Eds. Springer, 2007, pp. 9–17.

[26] D. Hofstadter, *Gödel, Escher, Bach: An Eternal Golden Braid.* Penguin, 2000.

[27] K. Gödel, "Unentscheidbare sätze der principia mathematica und verwandter systeme, i," *Monatshefte für Mathematik und Physik*, vol. 38, pp. 173–198, 1931, original publication.

[28] ——, *On formally undecidable propositions of Principia Mathematica and related systems.* Courier (Dover), 1992.

From epistemology to ethics of deep networks

Juraj Hvorecký
Department of Applied philosophy of ethics
Institute of Philosophy, Czech Academy os Sciences
Prague, Czech Republic
hvorecky@flu.cas.cz

***Abstract*—Blame for various missteps of AI is often assigned to either programmers or users. The discovery adversarial attacks and the complex epistemology behind neural networks leads us to skepticism toward their capacity to work efficiently in domains where robust human categories are central. Ethics is one such a domain and AI systems might be particularly unsuitable for operations in this domain.**

***Keywords*—neural networks, adversarial attacks, epistemology, ethics, blackboxing**

I. THE PROBLEM

The discriminatory practices of various artificial intelligence systems have been widely documented in academic literature and popular media, such as the documentary film "Coded Bias." Examples of unintended consequences of the use of artificial intelligence systems include biased Amazon hiring practices, major flaws in policing with automated surveillance cameras, and algorithmic stock market blunders. These issues have led to justified public outcry, and the resulting indignation has often resulted in either the abandonment of the original approach or improvements in the technology.

From a philosophical perspective, it is interesting to observe how blame is assigned for various mishaps and to what degree artificial systems themselves are held accountable. This responsibility assignment takes place against the backdrop of a larger debate about whether human beings are primarily responsible for any misuses of technology or whether the technology itself is to blame. This paper provides insights into why AI systems themselves may be deemed unusable and, thus, blameworthy for certain ethically sensitive tasks.

When presented with missteps of AI and deep networks that are usually at the backend of all AI systems, we are given two types of explanatory arguments. First, the training set of the network was flawed. There was insufficient diversity of certain crucial categories, and the results of optimization strategies within the network were thereby skewed in a particular direction, mis-categorizing important sections of the target groups. Examples of this ethical failure include Amazon hiring practices or the demonstrated lack of black and female face samples in face recognition software evaluated by Gender Shades [1].

Second, algorithms might turn out to be well trained on a balanced data set and eventually become efficient in solving a designated task, yet upon completion of their training, they are inappropriately applied to a domain on which they were not trained. While partial transferability of categorization efficiency within networks is possible, it is generally the case that trained networks struggle with genuinely novel input. The paper will provide further insights into why that tends to be the case.

It is important to note that both explanations effectively blame the human programmers and operators, not the AI technology itself. On the one hand, intentionally or not, the training set has been wrongly selected and designed, resulting in a biased network. On the other hand, a finished network is misapplied outside of its area of competence, leading to major categorization mistakes. In both scenarios, the faulty result is brought about by a procedure that takes place either before or after the network's constitution. Although the network is co-constituted by its training set, the main argument rests on the fact that the training set, in its skewed form, was in place before the resulting network emerged. Therefore, it is the fault of humans that networks have been trained in a biased way or applied incorrectly

II. EPISTEMOLOGICAL CONCERNS

In order to shift the debate and responsibility away from human operators and towards the nature of AI systems, it is necessary to consider the epistemology of networks and their inner workings. Networks are complex structures consisting of multiple layers that are trained on large amounts of data and evaluated based on resulting categorizations. Positive evaluations strengthen connections between input characteristics and categorization outcomes, while negative evaluations weaken them. Networks continually update their inner structure through the learning process by adjusting the weights of connections between nodes. Once a network achieves a high success rate in a categorization task during training, it is ready for implementation, which involves processing novel input of the same category. However, the constant adjustment of a network's inner structure leads to the blackboxing problem, where external observers and even network creators are unaware of the specific functional roles of individual connections or their subsets.

The discovery of adversarial attacks, which produce counterintuitive results when fed by slightly altered input, has highlighted the blackboxing problem [2], [3]. Researchers have found that perfectly trained networks are susceptible to such attacks, and that the alterations to the input can be imperceptible to humans (though see [4]). The root causes of this susceptibility have been widely debated, as have potential defense strategies. Evidence has shown that the susceptibility of networks to adversarial attacks is not related to a format of training material, nor is it due to intrinsic low quality of training sets. Additionally, it has been discovered that adversarial attacks are transferrable across networks trained on the same data set, regardless of the algorithms employed

[5]. This suggests that the inner workings of individual networks do not play a significant role in determining success rates, and raises important questions about the reliability of AI systems.

Epistemological explanations of network behavior can provide insight into this complex phenomenon [6]. Networks function as optimization machines, processing large sets of input and receiving assessments on their performance. Their task involves searching for the function that best approximates the relation between input properties and desired output. Networks achieve this through the internal adjustment of weights between their nodes, resulting in a highly complex function that maps inputs onto outputs. This complexity and perpetual change contribute to the problem of blackboxing, as external observers may not know what properties the function tracks.

This epistemological insight reveals that networks may not rely on humanly recognizable features in their search for an optimal solution to functional linkage problems. Instead, networks may track more efficient, but non-humanly perceptible properties. The transferability feature indicates that these properties may be optimal for achieving desirable categorization results.

Recent work by [7] highlights the non-humanly perceptible features that networks track. They were able to distil these features from inputs, contrast them with humanly accessible properties, and show that networks rely on non-robust features that have no humanly recognizable categories. Their experiments demonstrate that networks can categorize an image correctly, even if the image does not resemble anything that a human observer would categorize as such.

.

III. Ethics

However, the concern arises when networks are used in domains with ethical consequences. Ethical decision-making relies on a long list of humanly graspable categories that guide our actions, such as age, race, vulnerability, privilege, species, and artificiality. Networks may not operate on humanly detectable properties and categories, making it questionable whether they should be delegated to make ethically relevant decisions.

In summary, while networks may achieve unprecedented speed and efficiency through tracking non-humanly perceptible properties, their limitations in dealing with humanly graspable categories should be taken into account when considering their use in domains with ethical consequences.

This observation should not be regarded as a sweeping critique of AI adoption in all areas of application. Certain tasks, such as identifying optimal perfume or protein structures, are likely to be well-suited to AI and have no clear ethical implications. However, my assessment of AI's limitations is focused specifically on domains of human activity where ethical considerations are crucial.

The recognition of the threat of adversarial attacks has led to a significant focus on developing mechanisms that can mitigate their impact. While I commend these efforts, I must note that they are primarily reactive. They involve the introduction of various filters and sub-mechanisms into networks to prevent susceptibility to the issue, rather than addressing the more general concern that networks may not be tracking robust human categories. These solutions attempt to modulate internal optimization functions to avoid falling prey to specific adversarial threats.

Finally, the strategy employed in this paper may have implications for a larger problem, alluded to in section I. Previous attempts to explain problematic AI behavior have often blamed human operators or creators of artificial systems, who either provided biased data or misapplied them to inputs they were not trained for. However, this analysis suggests that the problem may actually lie with the networks themselves. The way they search for an optimal function within a task may lead to the adoption of strategies that are unsuitable for specific domains, including those where humanly observable criteria are of utmost importance, such as in ethical and ethically sensitive issues. The use of artificial networks in ethics may be a case where blame does not necessarily fall on a human administrator, but on the technology itself.

Acknowledgements

This article is the outcome of the Czech Science Agency (GAČR) project n. 20-14445S "Dual Models of Phenomenal Consciousness".

References

[1] I. Raji, I and J. Buolamwini, “Actionable Auditing: Investigating the Impact of Publicly Naming Biased Performance Results of Commercial AI Products”. Conference on Artificial Intelligence, Ethics, and Society, 2019.

[2] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna J, D. Erhan, I. Goodfellow, R. Fergus (2013) “Intriguing properties of neural networks”, 2013 arXiv: 1312.6199I

[3] S. Huang, N. Papernot, I. Goodfellow, Y. Duan, P. Abbeel, “Adversarial attacks on neural network policies”, 2017, arXiv: 1702. 02284 Maxwell, A Treatise on Electricity and Magnetism, 3rd ed., vol. 2. Oxford: Clarendon, 1892, pp.68–73.

[4] C. Tan, F. Sun, T. Kong, W. Zhang, C. Yang, C. Liu (2018) October) A survey on deep transfer learning. International conference on artificial neural networks. Springer, Cham, pp 270–279 Z. Zhou, C. Firestone, “Humans can decipher adversarial images”. Nat Commun 10, 1334 (2019). https://doi.org/10.1038/s41467-019-08931-6.

[5] C. Tan, F. Sun, T. Kong, W. Zhang, C. Yang, C. Liu (2018) October) A survey on deep transfer learning. International conference on artificial neural networks. Springer, Cham, pp 270–279

[6] C. Buckner “Empiricism without magic”, 2018, Synthese 195, pp. 5339–5372.

[7] A. Ilyas, S. Santurkar, D. Tsipras, L. Engstrom, B. Tran, A. Madry, “Adversarial examples are not bugs, they are features”, 2019, arXiv: 1905. 02175.

My belief or Alexa's? Belief attribution and AI extension

Hadeel Naeem
School of Philosophy, Psychology and Language Sciences
University of Edinburgh

Abstract

This paper investigates how we ought to attribute beliefs in the kinds of human-AI interactions that give rise to extended beliefs. Compared to classical cases of extension, AI-extended agents have a less active role in forming their beliefs, and this paper explores how we might nonetheless be able to ascribe beliefs to them. Toward this goal, I look for a suitable account of belief attribution in the extended mind literature. I examine the dynamical systems theory (DST), the boundary of the mind according to Markov blankets, and the virtue reliabilist concept of cognitive integration. I find all of these wanting. Of all the examined accounts, cognitive integration is best suited to explain how we attribute beliefs in non-AI extension cases, but it still ultimately fails in cases of AI extension. I show that AI systems will soon monitor their own cognitive integration, leaving no room for agents to manifest sufficient cognitive agency for these extended epistemic states to be attributed to them.

Who Needs Needy Machine Consciousness?

Max Jones
Department of Philosophy
University of Bristol
Bristol, UK
max.jones@bristol.ac.uk

***Abstract*— Much of AI Ethics has focused on the dangers of superintelligent machine consciousness. However, recent work on the homeostatic/affective basis of consciousness and on developments in soft robotics suggest that we may be able to build simple conscious machines in the relatively near future. As such it is pressing to ask whether we should build machines that may be conscious according to this theory. Here, I argue that we should avoid doing so because we don't need and certainly don't want needy conscious machines.**

Keywords—AI Ethics, Soft Robotics, Homeostatic consciousness.

I. FEELINGS, GENUINE INTELLIGENCE, AND CONSCIOUSNESS

Cantwell Smith argues that we may be further from creating genuinely intelligent AI than many assume [1]. While our current and emerging sophisticated AI systems may be capable of supreme reckoning abilities, they lack the capacity for judgement. In order to be capable of true judgement, AI systems must be capable of representing the world in a way that matters to them. Their decisions must have consequences for their own survival. In short, they must have needs.

Recent work on the neuroscience of consciousness suggests that the possession of needs and the ability to act in relation to them may be central to the function of consciousness [2, 3]. On this homeostatic/affective view, consciousness is inherently tied to affect, with a subject's affective states intimately linked to their overarching goal of maintaining homeostatic viability. This suggests that creating what Cantwell Smith sees as genuine intelligence may require the creation of (at least rudimentary) machine consciousness.

Man & Damasio have argued that soft robotics may provide a route to both genuine intelligence and machine consciousness [4]. By designing robots to be vulnerable in their exchanges with the environment and with a drive for self-preservation, it may be possible to render their interactions with the environment as meaningful in a way that rigid robots' interactions are not. One can similarly expand the repertoire of robots "feelings" by introducing other vulnerabilities, such as hunger or the need to care for offspring [5].

If these theorists are correct and genuine intelligence and machine consciousness just require us to build robots with needs, then the technologies for achieving these goals may not be as distant a prospect as some assume. There is already work on vulnerable soft robots that can act in relation to their own self-healing properties [6]. Robots with artificial guts that power themselves with energy from the environment have been developed [7, 8]. Self-replicating robots are already a possibility [9]. Moreover, one could presumably also engineer robots to care for the survival of their progeny (e.g. by implementing a biased ethical model based on simulation theory akin to [10]). On could go as far as to say that, if we wanted, we could build a robot with needs akin to biological organisms in the next 10 years. Perhaps we should stop thinking of these near future robots as machines and instead conceptualise them as artificial organisms [11].

II. DENNETT'S "NO MORE COLLEAGUES" CHALLENGE

Supposing that the affective/homeostatic approach to machine consciousness is correct and engineering conscious robots isn't such a distant prospect, the pressing questions become whether and why we should want to build conscious machines of this kind. Dennett has argued that we should avoid creating conscious machines even if it is possible [12]. We have plenty of conscious colleagues already. "If you want a conscious agent, we've got plenty of them around and they're quite wonderful, whereas the ones that we would make would be not so wonderful." For Dennett, the costs of producing conscious machines may be great and in the end we just end up with more colleagues, which we were never short of in the first place. The value of AI systems and robots may lie precisely in their not being conscious, since they are able to carry out impressive feats of intelligence relatively cheaply without needing to experience anything and without having any needs that must be satisfied.

We don't have to worry about our AI systems and robots refusing to work because they don't *feel* like it or because they *feel* like they are being treated unfairly. We struggle to organise the economy in a way that allows us to look after the wellbeing and mental health of countless humans. Perhaps this should give us pause in aiming to introduce a whole swathe of new entities whose needs should be also met. "We want smart tools, intelligent tools, not artificial colleagues".

III. THE DANGERS OF NEEDY CONSCIOUS ROBOTS

Dennett's critique arguably applies to all forms of machine consciousness. However, it is particularly pressing in the context of the affective/homeostatic approach to machine consciousness. If machine consciousness requires robots with real needs that impact on their survival, then the continued existence of machine consciousness requires that those needs be met. This will clearly come with some significant dangers and costs. It's not hard to imagine the direct dangers to humanity that could result from creating vulnerable and hungry robots with a drive for self-preservation, a capacity to reproduce and a drive to also protect their vulnerable progeny. Since we are organisms with our own energetic needs, we would inevitably be in competition with such robots for energetic resources, so such robots would be right to infer that we are a threat to their own and their progeny's survival. Hungry robots can decide to eat you or to kill you to protect the resources that we are competing to secure.

One potential way to avoid these kinds of nightmare scenario and to mitigate the potential costs of machine consciousness is to build robots that can only meet their energy requirements by consuming materials that we are unable to consume. Even better would be to do this with materials that we are keen to get rid of.

For example, one could build robots that are hungry for and capable of digesting the plastics that we need to remove from the ocean [8]. In short, we can try to ensure that the robots can meet their needs in a way that aligns with our own needs. However, the problem with this line of thought is that it's not clear why conscious robots would be needed to fulfil tasks that happen to align with the presence of unwanted resources. Most that see applications for conscious robots tend to envisage them being better at interacting with humans, but we tend to avoid locations that are full of toxic of polluting resources. Moreover, this may be precisely the kind of application where machine consciousness could potentially be dangerous, as it might enable the kind of behavioural flexibility that allows the robot to set its own goals that deviate from what we desire. If there are ways of solving similar problems (perhaps using robotics) that nonetheless don't involve the costs and risks of machine consciousness, then these should be preferred.

Given these risks and costs, the potential benefits of creating machine consciousness would have to be large to make its development a worthwhile endeavour. Unfortunately, the benefits of machine consciousness are far from clear. Some have argued that machine consciousness may be necessary for robots to feel empathy, and that this may be a prerequisite for their being able to engage in moral deliberation [4]. This may be true, but it just points to the further question of why we would want robots to be capable of such deliberation. The capacity for moral deliberation is not the same as a guarantee of morally acceptable action. After all, we have plenty of examples throughout history of conscious humans, capable of empathy and moral deliberation, doing abhorrent things to one another.

IV. PROTECTING AGAINST NEEDY ROBOTS

The deeper lesson that we should learn from considering the homeostatic/affective approach to machine consciousness is that we shouldn't assume that consciousness is aligned with superintelligence. Most work concerning consciousness in AI ethics has focused on worries about the singularity and super-intelligent conscious AI [13], but the homeostatic/affective approach to consciousness in living things takes it to be relatively widespread and possessed by relatively simple and unintelligent creatures. We therefore need to be careful, as building needy and not so intelligent conscious robots may be something already or nearly within our grasp, as a result of developments at the intersection of robotics and artificial life.

What can we do to protect against these costs and dangers? One option is to give robots needs without giving them the capacity to register those needs. One can build a robot that is, in a sense, hungry (in that it is energetically autonomous and must consume environmental resources to maintain its own viability) without building a robot that represents its own homeostatic state. In doing so, we may have to forgo the potential benefits, for example, in terms of behavioral complexity, of a robot that, in some sense, genuinely *feels* hungry. Yet, this may be a price worth paying to avoid the dangers of a robot with needs that it is aware of.

A second option is to accept that the advent of needy robots is inevitable and to ensure that we have the requisite moral framework and safeguards in place. Importantly, the relevant moral framework is unlikely to emerge from considerations of superintelligence and the singularity. Rather, we should turn to established issues in non-human animal ethics and the bioethics of artificial life [14, 15].

Most importantly, any discussion of whether we should create potentially conscious machines with their own homeostatically driven needs, should be conducted in the context of considering the needs of already existing humans and nonhuman animals. By creating new entities with needs of their own, we inevitably create more needs that must be satisfied with a still finite pool of resources. Conscious machines may be just around the corner. All we need to do is give them needs and the capacity to register and act upon those needs. What isn't clear is whether and why *we* need a proliferation of needy companions, particularly when we already struggle to cater for the needs of those that already exist.

REFERENCES

[1] B. Cantwell Smith, The Promise of Artificial Intelligence: reckoning and judgement. Cambridge, MA: MIT Press, 2019.

[2] M. Solms, The Hidden Spring: A journey to the source of consciousness. London: Profile Books, 2021.

[3] A. Damasio and H. Damasio, "Homeostatic feelings and the biology of consciousness," Brain, vol. 145, pp. 2231-2235, May 2022.

[4] K. Man and A. Damasio, "Homeostasis and soft robotics in the design of feeling machines," Nature Machine Intelligence, vol. 1, no. 10, pp. 446-452, 2019.

[5] D. Parisi and G. Perosino, "Robots that *have* feelings," Adaptive Behavior, vol. 18, no. 6, pp. 453-469, 2010.

[6] E. Roels et al., "Processing of self-healing polymers for soft-robotics," Advanced Materials, vol. 31, no. 6, 2022.

[7] C. Melhuish, I. Ieropoulos, J. Greenman, and I. Horsfield, "Energetically autonomous robots: Food for thought," Autonomous Robots, vol. 21, pp. 187-198, 2006.

[8] H. Philamore, J. Rossiter, A. Stinchcombe, and I. Ieropoulos, "Row-bot: an energetically autonomous water boatman," IEEE/RSJ International Conference on Intelligent Robots and Systems, pp. 3888-3893, 2015.

[9] K. Lee and G. S. Chirikjian, "Robotic self-replication," IEEE robotics & automation magazine, vol. 14, no. 4, pp. 34-43, 2007.

[10] D. Vanderelst & A Winfield, "An architecture for ethical robots inspired by the simulation theory of cognition," Cognitive Systems Research, vol. 48, pp. 56-66, 2018.

[11] J. Rossiter, "Soft robotics: the route to true robotic organisms," Artificial Life and Robotics, vol. 26, no. 3, pp. 269-274, 2021.

[12] J. Brockman, D. Chalmers and D. Dennett, "Is Superintelligence Impossible? On Possible Minds: Philosophy and AI with Daniel C. Dennett and David Chalmers", Edge Conversations, 4.10.19.

[13] V. Müller, "Ethics of artificial intelligence and robotics," Stanford Encyclopedia of Philosophy, 2020.

[14] T. Douglas, R. Powell, and J. Savalescu, "Is the creation of artificial life morally significant?" Studies in History and Philosophy of Science Part C, vol. 44, no. 4, pp. 688-696, 2013.

[15] A. Christiansen, "Synthetic biology and the moral significance of artificial life," Bioethics, vol. 30, no. 5, pp. 372-379, 2016.

Artificial Life as Controlled Disequilibrium

Steve Battle
Computer Science and Creative Technologies
University of the West of England, Bristol
Bristol, UK
steve.battle@uwe.ac.uk

***Abstract*—Artificial Life can be modeled using simulated autopoiesis. Liquid Automata are used to define simulated chemical reaction systems; particle systems with rules governing how particles are transformed on collision with each other. Unlike cellular automata, there is no fixed grid or time-step, only particles moving about and colliding in continuous space/time. These systems may give rise to emergent artificial life, or they may be (artificially) lifeless. Can we distinguish between these systems by analysing their equilibria?**

I. INTRODUCTION

The dynamics of non-living systems is characterised by increasing entropy. A stone rolling to the bottom of a valley can be explained by purely physical forces that maximise entropy. By comparison, a life-form like a single-celled bacterium exploits external energy sources to minimise entropy locally, creating "order from disorder" far from chemical equilibrium [1]. According to James Lovelock, this principle can be used to look for life on alien worlds by analysis of planetary atmospheres [2]. A system that is far from equilibrium may provide evidence of life. On Mars, Curiosity rover discovered levels of oxygen 30% higher in spring and summer than expected. This may be evidence of life, or must be explained by other means.

II. ARTIFICIAL LIFE

How then are we to understand life? Is it just a smorgasboard of features such as the capacity for growth, reproduction, functional activity, and change? This approach is not particularly scientific. Maturana and Varela argue that life is no more and no less than a system that exhibits *autopoiesis* [3], the ability of a living system to self-organise and produce itself in the physical realm. A living system is a self-referential domain of interactions in the physical space, generally a network of 'chemical' relationships. According to Maturana and Varela, "An autopoietic machine is a machine organized (defined as a unity) as a network of processes of production (transformation and destruction) of components." Yet, there are many kinds of chemical networks that aren't alive, consider a chemical explosion characterised by a runaway chain reaction of positive feedback. The signature of life is the emergence of structure that distinguishes self from non-self, closing it off from its environment, "A universe comes into being when a space is severed into two. A unity is defined." This closure emerges from, and is dynamically maintained by the organism. This principle is also central to Spencer-Brown's 'Laws of Form', "a distinction is drawn by arranging a boundary with separate sides so that a point on one side cannot reach the other side without crossing the boundary. For example, in a plane space a circle draws a distinction." [4]

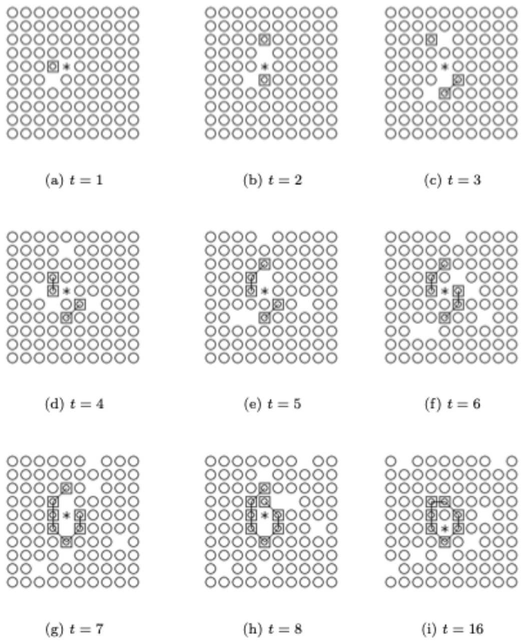

(a) $t = 1$ (b) $t = 2$ (c) $t = 3$
(d) $t = 4$ (e) $t = 5$ (f) $t = 6$
(g) $t = 7$ (h) $t = 8$ (i) $t = 16$

Fig. 1: A simulation of autopoiesis using a discrete time cellular automaton on a rectangular grid, based on Varela's original algorithm. In (a) a pair of substrate (circles) are transformed into a single link (squared circle) by the catalyst (asterisk). By (c) we see the first bonds forming, then in (i) these finally form a closed boundary around the catalyst.

There are two key features of autopoiesis:

1) Organisational closure: A network of the processes of production, e.g. a chemical reaction system.
2) Structural closure: The appearance of a homeostatically maintained boundary that divides self from non-self, also described as the maintenance of identity, e.g. a cell-wall.

We can simulate autopoietic processes computationally — not life as we know it, but *artificial* life. Conway's "Game of

Life" [5], [6] is a classic Cellular Automaton that exhibits emergent, self-organised behaviour. It gives rise to cyclic patterns that are self-reproducing, such as the 'blinker' and 'glider' patterns. But these patterns are not obviously life-like in the sense defined by Maturana and Varela as they lack a clear structural boundary (however, see Beer [7] for an alternative view). To demonstrate the process of autopoiesis, Francisco Varela devised a novel Cellular Automaton that shows how a simple system of rules can give rise to a 'cellular' structure [8]. It implements a simplified model of the chemical reaction rules found in living cells, demonstrating how these work together to form an organisationally closed system of production. Figure 1 shows output from this system; a catalyst (shown as an asterisk) is the 'seed'; substrate particles (circles) are consumed by the system, as pairs of them are composed by the catalyst to form new link particles (squared circles); and link particles self assemble into a compound structure — a primitive cell wall with bonds shown as lines drawn between them.

The rules implemented by Varela's Cellular Automaton are described as chemical reaction rules. A reaction rule has a left-hand side defining the *reactants*, separated by an arrow from the reaction *products* on the right-hand side. The appearance of a $+$ (plus) symbol between reactants indicates an event where all of the indicated particles must come together. Each particle type may be prefixed by a positive integer indicating a number of particles of the same type, so that $K + 2S$ is equivalent to $K + S + S$, the interaction of three particles. The use of plus between reaction products, indicates that the reaction produces multiple outputs. Bonds are indicated by multiplicative operators such as $L.L$, L^2, or L^n [9]. These reaction rules are summarised below, where K represents the catalyst, S the substrate, and L the link particles.

$$\text{composition}: K + 2S \rightarrow K + L \tag{1}$$

$$\text{disintegration}: L \rightarrow 2S \tag{2}$$

$$\text{concatenation}: L^n + L \rightarrow L^{n+1} \tag{3}$$

Figure 1 illustrates a number of steps from a single run. In step (a) at time, $t = 1$, a pair of substrate particles are composed into a single link particle by the catalyst. After the composition of a number of link particles, we see the first bonds forming between them in step (c) at $t = 3$. The concatenation rules in Varela's algorithm are constrained to forming only obtuse bond angles. This prevents the uppermost link particle at (h) $t = 8$, from bonding with the particle immediately below it. It is only when this particle disintegrates later at (i) $t = 16$, that enables the remaining links to re-bond, and form a closed boundary around the catalyst.

These same rules were re-implemented in a later program called SCL (Substrate, Catalyst and Link) by McMullin using the SWARM system [10] where chemical reaction rules are captured in a modular fashion enabling their reaction rates to be more precisely controlled. This program introduced a more configurable way to control the random motion of particles in 2D space, modelled as a square lattice with toroidal toplogy and discrete time.

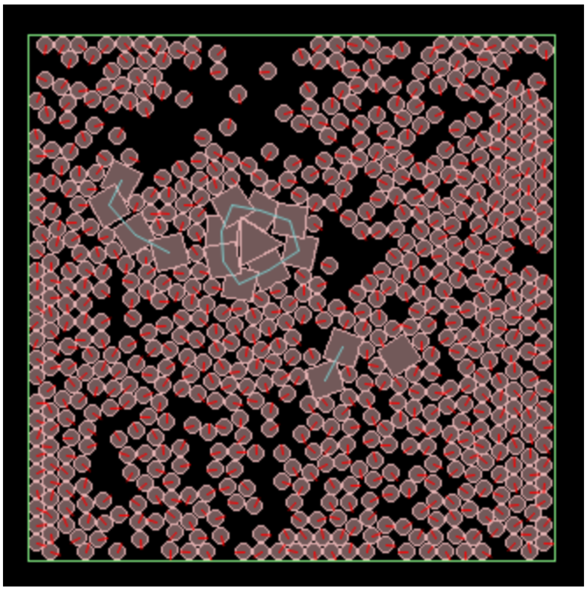
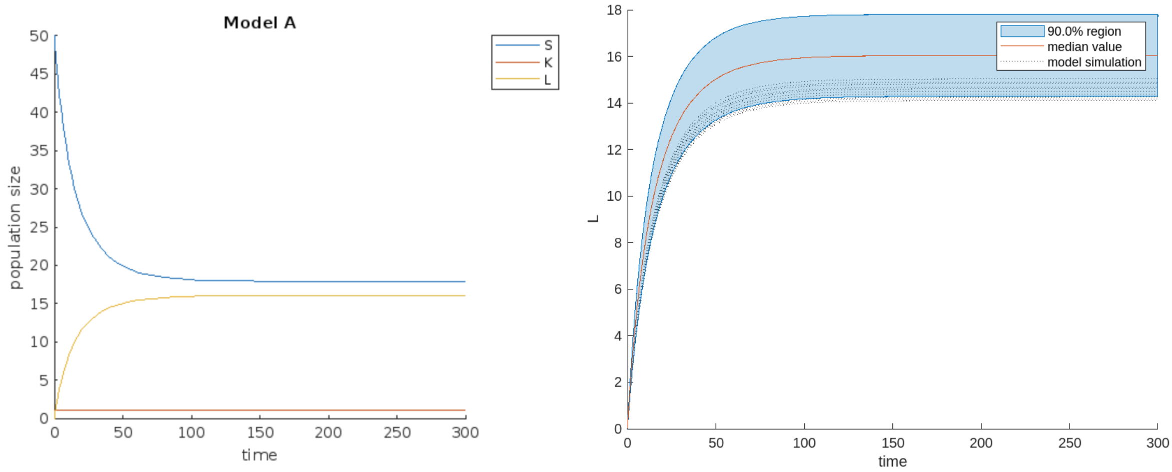


Fig. 2: Liquid automaton showing a boundary (blue links) forming around the catalyst (triangle), distinguishing self from non-self. The catalyst transforms the substrate (circles) into its structural building blocks (squares).

III. PARTICLE SYSTEMS

There is currently a great deal of interest in 2D particle systems, though the same ideas extend to three dimensions. Particle systems are game physics engines designed to reproduce naturalistic phenomena based on objects moving around, typically in a 2D space. The engine used for the Liquid Automata described in this paper is Box2D (specifically pybox2D), a rigid body simulation library for games. Interestingly, Box2D has been used as the game engine for a number of implementations of "Angry Birds." Each particle is a 2D body with mass and velocity, so particles have three degrees of freedom; translation along x,y axes, and rotation. Each body is associated with one or more shapes which can be any geometrical construct, such as the squares, circles, and triangles seen in Figure 2. Joints define constraints on the relative motion between two bodies, used here to create bonds between neighbouring links. Forces, torques, and impulses are applied to bodies to make them move. Box2D includes a high performance iterative constraint solver that resolves joint constraints, particle motion and resulting collisions [11]. Particles can bounce off each other in elastic collisions, or slide against each other based on a realistic simulation of the frictional forces between them. A world may, or may not, have gravity; the simulations described here do not.



(a) (b)

Fig. 3: (a) Model A runs to equilibrium with L converging at around 16. (b) This equilibrium is sensitive to variations in initial population size, S. 'Sobol' analysis shows the effect of this variation, with the blue region covering 90% of the observed variation.

IV. LIQUID AUTOMATA

We reproduce and extend Varela's simulations of autopoiesis [12] using so-called *Liquid Automata*. These are based on 2D particle systems with added support for chemical reaction rules. Unlike cellular automata which divide space into a regular fixed grid, and time into discrete steps, Liquid Automata implement a continuous model of space/time. There is no fixed grid, just particles moving about and colliding with each other in continuous space. By analogy with cellular automata, the Liquid Automaton is a variety of *collision-based* system [13]. Particles move freely in space and react 'chemically' when they collide. Rules (1) & (3) are invoked when the necessary particles come into contact with each other, while rule (2) occurs spontaneously.

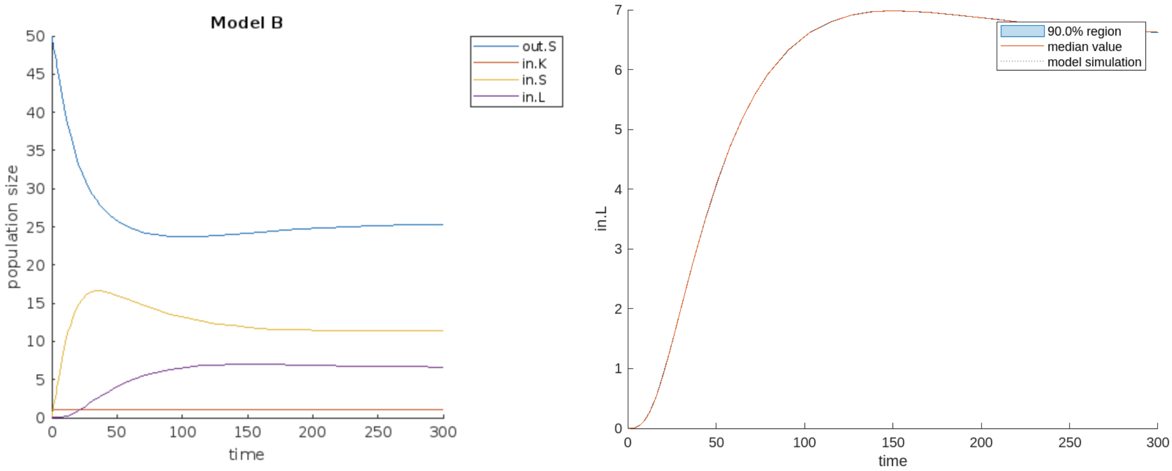
External energy is added to the system in the form of random 'Brownian' motion defined in terms of a Wiener process along the x,y dimensions [14]. The force applied to every particle along each dimension is a normally distributed random variable with zero mean, and variance, $(delta)^2 dt$, correlated with a single parameter delta, and time period, dt, which varies dynamically. It is this external energy source that drives the system to self-organise, and create a local island of order from the surrounding chaos.

A Liquid Automaton defines an organisationally closed chemical reaction system (CRS). Figure 2 shows output from a Liquid Automaton implementing reaction rules (1), (2), and (3). Bathed in a liquid substrate, the catalytic agent (triangle) composes substrate (circles) into its structural building blocks, or links (squares). Links are subject to decay, and may spontaneously disintegrate back into a pair of substrate particles. The link particles are able to self-organise (blue links) into a structure akin to a long-chain polymer. Links can make up to two connections, and a chain of concatenated links is able to wrap around and close in on itself. This emergent closed boundary, analogous to a cell wall, divides self from non-self; the signature of (artificial) life.

V. EQUILIBRIUM AND DISEQUILIBRIUM

If artificial life is to be recognised by its disequilibrium, we must first define what the equilibrium state would be in the absence of artificial life. If a key characteristic of life is the boundaries that it forms, then we need to study the system in the absence of those boundaries. Deleting the concatenation rule (3) that bonds particles together, disrupts the internal organisation of the system. The network of relations no longer performs a complete cycle that produces an enclosed structure so the boundary never forms. We will call the this simpler 'chemical' system, System A, while the full system with boundaries will be known as System B. For experimentation we need a measurable proxy for this emergent structure. By observation, the boundary appears to regulate the number of substrate particles reaching the catalyst, with the knock-on effect of throttling back the number of link particles produced by the reaction. We therefore use the number of link particles as a proxy for this structure and *hypothesise that the mean number of link particles is higher in System A than it is in System B.*

The Liquid Automaton for System B maintains a mean of 12.46 link particles, based on samples taken every second over a period of 30 minutes, with an initial substrate population of 500 substrate particles. Deleting the concatenation rule in System A, we see an increased mean of 17.56 link particles over the same time period, given the same initial conditions.



(a) (b)

Fig. 4: (a) Model B runs to equilibrium with $in.L$ converging around 6.6. There is an initial surge of substrate, $in.S$, into the interior before the boundary is fully formed. (b) The ‘Sobol’ analysis is absent of any variation, so the equilibrium of $in.L$ is independent of the initial population size, $out.S$.

This 29% difference from artificial chemistry is the ‘tell’ of artificial life at work. Its emergent structure shifts it away from ‘chemical’ equilibrium.

We further hypothesise that the higher point of equilibrium for System A is sensitive to initial conditions. Simulation with Liquid Automata is computationally expensive, and the data it produces is too variable to perform sensitivity analysis reliably. Even without the concatenation rule, the halo of reaction products around the catalyst produces a barrier that impedes subsequent reactions with the substrate. Instead, we define simplified models using the Mathworks SimBiology package [15]. These models are more tractable as they do not require a physics engine, and only take population size into account. They cannot produce any organised states of matter. We define a simpler Model A, corresponding to the Liquid Automaton of System A. Parameters governing the reaction rates are set to simple ‘Mass Action’ kinetics, where the rate of a chemical reaction is directly proportional to the concentration of the reactants. Model parameters governing reaction rates are tuned only to the extent of achieving comparable qualitative behaviour, with results of the same order of magnitude. Finer tuning doesn’t appear to provide any additional insight.

Model A comprises just the reaction rules for composition (4) by the catalyst, K, and spontaneous disintegration (5) of link particles back into substrate, S, at the rates indicated. There is no rule for concatenation as the simulator does not support the notion of compound particles. We will continue to refer to ‘link’ particles for continuity with Varela’s nomenclature, but these numerical simulations do not perform any such linking. This ‘chemical’ simulation will naturally run towards a state of equilibrium, at a rate governed by the “Law of Mass Action”. These rules and their corresponding (mass-action) reaction rates are defined below.

$$\text{composition} : 2S + K \rightarrow K + L \qquad \text{rate = 0.5E-3} \tag{4}$$

$$\text{disintegration} : L \rightarrow 2S \qquad \text{rate = 0.01} \tag{5}$$

Figure 3(a) shows a simulation run of model A. In these simulations we only consider a single catalyst particle, K, remaining constant at 1. The simulations begin with no link particles, and it has an initial population size of $S = 50$. The system reaches equilibrium when the production and decay of link particles, L, reaches a balance, converging at around 16 particles. Sensitivity analysis allows us to explore the effects of variations in model quantities on a model response. We want to explore the time-dependent sensitivities of the number of link particles, L, with respect to the initial substrate population, S. Using Local Sensitivity Analysis (LSA) we compute, $\delta L/\delta S = 1.22$. This shows a clear sensitivity of the state of the output species L on the initial population size of the substrate, S. This is positive, showing that in Model A, L increases with S, as expected. Figure 3(b) illustrates the results of a variance-based sensitivity analysis using the ‘Sobol’ method, showing the simulation results and mean value. The shaded region shows the spread of outputs, covering 90% of the results, based on a random spread of perturbations to the initial population size. We conclude that the stability exhibited by Model A is contingent, being sensitive to initial conditions.

VI. CONTROLLED DISEQUILIBRIUM

By “controlled disequilibrium”, we mean a new point of equilibrium displayed by a self-organising system that is both distinct from the ‘chemical’ equilibrium we find in the absence of self-organisation, and is intrinsically stable. This intrinsic stability can be understood as an internally set goal or purpose. In other words, it behaves as a control system to maintain that goal. As before, we measure the mean number of link particles observed from second to second and investigate whether the goal criteria, the goal value for L, is stable in the face of perturbations in the initial conditions. *We hypothesise that the artificially living system maintains this new point of equilibrium over a range of initial conditions.*

As before, we build a simplified model of System B, using MathWorks SimBiology [15], and call this Model B. Whereas in the System B Liquid Automaton, the boundary is an emergent property, Model B introduces explicit compartments that separate particles inside the boundary from those outside. The catalyst, K, now only reacts with substrate, S, inside the boundary (rule 6). We also assume that when link particles disintegrate, the waste substrate is ‘dumped’ outside the boundary, ready for recycling (rule 7). The boundary is semi-permeable allowing substrate particles to diffuse across in either direction at a rate that is inversely proportional to the density of link particles forming the boundary (rule 8). Model B incorporates the reaction rules below.

$$\text{composition} : 2\,in.S + in.K \rightarrow in.K + in.L \;\text{rate=0.5E-3} \quad (6)$$

$$\text{disintegration} : in.L \rightarrow 2\,out.S \quad \text{rate=0.01} \quad (7)$$

$$\text{diffusion} : out.S \longleftrightarrow in.S \quad \text{rate=}1/(in.L + 1) \quad (8)$$

Figure 4(a) shows a simulation run of model B, with an initial population size of $out.S = 50$. There is an initial surge of substrate into the interior before the boundary is fully formed (the density of link particles increases). The quantity of catalyst, $in.K$, again remains constant at 1. The mean number of link particles, $in.L$, converges at around 6.6 particles, lower than the point of equilibrium for Model A. Again, using Local Sensitivity Analysis (LSA), we compute, $\delta in.L/\delta out.S = 0$, the time-dependent sensitivities of the output species $in.L$ with respect to the initial substrate population, $out.S$. This is 0, demonstrating that $in.L$ does not vary with $out.S$, and so there is no sensitivity to perturbations in $out.S$ over the range tested. Similarly, the ‘Sobol’ plot in Figure 4(b) shows no variance in $in.L$. The shaded region of variance is all but invisible.

We are in effect applying Ashby’s “test for independence” [16, p158] to demonstrate that Model B induces temporary independence (over the range tested) between the initial population of substrate, and the density of link particles; “constancy in a subsystem’s state implies that the state is one of equilibrium, and constancy in the presence of small disturbances implies stability.” We conclude that the stability exhibited by Model B is largely independent of initial conditions. The variable diffusion rate creates a feedback control loop based on the quantity of link particles, making it robust to disturbance.

VII. CONCLUSION

A central feature of living systems is that they exist out of equilibrium with their environment. All closed systems will ultimately reach thermodynamic or chemical equilibrium, but this may simply be contingent on initial conditions. A key aspect of this disequilibrium is that it is under control by the organism and maintained within a specific range necessary for the organism’s continued survival.

Liquid Automata enable us to investigate the mechanisms of artificial life based on particle simulations with added reaction rules. They are a tool for exploring emergent phenomena and the equilibrium states that arise. The equilibrium state of artificial life (simulated autopoiesis) falls far from the baseline equilibrium of an (artificially) lifeless chemical reaction system. This baseline equilibrium is also contingent on initial conditions. Artificial life self-organises, defining a new point of disequilibrium relative to this baseline. Furthermore, this new equilibrium is dynamically controlled through feedback, remaining stable in the face of perturbation. Artificial Life has a goal, its purpose to survive.

REFERENCES

[1] E. Schrodinger, *What is Life?* Cambridge, UK: Cambridge University Press, 1944.

[2] J. E. Lovelock, “A physical basis for life detection experiments,” *Nature*, vol. 207, pp. 568–570, 08 1965.

[3] H. Maturana and F. Varela, *Autopoiesis and Cognition: The Realization of the Living*, ser. Boston Studies in the Philosophy and History of Science. Dordrecht, Holland: D. Reidel Publishing Company, 1980.

[4] G. Spencer-Brown, *Laws of Form*. London: George Allen and Unwin Ltd., 1969.

[5] M. Gardner, “The fantastic combinations of john conway’s new solitaire game “life”,” *Scientific American*, vol. 223, pp. 120–123, 1970.

[6] E. R. Berlekamp, J. H. Conway, and R. K. Guy, “Winning ways for your mathematical plays,” *Journal of Multivariate Analysis*, 1982.

[7] R. D. Beer, “Characterizing autopoiesis in the game of life,” *Artificial Life*, vol. 21, no. 1, pp. 1–19, 2015.

[8] F. Varela, *Principles of Biological Autonomy*. 52 Vanderbilt Avenue, New York, New York 10017: Elsevier North Holland, Inc., 1979.

[9] W. Hordijk and M. Steel, “Autocatalytic sets and boundaries,” *Journal of Systems Chemistry*, vol. 6, no. 1, 2015.

[10] B. McMullin, “SCL: An artificial chemistry in Swarm,” *Santa Fe Institute*, 02 1997.

[11] E. Catto, “Iterative dynamics with temporal coherence,” *Game Developer Conference*, 03 2005.

[12] F. Varela, H. Maturana, and R. Uribe, “Autopoiesis: the organization of living systems, its characterization and a model,” *Currents in modern biology*, vol. 5, no. 4, pp. 187–196, may 1974.

[13] T. Toffoli, “Symbol super colliders,” in *Collision-Based Computing*, A. Adamatzky, Ed. London: Springer London, 2002, pp. 1–23.

[14] E. Nelson, *Dynamical theories of Brownian motion*. Princeton, New Jersey: Princeton university press, 1967.

[15] The MathWorks Inc., “Simbiology (r2020a),” Natick, Massachusetts, United States, 2020. [Online]. Available: https://uk.mathworks.com/help/simbio/

[16] W. Ashby, *Design for a brain*, ser. Science Paperbacks. London: John Wiley & Sons Inc., 1960.

Part II.

General Track

NLP Based Framework for Recommending Candidate Ontologies for Reuse

Reham Alharbi
Department of Computer Science
University of Liverpool
Liverpool, UK
R.Alharbi@liverpool.ac.uk

Valentina Tamma
Department of Computer Science
University of Liverpool
Liverpool, UK
V.Tamma@liverpool.ac.uk

Floriana Grasso
Department of Computer Science
University of Liverpool
Liverpool, UK
F.Grasso@liverpool.ac.uk

***Abstract*—Ontology reuse is a complex process that requires the support of methodologies and tools to minimise errors and to keep the ontologies consistent and coherent. Although the vast majority of ontology engineering methodologies include a reuse phase, and reuse has been investigated for different tasks and purposes (e.g. ontology integration), this body of work does not seem to translate into practice, neither in the form of strict criteria for reuse nor as a set of community proposed guidelines. This study proposes a novel process, to be embedded in existing ontology engineering methodologies, that determines a set of candidate ontologies to be reused based on the similarity between the requirements of the new ontology being developed and those of an existing set of potential candidates. We evaluate the accuracy of our method with a case study of developing an ontology for a role game. The preliminary results of this work confirm that the proposed approach can identify ontologies with similar requirements to those of an ontology under development.**

***Index Terms*—Ontology reuse and development, Requirement engineering**

I. INTRODUCTION

A cornerstone of many ontology engineering methodologies is the reuse of ontologies in order to reduce development time and to exploit fragments of knowledge that have been already validated through use [1]–[5]. However, while the research community has extensively studied ontology reuse and contributed several definitions and methodologies, ontology reuse is not a widespread practice, neither in general nor within a given domain [6], [7]. For example, an analysis of the biomedical ontologies in Bioportal concluded that the reuse of ontologies is quite limited (less than 5%) [8].

The task of developing an ontology through reuse, however, is complex, requiring expertise. A major challenge is posed by the choice of a suitable ontology to reuse. Recent studies confirmed that this choice is highly subjective, based on the experience of ontology engineers who typically select ontologies that intuitively fit the purpose, with no effective solutions to support this complex decision-making [6], [9].

Such case-by-case, often biased, approach to the task is not conducive to the establishment of rules and protocols that can lead to a 'shareable good practice' [6], [9].

This is in contrast with other disciplines, such as software engineering, where, for a similar need to reuse artefacts, tools were developed to support this decision-making process within large information spaces [10]. In particular, *Recommender Systems* have been used as an information filtering mechanism that guides software developers in identifying relevant code fragments out of a large dynamically generated corpus, on the basis of the developer preferences [11], [12]. Several applications of this approach have been proposed in the literature, especially to support Requirements Engineering (RE) activities: e.g. stakeholder recommendation for requirements discussions [13], refactoring recommendations based on feature requests [14] and requirement retrieval [10], [15].

Of particular interest is the requirement retrieval scenario, where the analyst, given a newly identified requirement, searches a corpus of requirements to find those that are similar to it, in order to adopt previously developed models and implementations [16], [17]. For this task, content-based recommender systems proved successful in returning recommendations of existing requirements from a historical database of product releases, together with the associated artefacts [18]. The approach is based on the premise that there is a correlation between requirements similarity and software similarity, therefore similar requirements can be used as proxies to retrieve similar software fragments using Natural Language Processing (NLP) techniques [16].

While we acknowledge that reusing software and reusing ontologies are inherently different tasks, we argue that there are lessons to be learned from software reuse practices. Therefore, in this paper we propose a novel process to select ontologies to reuse, based on assessing the similarity between requirements, that is partially inspired by the use of recommender systems in software engineering.

Specifically, we augment the ontology reuse phase, as included in most of the main methodologies for developing ontologies, with a new step: *Reuse Recommendations*. The input to this step is the ontology artefacts that are produced from the *"Ontology Requirements Specification"* phase, which is the phase of collecting the requirements that the ontology should fulfil [19].[1] As output, *Reuse Recommendations* generates a list of candidate ontologies to reuse based on the similarity between the requirements identified for the

---

[1] We refer to the name of phases and activities in the Neon methodology [19] for those activities that are common across ontology development methodologies.

ontology being developed and those in an existing corpus of requirements used in the development of other ontologies. The analysis of ontological requirements has been used in several tasks, such as to improve ontology testing, to improve the requirements specification activity, and to define patterns that implement such requirements [20]. However, to the best of our knowledge, this is the first study that analyses the linguistic aspects of requirements to support and facilitate reuse.

In the proposed approach we analyse requirements through a quantitative lens and investigate if an association can be identified between requirements similarity and candidate ontologies for reuse so that similar requirements can be used as an indicator for the suitability of reuse. We, therefore, proceed to answer the research question *“To what extent the similarity between requirements indicate the reusability of an ontology?”* The paper describes the workflow of the methodology and presents the preliminary results of evaluating the accuracy of the recommendations with respect to a manually created benchmark. Thus, the contribution of this paper is two-fold: (i) providing developers with a workflow for analysing requirements based on NLP techniques, (ii) establishing requirement similarity as an indicator of the reusability of ontologies or ontological fragments that have been independently validated through use.

The paper is organised as follows: after framing our research with a discussion on the background in section II, we present the proposed method in Section III. Section IV presents the case study and shows how the approach is used in practice. Section V offers a discussion of such results. Conclusion and future trends are outlined in Section VI.

II. BACKGROUND

Ontology reuse is an established and extensively studied activity in ontology engineering [7]. Recent studies have focused on how ontology developers approach this activity: Carriero and colleagues survey the different approaches to ontology reuse with the aim of identifying their motivations, strategies, benefits and limits. In their study, they propose two prototypical use cases and analyse them with respect to their benefits and limitations [6]. They conclude that there are no effective solutions for supporting the decision-making process behind the choice of an ontology reuse strategy. Similarly, Alharbi and colleagues investigate the gap between the theory and practice of ontology reuse by consulting directly ontology developers [9]. They investigate ontology reuse practices in relation to the expertise of the developers and conclude that the level of expertise plays a significant role in selecting an ontology to reuse.

The analysis of requirements is also a critical activity in many of the ontology engineering methodologies that involve a reuse phase: e.g. METHONTOLOGY [21], NeON [19], eXtreme Design methodology (XD) [2], the Modular Ontology Modeling methodology [22] and, recently, the LOT methodology [23]. These methodologies typically support the reuse either of terms in the ontology (including classes, attributes, and relations) or of patterns modelling some associated requirements. However, this support typically differs between methodologies (e.g. some support hard reuse, whereas others provide guidelines for hard and soft reuse [7]) and there are very few tools that provide at least semi-automatic reuse functionalities, with the notable exception of ROBOT [24], [25]. This is in contrast to what happens in other disciplines, e.g. software engineering, where code reuse is an established practice supported by different methodologies and editing and versioning tools (e.g. GitHub).

Content-based recommender systems, e.g. those based on clustering and similarity assessment, have been used in software engineering to recommend the reuse of existing artefacts, dependencies, or traceability links [12], [26]. Various applications of recommender systems for requirement engineering have been proposed in the literature, such as requirements retrieval [10], [15], where the vector-space model and cosine similarity are used to assist retrieval of existing requirements on a large industrial data set. The OpenReq EU project’s main objective is the development of recommendation and decision technologies that support requirement engineering tasks in large and distributed software projects [27].[2] As part of its deliverables, the project has released a specific service for similarity computation among requirements based on the *tf-idf* metrics. The underlying hypothesis in these approaches is that artefacts associated to similar requirements are themselves similar and can be reused; for instance, if the descriptions for products A and B are similar, a software implementation of A can be used for B and vice versa. The study in [16] empirically proved a positive correlation between requirement and software similarity.

In order to compute the similarity between requirements, first, we need to decide on a language model. This is a statistical representation of the frequency and relationship between words in a language. In more detail, a requirement will be converted into a numerical vector, and then the similarity is determined by measuring the distance between vectors using the cosine similarity, which measures the cosine of the angle between the vectors [28], [29]. Therefore, the accuracy of the similarity computation is heavily reliant on the language model used to calculate feature vectors [16].

There are a number of state of the art language models, ranging from simple term frequency-based approaches to more advanced that generate neural network-based embedding. In this study, we consider Deep Learning (DL)-based models to semantically represent the requirements and compute their similarity. The application of these models is gaining traction in software engineering, where they are used to provide a contextual representation of expressions through deep learning architectures [16]. One of the most popular language models is the Bidirectional Encoder Representations from Transformers (BERT) [30]. BERT is a pre-trained bidirectional encoder-based transformer model that considers positional and contextual information of words and is trained on very

---

[2] https://openreq.eu/

large corpora in English (e.g English Wikipedia with 2500M words). The BERT model takes as input a text sequence and breaks it down into token embedding, sentence embedding, and positional embedding. Sentence embeddings can be used to derive meaning-rich representation vectors that capture the semantics of requirements, and has shown promising results in different software engineering tasks [16], [31], [32].

The choice of language model together with the choice of effective similarity and relatedness measures heavily influences the performance of the recommenders.

III. APPROACH

Existing ontology development methodologies, for example METHONTOLOGY [21] and LOT [23] involve a requirement gathering and specification phase and an ontology reuse phase; however, typically the requirements gathered in order to scope a new ontology are not exploited to guide the possible reuse of existing ontologies. As identified in [9], the decision of whether to reuse an existing ontology is a subjective task that is heavily dependent on the expertise of the ontology engineer. We propose to augment ontology development methodologies with a new phase, *Reuse Recommendation* that recommends candidate ontologies to reuse on the grounds of the similarity between their requirements and those of the ontology under development.

Figure 1 illustrates this phase as an extension of the ontology implementation phase in the LOT methodology [3] However, this phase can be included in any methodology since it uses artefacts that are commonly generated by all the methodologies. The Reuse Recommendation phase follows the requirement specification in the development process and involves *ontology developers (ontology engineers)* and *domain experts* as actors. This phase takes as input the ontology artefacts produced during the requirement gathering and specification phase of the ontology development process and a set of ontologies that are potential candidates for reuse, together with their requirements. The output is a list of recommended ontologies to reuse based on their requirements. The set of potential candidates can be determined manually by the ontology engineers, but we also envisage constructing this set semi-automatically by processing ontologies in domain-specific repositories such as Bioportal [4].

The recommendation of ontologies to reuse is based on the assessment of the pairwise semantic similarity and relatedness between the requirements of each of the ontologies in the set of potential candidates for reuse (Existing Requirements – ERs) and the requirements of the ontology being developed (Developer Requirements – DRs). By semantic relatedness, we refer to the degree to which two concepts are related with respect to some semantic network [33].

The NLP-based framework workflow for determining the recommendation is illustrated in Figure 2. It starts by encoding both the ERs and DRs using BERT and then by computing their spatial distance using the cosine similarity between the two semantic embedding vectors in a low dimensional space. Only those requirements whose similarity is above a given threshold are retained and passed to the next phase. Each DR might have either zero or several ERs that are above the threshold. This happens when there are a number of candidate ontologies for reuse whose scopes partially overlap, and this affects the similarity scores.

The pairs (DRs, ERs) whose similarity is above the threshold are passed to the next step, POS tagging, to determine the entities and predicates that are used to express the requirements and whose choice determines the context. Following the POS tagging step we compute the relatedness between the chunks, and we then discard those requirements that are irrelevant. Finally, we ask the ontology engineer to analyse, with the support of the domain expert, the ontology modelling the ERs, thus validating the suggestion and then adding it to the ranked list of the candidate ontologies for reuse, which is the output of the Reuse Recommendation phase.

IV. WALK THROUGH SCENARIO

We designed our walkthrough scenario following the guidelines in [34] for conducting and reporting case studies. The study is designed to collect quantitative data that can support the hypothesis we want to test: i.e. the selection of ontologies to reuse can be based on assessing the similarity between their requirements and the ones of the ontology under development. Our systematic approach can help the developers identify not only obvious candidate ontologies for reuse, but also those that, whilst covering different domains, should still be considered for reuse, because they provide terms or patterns whose semantics are similar to those in the domain of interest.

In this walkthrough, we model an ontology for a domain of interest, the Dungeons and Dragons (DnD) role game [5], for which we gather requirements and competency questions. We then apply the Reuse Recommendation process described in Section III.

A. Domain Overview for the Case Study

In our case study, the target ontology required is a model of Dungeons and Dragons[6], and in particular, we narrow the DnD ontology scope to cover only the character features and the game environment. The source of external requirements used in the scenario is the CORAL dataset [20]; we decided to use CORAL because it is one of the richest corpora with a comprehensive collection of CQs and requirements, thus addressing the scarcity of sufficient documentation for published ontologies that includes both CQs and (functional and non-functional) requirements.

DnD is a fantasy role-playing game, played by a number of players, where one adopts the role of the Dungeon Master (DM), who is the lead storyteller and referee of the adventure, and the other players are characters that navigate the adventure's hazards and decide where to explore.

---

[3] We refer to the LOT's activities phase for simplicity of reporting.

[4] https://bioportal.bioontology.org

[5] https://dnd.wizards.com/what-is-dnd/basic-rules

[6] https://dnd.wizards.com/what-is-dnd/basic-rules

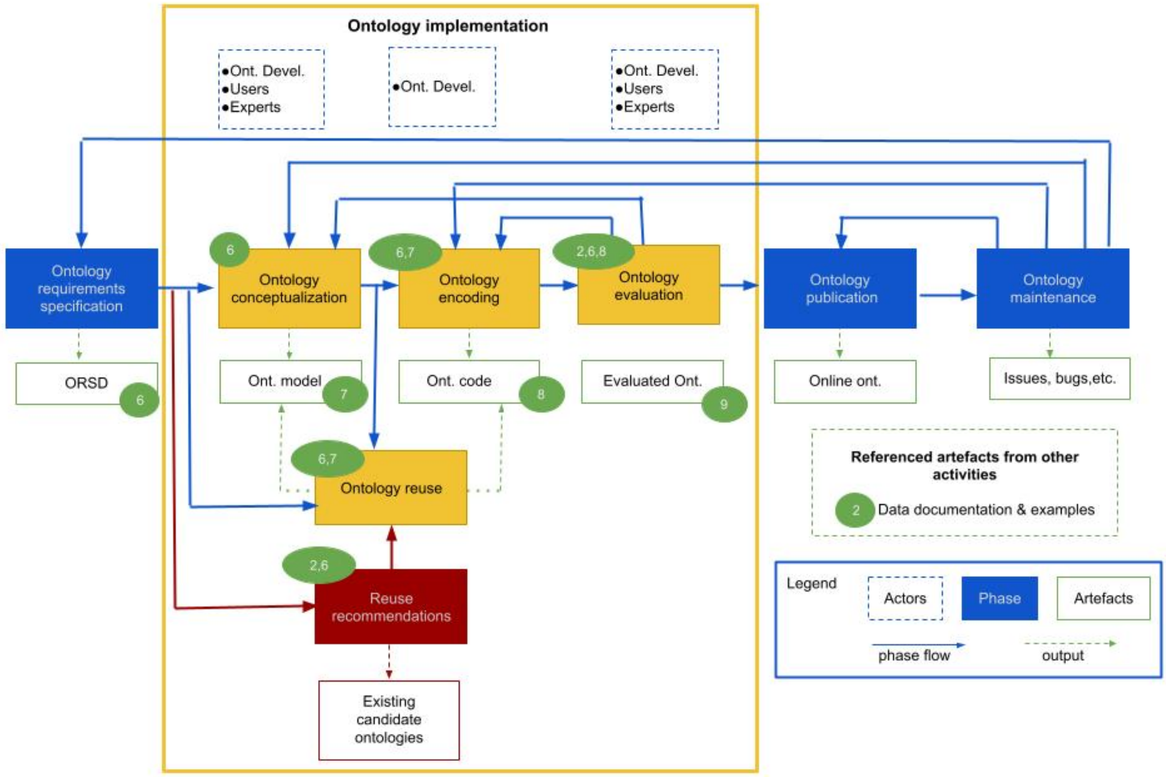


Fig. 1. Reuse recommendation phase (burgundy) as an extension of the LOT methodology.

DnD gameplay is defined by several components: the characters, the rules, the nature of magic, and the DM tools. Each adventurer starts by creating a character by selecting race, class, background, equipment, and other customisation options as outlined in the Player's Handbook [35]. The rules determine the types of dice used to determine success or failure of a character's goal. It also describes the three broad categories of activity: exploration, interaction, and combat.

The DM runs a game based on a description of an adventure that includes information and stat blocks for monsters, advice for building combat encounters, and magic items.

B. Experiment

To collect the requirements for the DnD ontology, we used the official DnD digital game companion, *D&D Beyond* [7], as well as interviews conducted with domain experts to gather more details about different aspects of the game.

We gathered a set of 48 requirements in the form of CQs and statements. Table I presents the first 10 requirements, for illustration purposes. All the requirements and the relevant data used in this walkthrough are available at: https://doi.org/10.5281/zenodo.7454530.

dnd_Identifier	Competency questions/NL sentence
dnd_01	When is the game scheduled to play?
dnd_02	What is a campaign?
dnd_03	What is an adventure?
dnd_04	What is a session?
dnd_05	How long does a campaign last?
dnd_06	How long does a session last?
dnd_07	What is the difficulty of session X?
dnd_08	A campaign has a team
dnd_09	One of the players is the leader
dnd_10	Players have different skills in the campaign

TABLE I: The first 10 requirements

Given a set of Developer Requirements (DRs) and a set of existing ontologies with their requirements (ERs), the Reuse Recommendation phase identifies a subset of similar ERs,

---

[7] https://www.dndbeyond.com/

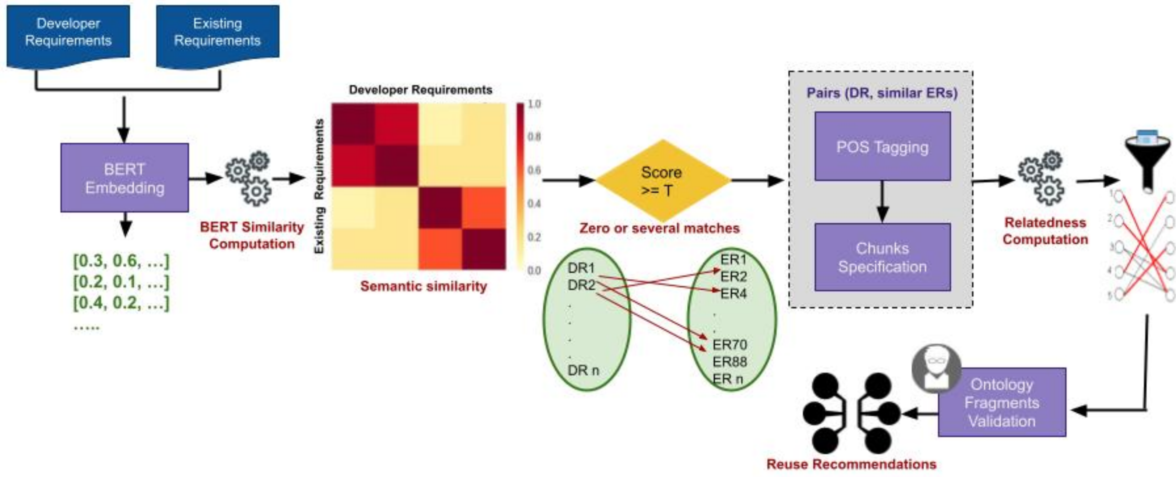


Fig. 2. The workflow of the reuse recommendation method

which then determines the recommendation of the ontologies to reuse. In the remainder of this section, we describe how the approach is applied to the DnD ontology, and we then analyse some preliminary performance results.

In our scenario, the Development Requirements (DRs) are the 48 previously gathered DnD requirements, and the Existing Requirements (ERs) are those in the CORAL dataset [20]. The underlying assumption is that the DnD rules and roles in the games are potentially similar to the requirements in CORAL.

In order to compute the similarity between the DnD requirements and CORAL, we use BERTSimilarity 0.1 [8], a Python library for computing the cosine similarity between the semantic embeddings of each of the DRs against all the ERs. This provided 40,032 pairs of requirements along with their similarity scores.

Following this, the selection of the requirement pairs with similarity above an inclusion threshold, as determined by the ontology engineer. In the course of this walkthrough, we experimented with four inclusion thresholds: 0.5, 0.6, 0.7 and 0.8. We chose these four thresholds to investigate how accurate the proposed approach is, whilst making it more restrictive by requiring a stricter similarity for requirements to be recommended.

The new set of pairs is then subject to the analysis process to identify entities and relationships. We consider the fragments of texts identified by SpaCy [9], a free open-source library for Natural Language Processing in Python, specifically, we only consider nouns and verbs in the tagging. These are the text chunks that are likely to represent entities and relationships in ontologies and we label them as Entity Chunks (ECs) or Predicate Chunks (PCs) depending on whether the chunk computed has a noun as the head of the phrase (ECs) or a verb (PCs) with a possible adposition (in the English language mainly prepositions, it indicates how a sentence should be interpreted in the surrounding context). Each of these chunks is checked for validity, so we reject those ECs that are unlikely to be a class or an individual name, or those PCs that are vague and unlikely to be a relation name.

Word	**Part of Speech**
players	NNS- noun plural
have	VBP- verb, sing. present, non-3d
different	JJ- adjective
skills	NNS- noun plural
in	IN- preposition/subordinating conjunction
the	DT- determiner
campaign	NN- noun, singular

TABLE II: POS tagging for requirement dnd_10

For example, let us consider the requirement dnd_10: ‘Players have different skills in the campaign’, Table II represents the words and corresponding POS tags computed for this requirement [10]. The output from this step is the terms used in the requirement with their corresponding role in the sentence, for each of the requirement in the pairs. These terms are then passed to the next task, that determines the entities and predicates chunks.

Then the relatedness between chunks is computed in order to identify chunks that are relevant wrt the DRs, and determines the relationship between chunks is based on both the background knowledge and the similarity measure. We choose *WordNet* as background knowledge, noting that its primary structure is based on synonymy. Thus, every word has some synset according to the meaning of the word in the context of a statement. The most frequently encoded relation

---

[8] https://pypi.org/project/BERTSimilarity/

[9] https://spacy.io/

[10] We use the POS tags defined by the Penn Treebank [36].

among synsets is the super-subordinate relation (hyperonymy, hyponymy, or ISA relation) [37].

Given two chunks $c_1 \in DRs$ and $c_2 \in ERs$, we find a match if a semantic relation holds between at least one pair of their synsets S_{c_1} and S_{c_2} according to the background dataset A that is used, i.e. if the two corresponding synsets are synonyms, or one is the hypernym (hyponym) of the other.

$$isMatch_A(c_1, c_2) \text{ iff } \begin{cases} isSynonymous_A(S_{c_1}, S_{c_2}) \\ \\ isHypernym_A(S_{c_1}, S_{c_2}) \vee \\ \quad isHypernym_A(S_{c_1}, S_{c_2}) \\ \\ isHyponymy_A(S_{c_1}, S_{c_2}) \vee \\ \quad isHyponymy_A(S_{c_1}, S_{c_2}) \end{cases} \tag{1}$$

In order to determine if a semantic relation holds, we measure the relatedness of corresponding chunks using the *Wu-Palmer* measure, that calculates relatedness based on how similar two synsets (s_1, s_2) are, and where the synsets occur relative to each other in the hypernym tree.

$$Wu - Palmer = \frac{2 * depth(lcs(S_{c_1}, S_{c_2}))}{(depth(S_{c_1}) + depth(S_{c_2}))} \in]0, 1] \tag{2}$$

The chunks produced in the previous step are manually validated by the ontology engineer to verify whether the labels of the ontology elements match the chunks identified in the requirements.

V. DISCUSSION

In order to answer our research question, *"To what extent the similarity between requirements indicate the reusability of an ontology?"* we measure the accuracy of our method against the expert manually created corpus.

This approach mimics a software requirement retrieval scenario, where requirements from DnD are considered queries to be answered against CORAL. As in an actual scenario, requirements that are considered as "most similar" according to a particular metric may not actually be similar in practice from the developer's viewpoint.

We measure the accuracy by computing different inclusion thresholds: 0.5, 0.6, 0.7, and 0.8 and after having completed the matching chunk detection and validation. Table III also reports the number of requirements matched for each inclusion threshold.

Our method recommends 5 ontologies out of the 14 in the CORAL dataset, based on the similarity between requirements. In all but one of these, the inclusion threshold values behave consistently, with the value of 0.5 which does not seem to be sufficiently discriminative, returning more matching requirements than the ones validated by our expert in the validation corpus, and a value between 0.7 and 0.8 returning a number of matched requirements increasingly close to the optimal one in the corpus. The remaining case is worth a more in-depth discussion: for the BTN100 ontology, our approach returns a very high number of false positives (i.e. requirements that are incorrectly identified as matching). It is interesting to note that BTN100 models geographic information about topographic and thematic data about different themes: administrative units, protected zones, buildings and population entities, transport networks, energy and conduction, geodetic vertices, altimetry and hydrography. From this description, it seems evident that only part of this ontology could be reused in the DnD domain. However, the vocabulary of the ontology, which affects the way relatedness is computed, has similarities with the terms used to describe the DnD requirements. This might explain why the number of matched requirements is only marginally affected by the threshold value becoming more stringent.

This example serves as a reminder that the evaluation at this stage is very much affected by the selected similarity measure, and that care must be taken, in applying the methodology described in this paper, to adopt a similarity measure that provides the best impact. For example, recent work has analysed experimentally the impact of background knowledge in evaluating concept similarity [38]. A natural extension of our work would therefore involve a systematic analysis of several similarity approaches to determine the ones that, for a given scenario, are most likely to provide the ontology engineer with the best chance to identify a set of ontologies with the right structure, because they answer competency questions or address requirements that are similar to those defined for the ontology being written. However, we note that such a systematic evaluation would be greatly facilitated by a fundamental shift in ontology documentation practices, mandating that CQs and requirements are always made publicly available.

VI. CONCLUSION AND FUTURE WORK

We proposed a process to recommend candidate ontologies for reuse when developing a new ontology. This recommended set of ontologies is determined by assessing the similarity (including relatedness) of their requirement with respect to those gathered by an ontology engineer for a new ontology to develop. The methodology is general and can be seen as an additional step in many of the current ontology development methodologies. The methodology is also independent of the choice of specific similarity and relatedness measure implementation tools. the preliminary results seem to support the premise of this study, i.e. that evaluating the similarity between requirements can be an effective proxy for selecting ontologies to reuse during the development process, thus reducing the time an ontology developer would need to spend analysing requirements and code of candidate ontologies to reuse. We aim to evaluate the methodology in real-case scenarios, and with a systematic evaluation, thus gaining valuable feedback for future refinements.

Ontology name	**No. of Req.**	**No. of matched Req. on:**				
		Benchmark	**0.5**	**0.6**	**0.7**	**0.8**
Video Game	66	15	17	17	15	13
Building Topology Ontology	18	4	4	4	4	3
BTN100	77	10	65	63	63	56
WoT VICINITY	24	8	17	15	12	6
VICINITY Core	127	5	19	16	8	4

TABLE III: The candidate ontology for reuse with the number of matched requirements for different inclusion thresholds, and those in the validation corpus.

References

[1] T. R. Gruber, “A translation approach to portable ontology specifications,” *Knowledge Acquisition*, vol. 5, no. 2, pp. 199–220, 1993.

[2] V. Presutti, E. Daga, A. Gangemi, and E. Blomqvist, “Extreme design with content ontology design patterns,” in *Proceedings of the 2009 International Conference on Ontology Patterns*, ser. WOP’09, vol. 516. CEUR-WS.org, 2009, p. 83–97.

[3] N. F. Noy and D. L. McGuinness, “Ontology development 101: A guide to creating your first ontology,” Stanford knowledge systems laboratory, Tech. Rep. KSL-01-05, 2001.

[4] R. Neches, R. E. Fikes, T. Finin, T. Gruber, R. Patil, T. Senator, and W. R. Swartout, “Enabling technology for knowledge sharing,” *AI magazine*, pp. 36–56, 1991.

[5] R. Studer, V. R. Benjamins, and D. Fensel, “Knowledge engineering: Principles and methods,” *Data & Knowledge Engineering*, vol. 25, no. 1-2, pp. 161–197, 1998.

[6] V. A. Carriero, M. Daquino, A. Gangemi, A. G. Nuzzolese, S. Peroni, V. Presutti, and F. Tomasi, “The landscape of ontology reuse approaches,” in *Applications and Practices in Ontology Design, Extraction, and Reasoning*, G. Cota, M. Daquino, and G. Pozzato, Eds. IOS Press, 2020, vol. 49, pp. 21–38.

[7] M. Fernández-López, M. Poveda-Villalón, M. C. Suárez-Figueroa, and A. Gómez-Pérez, “Why are ontologies not reused across the same domain?” *Journal of Web Semantics*, vol. 57, p. 100492, 2019.

[8] G.-Q. Zhang, M. R. Kamdar, T. Tudorache, and M. A. Musen, “A systematic analysis of term reuse and term overlap across biomedical ontologies,” *Semant. Web*, vol. 8, no. 6, p. 853–871, 2017.

[9] R. Alharbi, V. Tamma, and F. Grasso, “Characterising the gap between theory and practice of ontology reuse,” in *K-CAP ’21: Proceedings of the 11th Knowledge Capture Conference*. Association for Computing Machinery, 2021, p. 217–224.

[10] M. Irshad, K. Petersen, and S. Poulding, “A systematic literature review of software requirements reuse approaches,” *Information and Software Technology*, vol. 93, pp. 223–245, 2018.

[11] F. Isinkaye, Y. Folajimi, and B. Ojokoh, “Recommendation systems: Principles, methods and evaluation,” *Egyptian Informatics Journal*, vol. 16, no. 3, pp. 261–273, 2015.

[12] A. Felfernig, G. Ninaus, H. Grabner, F. Reinfrank, L. Weninger, D. Pagano, and W. Maalej, “An overview of recommender systems in requirements engineering,” in *Managing Requirements Knowledge*, W. Maalej and A. K. Thurimella, Eds. Springer Berlin Heidelberg, 2013, pp. 315–332.

[13] C. Castro-Herrera, J. Cleland-Huang, and B. Mobasher, “Enhancing stakeholder profiles to improve recommendations in online requirements elicitation,” in *2009 17th IEEE International Requirements Engineering Conference*, 2009, pp. 37–46.

[14] A. S. Nyamawe, H. Liu, N. Niu, Q. Umer, and Z. Niu, “Automated recommendation of software refactorings based on feature requests,” in *2019 IEEE 27th International Requirements Engineering Conference (RE)*, 2019, pp. 187–198.

[15] J. Natt och Dag, B. Regnell, V. Gervasi, and S. Brinkkemper, “A linguistic-engineering approach to large-scale requirements management,” *IEEE Software*, vol. 22, no. 1, pp. 32–39, 2005.

[16] M. Abbas, A. Ferrari, A. Shatnawi, E. Enoiu, M. Saadatmand, and D. Sundmark, “On the relationship between similar requirements and similar software,” *Requirements Engineering*, vol. 28, pp. 23–47, 2023.

[17] A. Shatnawi, A.-D. Seriai, and H. Sahraoui, “Recovering software product line architecture of a family of object-oriented product variants,” *Journal of Systems and Software*, vol. 131, pp. 325–346, 2017.

[18] P. Lops, M. de Gemmis, and G. Semeraro, “Content-based recommender systems: State of the art and trends,” in *Recommender Systems Handbook*, F. Ricci, L. Rokach, B. Shapira, and P. B. Kantor, Eds. Springer US, 2011, pp. 73–105.

[19] M. C. Suárez-Figueroa, A. Gómez-Pérez, and M. Fernández-López, “The NeOn Methodology framework: A scenario-based methodology for ontology development,” *Applied ontology*, vol. 10, no. 2, pp. 107–145, 2015.

[20] A. Fernández-Izquierdo, M. Poveda-Villalón, and R. García-Castro, “Coral: A corpus of ontological requirements annotated with lexico-syntactic patterns,” in *The Semantic Web*, P. Hitzler, M. Fernández, K. Janowicz, A. Zaveri, A. J. Gray, V. Lopez, A. Haller, and K. Hammar, Eds. Springer International Publishing, 2019, pp. 443–458.

[21] M. Fernández-López, A. Gómez-Pérez, and N. Juristo, “Methontology: From ontological art towards ontological engineering,” in *AAAI Conference on Artificial Intelligence*. American Association for Artificial Intelligence, 1997, pp. 33–40.

[22] C. Shimizu, K. Hammar, and P. Hitzler, “Modular ontology modeling,” *Semantic Web*, vol. Pre-Press, no. Pre-Press, pp. 1–31, 2022.

[23] M. Poveda-Villalón, A. Fernández-Izquierdo, M. Fernández-López, and R. García-Castro, “Lot: An industrial oriented ontology engineering framework,” *Engineering Applications of Artificial Intelligence*, vol. 111, p. 104755, 2022.

[24] J. A. Overton, H. Dietze, S. Essaid, D. Osumi-Sutherland, and C. J. Mungall, “ROBOT: A command-line tool for ontology development,” in *International Conference on Biomedical Ontology*, ser. CEUR Workshop Proceedings, F. Couto and J. Hastings, Eds., no. 1515, 2015.

[25] R. C. Jackson, J. P. Balhoff, E. Douglass, N. L. Harris, C. J. Mungall, and J. A. Overton, “ROBOT: A Tool for Automating Ontology Workflows,” *BMC Bioinformatics*, vol. 20, no. 1, 7 2019.

[26] A. Danylenko and W. Löwe, “Context-aware recommender systems for non-functional requirements,” in *2012 Third International Workshop on Recommendation Systems for Software Engineering (RSSE)*. IEEE, 2012, pp. 80–84.

[27] A. Felfernig, M. Stettinger, A. Falkner, X. Franch, and C. Palomares, “Openreq: Recommender systems in requirements engineering,” in *Proceedings of the Workshop Papers of I-Know 2017 (I-Know 2017)*, ser. CEUR Workshop Proceedings, M. Ähnelt *et al.*, Eds., vol. 2025, 2017, pp. 1–4.

[28] C. D. Manning, P. Raghavan, and H. Schütze, *Introduction to Information Retrieval*. Cambridge University Press, 2008.

[29] J. M. Ponte and W. B. Croft, “A language modeling approach to information retrieval,” in *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, ser. SIGIR ’98. Association for Computing Machinery, 1998, p. 275–281.

[30] J. Devlin, M. Chang, K. Lee, and 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, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, J. Burstein, C. Doran, and T. Solorio, Eds. Association for Computational Linguistics, 2019, pp. 4171–4186.

[31] J. Lin, Y. Liu, Q. Zeng, M. Jiang, and J. Cleland-Huang, “Traceability transformed: Generating more accurate links with pre-trained bert models,” in *2021 IEEE/ACM 43rd International Conference on Software Engineering (ICSE)*, vol. 43. IEEE, 2021, pp. 324–335.

[32] H. Guo and M. P. Singh, “Caspar: Extracting and synthesizing user stories of problems from app reviews,” in *2020 IEEE/ACM 42nd International Conference on Software Engineering (ICSE)*, 2020, pp. 628–640.

[33] A. Budanitsky, “Lexical Semantic Relatedness and its Application in Natural Language Processing,” Department of Computer Science, University of Toronto, Tech. Rep. CSRG-390, Aug. 1999.

[34] P. Runeson and M. Höst, “Guidelines for conducting and reporting case study research in software engineering,” *Empirical software engineering*, vol. 14, pp. 131–164, 2009.

[35] D. Dragons, *Dungeons & Dragons Player’s Handbook (Core Rulebook, D&D Roleplaying Game)*, ser. Dungeons & Dragons. Wizards of the Coast Publishing, 2014. [Online]. Available: https://books.google.co.uk/books?id=TARPEAAAQBAJ

[36] M. P. Marcus, M. A. Marcinkiewicz, and B. Santorini, “Building a large annotated corpus of English: The Penn treebank,” *Computational Linguistics*, vol. 19, no. 2, p. 313–330, 1993.

[37] G. A. Miller, “Wordnet: A lexical database for english,” *Communications of the ACM*, vol. 38, no. 11, p. 39–41, 1995.

[38] J. Portisch, M. Hladik, and H. Paulheim, “Background knowledge in schema matching: Strategy vs. data,” in *The Semantic Web – ISWC 2021*, A. Hotho, E. Blomqvist, S. Dietze, A. Fokoue, Y. Ding, P. Barnaghi, A. Haller, M. Dragoni, and H. Alani, Eds. Springer International Publishing, 2021, pp. 287–303.

Trust in Cognitive Models: Understandability and Computational Reliabilism

Noman Javed
Centre for Philosophy of Natural and Social Science
London School of Economics and Political Science
London, UK
n.javed3@lse.ac.uk

Angelo Pirrone
Centre for Philosophy of Natural and Social Science
London School of Economics and Political Science
London, UK
a.pirrone@lse.ac.uk

Laura Bartlett
Centre for Philosophy of Natural and Social Science
London School of Economics and Political Science
London, UK
l.bartlett@lse.ac.uk

Peter Lane
School of Physics, Engineering and Computer Science
University of Hertfordshire
Hatfield, UK
p.c.lane@herts.ac.uk

Fernand Gobet
Centre for Philosophy of Natural and Social Science
London School of Economics and Political Science
London, UK
f.gobet@lse.ac.uk

***Abstract*—The realm of knowledge production, once considered a solely human endeavour, has transformed with the rising prominence of artificial intelligence. AI not only generates new forms of knowledge but also plays a substantial role in scientific discovery. This development raises a fundamental question: can we trust knowledge generated by AI systems? Cognitive modelling, a field at the intersection between psychology and computer science that aims to comprehend human behaviour under various experimental conditions, underscores the importance of trust. To address this concern, we identified understandability and computational reliabilism as two essential aspects of trustworthiness in cognitive modelling. This paper delves into both dimensions of trust, taking as case study a system for semi-automatically generating cognitive models. These models evolved interactively as computer programs using genetic programming. The selection of genetic programming, coupled with simplification algorithms, aims to create understandable cognitive models. To discuss reliability, we adopted computational reliabilism and demonstrate how our test-driven software development methodology instils reliability in the model generation process and the models themselves.**

***Index Terms*—Trust, Computational reliabilism, Understandability, Genetic Programming**

I. INTRODUCTION

Collaboration between humans and AI has significantly advanced various fields, particularly science [1] [2]. As a result, there is growing interest in accelerating scientific research by automating certain aspects of the process, increasing AI involvement and reducing human intervention [3]. However, this raises a critical issue regarding the trustworthiness of AI systems, as there have been cases where relying on AI has led to undesirable outcomes [4]. A key factor contributing to this mistrust is the "black box" nature of some AI technologies, which are often complex and difficult to comprehend, limiting transparency and the ability to validate their results. This, in turn, raises concerns about the reliability of AI-generated knowledge.

Addressing the trustworthiness concerns associated with "black-box" systems necessitates a thorough examination of their internal mechanisms. Potential solutions include the use of transparent white-box techniques, the development of explainable AI models, the application of rigorous verification and validation processes, and the establishment of interactive systems that combine human expertise with AI capabilities. By addressing these trust-related challenges, the full potential of AI in scientific discovery can be harnessed, ultimately resulting in the generation of reliable and accurate knowledge.

We are constructing a semi-autonomous system that generates cognitive models in the form of computer programs that represent scientific theories [5] [6]. As these models represent the knowledge produced by our system, the question of trust can be applied to them. The ultimate goal is to create a system that produces trustworthy knowledge; thus, it is essential to examine the factors contributing to the trustworthiness of these models. We identified understandability and reliability as the two key aspects of trust. This paper aims to justify and ground our choices and strategies by exploring the question of trust in terms of these two aspects.

Understandability refers to the ease with which humans can comprehend the inner workings of cognitive models. When a model is comprehensible to domain experts and stakeholders, it promotes transparency and fosters trust in the system. In order to explain how our system enhances understandability, we will provide a rationale for the selection of its basic components and the array of techniques implemented for post-processing the generated models. The main idea involves employing white-box approaches to generate cognitive models and subsequently refining them to enhance their understandability.

Reliability refers to the consistency and accuracy of cognitive models in generating valid, empirically verified results, which is crucial for building trust. In this paper, we embrace computational reliabilism [7] as the philosophical foundation for reliability. This approach proposes that knowledge is deemed valid when produced by reliable processes. In other words, the credibility of the knowledge depends on the consistency and precision of the underlying methods, techniques, or models used in its creation. Thus, it is essential to examine and improve these processes, verifying their efficiency and capability to generate accurate and consistent results. This ultimately enhances the credibility of the knowledge, promoting trust among stakeholders. As the nature of our system's is computational modelling, we aim to explore the software development methodology employed in its construction and evaluate the results it generates. We will provide an in-depth examination of how these software engineering practices correspond to the sources of reliability identified in computational reliabilism.

This paper investigates the factors influencing trust in the context of understandability and reliability, aiming to establish a robust foundation for developing a dependable system that generates cognitive models representing scientific theories. The following section outlines our system's fundamental components and operation. Section III addresses the issue of understandability and discusses how our design choices and strategies can improve the comprehensibility of the generated models. Section IV tackles the reliability concern by initially presenting it within the framework of computational reliabilism. Subsequently, it elucidates how our system development methodology can contribute to enhanced reliability. In the discussion section, we address a couple of significant points before concluding the article.

II. Genetically Evolving Models of Science

Most theories in psychology are expressed informally, using natural language. However, this type of representation suffers from a number of weaknesses: due to the lack of precision of natural language, it is hard to make clear-cut predictions; verbal theories do not really explain empirical data, but are just a redescription of them; and they cannot be tested, because the lack of specification offers an indefinite number of possible interpretations of any verbal theory.

By contrast, computational models – models implemented as computer programs – offer a number of strengths [8], [9]: the theoretical mechanisms and parameters must be rigorously and unambiguously specified, with the consequence that testable predictions can be derived; behaviours can be simulated, even when multiple variables are implicated; it is possible to systematically manipulate the mechanisms and parameters of a model, on the one hand, and the features and statistics of the environment, on the other hand, in order to ascertain how they affect behaviour; finally, it is possible to manipulate not only numerical parameters, such as learning rates, but also qualitative parameters such as heuristics.

In spite of these advantages, one serious drawback of computational modelling is that it is difficult and time consuming to develop cognitive models. To alleviate this difficulty, one approach is to develop algorithms that write the code of computer programs. This is the approach followed by our group, building on previous work [5] [10]. The idea is to (semi-) automatically develop models in psychology, giving as input a description of the experiment to simulate, the experimental results, and a list of theoretical constructs that can be used to build theories.

A. Model Generation System

In order to address the cognitive limitations (such as cognitive biases) that hinder scientists' pursuit of successful solutions and to implement the ideas mentioned earlier, we suggest a meta-modelling system that creates cognitive models in the form of computer programs. As our system generates computer programs, it can be considered a program synthesis system. The figure 1 outlines the system's basic components. Before model generation, a human expert must set up the system by supplying several elements:

Set of operators: Basic operators implemented as programming functions. These operators serve as the fundamental building blocks of the models and are arbitrarily combined to create a model.

Experimental data: Data used in real experiments, typically percentage correct and response times.

Set of initial models: Models already developed by researchers or human experts who believe they are suitable candidates for the target theory.

Model evaluation environment: Human experts must implement an experimental protocol to simulate the actual experiment. Models at this stage represent human participants and face identical experimental conditions. They receive real experimental data, and their responses and timings are recorded. The main idea is to closely simulate the actual experimental conditions.

GP-Algorithm: Serving as the model generation engine, it uses the set of operators and initial models to create a population of models through random combinations. At this stage, models are hypotheses subject to experimental evaluation. The advantage of using GP is its ability to perform multiple parallel searches within the search space. It gradually evolves models using genetic operators, such as crossover and mutation, following natural selection principles. The main difference between our approach and others is how we evaluate the fitness of our models. As mentioned earlier, models undergo experimental evaluation within the model evaluation environment, simulating real experimental conditions.

Phased-evolution Occasionally in our experiments, the model's fitness consists of several components, making it challenging to optimize them collectively. There are different methods to address these issues. We proposed a phased-evolution system that breaks down the initial fitness criterion into distinct components. In the first phase, GP aims to optimize one component, while in the subsequent phases, it optimizes the first and second components, and so forth. This part is not depicted separately in the figure, as it is not a separate part, but rather a modification of the traditional GP system.

Post-processing: This step occurs at the end of the evolutionary run to convert candidate models into a more compact and understandable form.

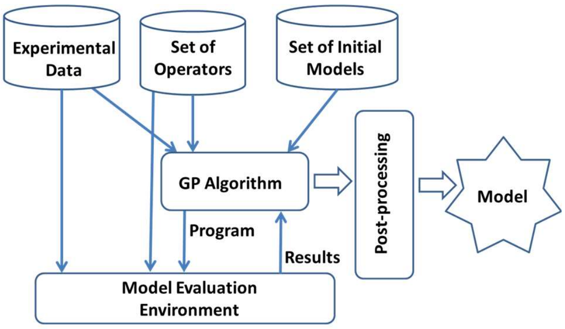


Fig. 1. Architecture of our system

B. Model generation process

To initiate the process of model generation, human experts must first provide the system with several fundamental components. This involves implementing a set of operators and defining their interpretation rules for the embedded interpreter. While some basic cognitive operators are already implemented with their interpretation rules, human experts may need to supply definitions and rules for any additional operators. Next, experts must create a simulated version of the experiment using experimental stimuli. This simulation will be used to calculate model fitness later on.

Once these initial steps are completed, the system is ready for launch, following the primary genetic programming operational procedure. This involves randomly combining the set of operators to generate an initial population of individuals. The fitness of each individual is determined by running the previously defined experiment and comparing the simulated results with the actual human data. Those individuals with higher fitness levels are chosen to breed and produce the next generation. This process continues until a predefined criterion set by the human expert is met or the maximum number of generations is reached. Through this evolutionary process, the fitness of individuals is gradually improved, following the principles of natural selection.

Upon completion of the run, models are post-processed to eliminate dead code and identify similarities among them. These candidate models represent the strategies employed by humans when taking part in the actual experiment.

C. Software Development Methodology

Test-Driven Development (TDD) is a software development approach that prioritises creating tests before developing the actual code. This ensures that the code aligns with the desired requirements, ultimately enhancing the software's overall quality. TDD is an iterative process that begins with the developer writing test cases for a specific functionality before implementing the actual function. The next step involves writing the least amount of code necessary for the test cases to pass. The final step consists of optimising the code and eliminating redundancies, a process known as code refactoring. Kent Beck introduced it in 2003 [11], and since then, it has become a crucial practice within the software development community. TDD supports the agile principles of iterative development, continuous improvement, and high-quality code.

Some of our team members have adapted this approach specifically for developing cognitive architectures [12], as well as other scientific architectures. They have identified three distinct test categories that correspond to various requirements of scientific software [13]. Unit tests ensure the accuracy of individual code segments at the algorithmic level, while process tests, which are implementation-independent, validate the proper functioning of a theoretical process by confirming the generated results. Lastly, canonical tests capture the empirical outcomes accounted for by a model.

We employ their testing framework for our system's development. The first two test types are vital for correctly implementing system components, while the third facilitates the empirical comparison of candidate models generated by our system. We utilise this as a fitness measure for the models, allowing the genetic programming (GP) system to evaluate all models against this metric and evolve the superior ones.

The primary advantages of using this methodology are the clear separation of the implementation level correctness and the correctness of results. Additionally, because the developers have to identify test cases before writing the actual code for a function, it promotes code refactoring while ensuring the correctness of implementation. Another advantage is the clear separation of duties between machines and humans. A final advantage is the clear division of responsibilities between machines and humans. This is in line with the ideas of researchers advocating for interactive machine learning as a means to expedite the process of scientific discovery [14].

III. UNDERSTANDABILITY

Understandability refers to the capacity of an artificial intelligence system to clearly and effectively communicate its decision-making process, reasoning, and results to humans. This idea is closely connected to the concepts of interpretability and explainability in AI, which emphasise the ease with which human users can grasp and interpret the internal mechanisms, models, and choices made by AI systems. Gaining insight into the model's inner workings requires humans to have access to the model's details, which can be challenging or nearly impossible with neural networks and numerous other machine learning techniques. Since our ultimate goal is to create comprehensible and trustworthy models, we chose to employ a transparent algorithm known as genetic programming [15].

A. Genetic Programming

Genetic programming (GP) is an evolutionary algorithm-based approach in which computer programs are evolved to solve a particular problem. It can potentially improve understandability compared to other black-box approaches because of its inherent structure and the way it evolves solutions. Here are some reasons why genetic programming may offer better understandability:

Human-readable representation: GP typically uses tree-based structures to represent candidate solutions, which are often more interpretable than other representations, like neural networks or support vector machines. These tree structures can be easily converted into human-readable formats, such as mathematical expressions, decision trees, or if-then rules, making it easier for humans to understand the underlying logic of the evolved solutions.

Explicit feature selection: Genetic programming often performs feature selection implicitly during the evolution process. It tends to select and use only relevant features while constructing the solution, which can help users identify which features contribute to the final decision. This contrasts with some black-box approaches, like deep learning, where it is difficult to determine the specific contribution of individual features.

Evolvable heuristics: Genetic programming can evolve human-understandable heuristics or rules that can be easily inspected, understood, and modified by humans if necessary. This is in contrast to other black-box approaches, such as deep learning, where the internal logic of the model is not readily accessible or modifiable.

Traceability: The evolutionary process in genetic programming is typically recorded step-by-step, allowing users to trace back the development of the final solution. This traceability provides insights into the evolution of the solution and can help users understand the decision-making process, unlike some black-box approaches where the training process is less transparent.

Although genetic programming might provide improved understandability in comparison to other black-box methods, it is crucial to recognise that the degree of interpretability can vary depending on the particular problem being addressed and the complexity of the developed solution. Sometimes, the complexity escalates due to excessive growth in solution size, a phenomenon referred to as “bloat”. As the size of the solutions expands, even domain experts may struggle to comprehend them. This growth in solution size is frequently associated with an increase in the program’s complexity. We implemented several techniques to simplify complex models into more understandable by reducing bloat.

B. Simplification of genetic programs

Simplification is a widely recognised technique for controlling bloat [16]. Initially, it aimed to decrease the size of final solutions by editing them, which in turn made them more comprehensible. Subsequently, researchers adapted this method in various ways to manage bloat. Some simplification strategies focus on syntactic similarity, while others emphasise semantic equivalence. Despite their differences, all simplification techniques share the common goal of removing non-functional genetic material by editing individuals. Consequently, simplification not only helps control bloat but also eliminates the internal redundancy of individuals. We implemented three simplification techniques, one offline and two online to control bloat and enhance understandability of the models.

1) Offline simplification:: The most basic form of simplification involves detecting and eliminating dead code. Dead code refers to segments of code that are not executed during an individual’s execution. An example of unreachable code includes statements written after a $return$ statement in programming languages such as $C++$, $Java$, and others. To identify these dormant nodes, we ran an experiment on these individuals, marking the unreachable nodes [6]. Subsequently, we reconstructed the same individual after removing these nodes. We carried out this procedure in an offline mode for the post-processing of suitable candidate models. This approach not only diminishes the size of the models but also aids in analysing their similarities and ultimately reducing the overall number of candidate models.

2) Online simplification:: Online refers to the process of employing simplification algorithm while running the evolutionary process. Both of our online algorithms [17] thrive from the inherent working of genetic programming. GP generates individuals through crossover, so they contain the genetic material of both the parents. Some of this genetic material is important and contributing to the better fitness of the individual. However, some of it may be of no use. Thus, there is a possibility of reducing the size of an individual by keeping a copy of good genetic material and removing the one that is not contributing. Both of our algorithms exploit this property of crossover. They do so by unlocking the good genetic material and getting rid of the non-contributing part.

Generation-wide simplification (Gws) takes inspiration from the population-wide bloat control techniques. So, instead of working on just one individual at a time, this method operates at the whole population at once. It halts the routine process of producing the new generation – by selection, crossover, and mutations – after every kth generation. Instead, the application of rupture routine to every individual now generates the new generation. In this way, every individual creates several children. The fitness and hash-values of all these children are then computed and stored. If the computation of hash-value results in a collision, this indicates that an identical individual is already present in the population. So, there is no need for multiple copies of an individual, and hence this child is discarded. This process creates a new population of individuals that may or may not exceed the specified population size. If the number of newly generated individuals is greater than the population size, we can discard the ones with the lowest fitness. Otherwise, we can fill the gap following the routine process of crossover and mutation.

Pruning operates at the individual level rather than at the population level. The motivation behind this is to prune only those individuals who have a high probability of being selected as parents. These are the individuals of relatively higher fitness than the rest of the population. So, the precondition of pruning as an operator is the sorting of the population according to fitness. It takes top $k\%$ individuals from the sorted population as input and ruptures them one by one to generate their children. It then computes the fitness of the children and replaces an individual with its fittest child. However, it only performs the replacement when the child's fitness surpasses the fitness of the parent. So, after completion of this procedure, at most $k\%$individuals in the population give their places to their fittest children. This percentage is called "pruning rate". Since it is a one-to-one replacement algorithm, a very high pruning rate means applying it to most individuals, which is too costly. Therefore, we recommend using a lower rate of pruning. However, this rate can vary from problem to problem, and the best way to determine it is experimental.

IV. RELIABILITY

Reliability is one of several factors contributing to the overall trustworthiness of an AI system. We employed computational reliabilism as the philosophical foundation for reliability to guarantee the production of reliable knowledge. This section begins by defining computational reliabilism, then delves into our software development approach for creating cognitive models, and finally illustrates how this test-driven methodology serves as a practical method for implementing computational reliabilism.

A. Computational reliabilism

In philosophy, knowledge has traditionally been understood as justified true belief. Justification refers to the property of a belief that qualifies it to become knowledge. Edmund Gettier was the first to show that in certain situations, having a justified true belief does not necessarily qualify as knowledge [18]. Reliabilism offers an alternative account to justification, proposing that beliefs formed through a reliable process can be considered knowledge [19]. While there are some differences, most reliabilists argue that a belief can be justified or constitute knowledge even if the believer is unaware of or does not understand the process that makes the belief reliable. This interpretation of reliabilism creates room for the black-box machine learning methods. We will revisit this point later in the discussion.

In the contemporary world, a considerable number of knowledge-generating processes are computational in nature. Recognising this, Durán proposed computational reliabilism [7] as an extension to process reliabilism. This adaptation aims to address the trustworthiness and reliability of knowledge produced through computational methods, reflecting the growing importance of such methods in modern knowledge production. Durán was particularly interested in the trustworthiness of the knowledge produced by computer simulations. Since our work focuses on computational cognitive modelling, we adopt Durán's conceptualisation of computational reliabilism in our system.

Durán defines computational reliabilism in the following terms:

(CR) if S's believing p at t results from m, then S's belief in p at t is justified.

where S is a cognitive agent, p is any truth-valued proposition related to the results of a computer simulation, t is any given time, and m is a reliable computer simulation. Since our systems generates knowledge in the form of cognitive models, our adoption of his definition becomes

If S's believing c at t results from g, then S's belief in c at t is justified.

where S is a cognitive agent, c is the cognitive model generated by the system g, t is any given time, and g is a reliable cognitive model generator.

In contrast to Durán's conception, we are required to guarantee reliability at two distinct stages. First, since our system operates as a model generator, it is essential to verify its reliability. Additionally, as the models generated by our system act as potential candidates for scientific theories, empirical validation of their results is necessary. To accomplish this, we employed a test-driven software development methodology in building our system and subsequently in verifying the results. But before discussing this methodology, we shall first discuss the four sources of reliability identified by Durán.

B. Four sources of reliability

Durán identifies four external sources that can attribute reliability to a computational process, namely verification and validation, robustness analysis, history of (un)successful attempts, and expert knowledge. However, these sources are not of equal importance and their degree of reliability may vary depending on the circumstances, domain, and nature of the computation process. Moreover, their reliability measures are not yet quantified.

Verification and validation: Verification and validation represent two separate processes aimed at guaranteeing the quality, dependability, and accuracy of software. Setting aside their technical definitions and the differing perspectives of the scientific and software communities, verification can be considered the process of ensuring that software is developed correctly, adhering to its design and specifications. On the other hand, validation, particularly in the context of scientific software, concerns the accuracy of the generated results, which is evaluated by comparing them to real experimental outcomes. Both aspects are essential in ensuring the trustworthiness of the knowledge produced.

Robustness analysis: Robustness analysis is a method used in computational modelling and simulation to assess the stability and reliability of a model's results. It consists of two steps: first, examining a group of models to identify a common predicted result, called the robust property; and second, analysing the models to find the structures responsible for generating this robust property. Researchers need to investigate a group of similar, yet distinct models to identify a robust behaviour. The objective of this analysis is to develop a diverse range of models, ensuring that the identification of a robust property is not merely due to chance in the way the models were analysed, but because the property truly exists. By exploring various models, researchers can gain a deeper understanding of the core similarities that yield the robust property and enhance their confidence in the model's predictions and insights.

A history of (un)successful implementations: In line with scientific traditions, documenting the progress of computer simulations' development is crucial, encompassing both successes and failures. By continuously updating methods and techniques and incorporating the latest advances in knowledge, the field gradually progresses. This process helps establish reliability by maintaining a record of accomplishments and challenges, ensuring that researchers can learn from past experiences and build upon them to further enhance the reliability of computer simulations.

Expert knowledge: The fourth and last source is expert knowledge. It is crucial for the reliability of computer simulations, as it influences the underlying theories, assumptions, and computations. Experts from various disciplines contribute to the robustness of a simulation and its history of successful or unsuccessful implementations. Their expertise, acquired through training and social interactions, helps in identifying and judging relevantly similar structures, which is crucial for claims about robust properties. By influencing the assumptions built into the simulation model and determining the acceptable range of results, experts contribute significantly to the overall reliability of computer simulations. However, it is important to remember that expert knowledge can also be a source of errors, necessitating a careful balance between reliance on expertise and other sources of reliability.

C. *Reliability of our System*

So far, we have presented the notion of computational reliabilism and identified its sources. We ensure the reliability of the processes and the results in our system by adopting the software development practices and through various other strategies inbuilt into the design of our system. The following subsections present how our system ensures some form of the four sources of reliability.

Verification and validation: Our system creates computer programs representing candidate models for a theory, and we must verify its correctness at two levels. First, we need to ensure the proper functioning of the system generating the programs. Second, we need to confirm the correctness of the produced programs. To achieve this, we employ an extended version of test-driven development methodology, incorporating three tiers of tests [12] [13]. Unit tests check the implementation correctness, while process tests verify the accuracy of different theoretical processes. The generated programs' correctness is ensured through the genetic programming's generational process, which combines user-defined operators in tree structures, representing the program's syntax. An interpreter within our system interprets these programs. The test-driven methodology guarantees the correctness of generational and interpreter parts and their functioning. However, this does not guarantee that the evolved programs will perform well on real experimental data.

To validate the system's performance, we compare the evolved programs' performance with real experimental data, similar to conducting experiments on the original subjects. The model's performance is assessed by running these experiments. During the evolutionary process, effective programs advance while inferior ones are eliminated. At the end of the evolutionary run, we obtain models that perform on par with real subjects. This empirical testing corresponds to the validation aspect.

Robustness analysis: Robustness analysis involves examining a group of models to identify robust properties. We have incorporated a post-processing step to refine a set of models deemed as suitable candidates. The goal is to eliminate dead code in these models and identify their syntactic similarities. These similar syntactic structures could correspond to robust properties. However, currently, our approach is limited to detecting these similarities and grouping these models into the same family. In the future, we plan to implement additional similarity measures, such as semantic similarity, among others.

A history of (un)successful implementations: In line with Newell's concept of a unified cognitive architecture [20], our system assumes a fixed structure and conducts various experiments across different domains. Our approach is to use parameters supported by research literature and maintain their consistency throughout the experiments. This cross-experimental validation can be assessed using canonical tests. There are several ways to achieve this, reflecting a form of multi-objective optimisation. We have introduced phased evolution [6], which is currently limited to single experiments, but can be extended to multiple experiments.

At the implementation level, the unit testing framework enables continuous code refactoring while ensuring correctness. This continuous refactoring leads to consistent improvement in code quality, effectively preserving the history of code implementations.

V. DISCUSSION

Since our system aims to generate trustworthy knowledge following the account of computational reliabilism, it is crucial to address two key aspects. The first aspect pertains to the nature of the scientific method employed for knowledge generation. The second aspect emphasises the rationale behind distinguishing understandability from computational reliabilism.

A. *Computational reliabilism and falsification*

It is undeniable that AI has become a catalyst for propelling scientific discovery, revolutionising the practice of science in numerous ways. One significant shift encompasses the transition from hypothesis-driven to data-driven science [3]—a subject of ongoing debate with advocates on both sides [21]. Supporters of the hypothesis-driven approach adhere to the traditional Popperian view of science, which centres on falsification as the core of the scientific process. Popper posits that while scientific evidence cannot prove a theory to be true, it can demonstrate its falsehood. If a theory yields a false prediction, the theory itself must also be false. Popper's knowledge construction method relies on a process of elimination, considering surviving theories true until counter-evidence emerges.

Popper does not endorse deriving theories inductively from data. In contrast, many modern machine learning systems are built upon this inductive approach. Reliabilism offers a balanced perspective, embracing the possibility of data-driven theory development without outright dismissal. Unlike Popper, reliabilism permits inductive reasoning, provided reliable falsification mechanisms exist. As a result, it does not reject data-driven machine learning systems as long as reliable refutation opportunities are present. Reliabilism predominantly aligns with traditional hypothesis-driven science, where hypotheses are first formulated and later experimentally refuted using data [22]. Our system aligns well with this traditional account, as it begins by stochastically generating hypotheses. Each hypothesis undergoes experimental testing, and we move forward only with those that account for a relatively larger portion of the data, following the data-driven science approach. This process results in the gradual evolution of hypotheses, ultimately converging on those that fully explain the data.

B. *Decoupling understandability and reliabilism*

Duran's conception of computational reliabilism ascribes credibility to the knowledge generated by black-box machine learning models, offering a means to trust the results produced by these opaque systems. Although we adhere to his formulation of computational reliabilism, we choose to employ more transparent machine learning techniques. This departure from Duran's original formulation is based on several reasons:

Firstly, the main aim of cognitive science involves understanding the cognitive processes and strategies employed by humans in various experimental situations. By choosing genetic programming, we can encode these basic processes as cognitive operators, enabling human experts to consider different cognitive operators and examine their potential combinations to generate higher cognitive functions. Gaining such understanding is not easily achievable with black-box methodologies. Therefore, the primary argument here emphasises the development of scientific understanding, rather than merely obtaining reliable scientific discoveries.

Secondly, the argument is grounded in the critiques of autonomous machine learning. Numerous researchers advocate for interactive machine learning, where humans occupy a central role. Our system is inherently interactive, as it necessitates human involvement in coding basic cognitive operators, creating experimental code based on original protocols, and developing an empirical evaluation of hypotheses through a fitness function. The argument here is that such interaction fosters a better understanding, which is difficult in a fully automated system.

VI. CONCLUSION AND FUTURE WORK

This paper has explored the creation of trustworthy cognitive models by addressing understandability and reliability as two crucial aspects of trust in cognitive modelling. To develop understandable cognitive models, we designed a system for generating these models and explained how their representation as computer programs, the application of genetic programming, simplification, and post-processing contribute to enhanced understandability. To ensure reliable cognitive models, we embraced computational reliabilism as the philosophical foundation and discussed how our agile development methodology, comprising an extended form of test-driven development and continuous refactoring, establishes reliability in the system. This methodology guarantees not only the reliable generation of models but also the validity of these models as candidates for the theory through experimental evaluation.

Our approach strikes a balance between conventional falsifiable theory-driven science and the potential of data-driven methods, as promoted by computational reliabilism. Although the original concept of computational reliabilism was developed to foster trust in opaque yet reliable methods, we argued for transparent approaches to encourage scientific understanding.

This was our first attempt to apply computational reliabilism to cognitive modelling. As computational modelling and simulations share many similarities, we aimed to make minimal modifications to the original formulation. Moving forward, we plan to develop a customised version of computational reliabilism tailored for meta-modelling frameworks. Additionally, we intend to explore how ethical issues related to autonomous AI can become epistemic concerns, potentially creating new trust challenges in computational systems.

ACKNOWLEDGMENT

This article is part of the project "Genetically Evolving Models in Science (GEMS)" that has received funding from the European Research Council (ERC) under the grant agreement no. ERC-2018-ADG-835002.

REFERENCES

[1] Y. Xu, X. Liu, X. Cao, C. Huang, E. Liu, S. Qian, X. Liu, Y. Wu, F. Dong, C.-W. Qiu *et al.*, "Artificial intelligence: A powerful paradigm for scientific research," *The Innovation*, vol. 2, no. 4, p. 100179, 2021.

[2] L. K. Bartlett, A. Pirrone, N. Javed, and F. Gobet, "Computational scientific discovery in psychology," *Perspectives on Psychological Science*, vol. 18, no. 1, pp. 178–189, 2023.

[3] H. Kitano, "Nobel Turing Challenge: creating the engine for scientific discovery," *npj Systems Biology and Applications*, vol. 7, no. 1, p. 29, 2021.

[4] D. Siddarth, D. Acemoglu, D. Allen, K. Crawford, J. Evans, M. Jordan, and E. Weyl, "How ai fails us," *arXiv preprint arXiv:2201.04200*, 2021.

[5] E. Frias-Martinez and F. Gobet, "Automatic generation of cognitive theories using genetic programming," *Minds and Machines*, vol. 17, pp. 287–309, 2007.

[6] P. Lane, L. Bartlett, N. Javed, A. Pirrone, and F. Gobet, “Evolving understandable cognitive models,” in *Proceedings of the 20th International Conference on Cognitive Modelling*, 2022.

[7] J. M. Durán and N. Formanek, “Grounds for trust: Essential epistemic opacity and computational reliabilism,” *Minds and Machines*, vol. 28, pp. 645–666, 2018.

[8] F. Gobet, P. Chassy, and M. Bilalic, *Foundations of cognitive psychology*. McGraw Hill, 2016.

[9] H. A. Simon, *Models of thought*. Yale university press, 1979, vol. 352.

[10] F. Gobet and A. Parker, “Evolving structure-function mappings in cognitive neuroscience using genetic programming,” *Swiss Journal of Psychology*, vol. 64, no. 4, pp. 231–239, 2005.

[11] K. Beck, *Test-driven development: by example*. Addison-Wesley Professional, 2003.

[12] P. C. Lane and F. Gobet, “Developing reproducible and comprehensible computational models,” *Artificial Intelligence*, vol. 144, no. 1-2, pp. 251–263, 2003.

[13] ——, “A theory-driven testing methodology for developing scientific software,” *Journal of Experimental & Theoretical Artificial Intelligence*, vol. 24, no. 4, pp. 421–456, 2012.

[14] O. Mac Aodha, V. Stathopoulos, G. J. Brostow, M. Terry, M. Girolami, and K. E. Jones, “Putting the scientist in the loop–accelerating scientific progress with interactive machine learning,” in *2014 22nd International Conference on Pattern Recognition*. IEEE, 2014, pp. 9–17.

[15] J. R. Koza, *Genetic Programming: On the Programming of Computers by Means of Natural Selection*. Cambridge, MA: MIT Press, 1992.

[16] N. Javed, F. Gobet, and P. Lane, “Simplification of genetic programs: a literature survey,” *Data Mining and Knowledge Discovery*, vol. 36, no. 4, pp. 1279–1300, 2022.

[17] N. Javed and F. Gobet, “On-the-fly simplification of genetic programming models,” in *Proceedings of the 36th annual ACM symposium on applied computing*, 2021, pp. 464–471.

[18] E. L. Gettier, “Is justified true belief knowledge?” *Analysis*, vol. 23, no. 6, pp. 121–123, 1963. [Online]. Available: http://www.jstor.org/stable/3326922

[19] A. Goldman and G. Pappas, “What is justified belief?” *Epistemology: An Anthology*, vol. 2, pp. 333–347, 1979.

[20] A. Newell, *Unified theories of cognition*. Harvard University Press, 1994.

[21] F. Mazzocchi, “Could big data be the end of theory in science? a few remarks on the epistemology of data-driven science,” *EMBO reports*, vol. 16, no. 10, pp. 1250–1255, 2015.

[22] P. Lipton, “Popper and reliabilism,” *Royal Institute of Philosophy Supplements*, vol. 39, pp. 31–43, 1995.

Local Minima Drive Communications in Cooperative Interaction

Roger K. Moore
Dept. Computer Science
University of Sheffield
Sheffield, UK
r.k.moore@sheffield.ac.uk

***Abstract*—An important open question in human-robot interaction (HRI) is precisely *when* an agent should decide to communicate, particularly in a cooperative task. Perceptual Control Theory (PCT) tells us that agents are able to cooperate on a joint task simply by sharing the same 'intention', thereby distributing the effort required to complete the task among the agents. This is even true for agents that do not possess the same abilities, so long as the goal is observable, the combined actions are sufficient to complete the task, and there is no local minimum in the search space. If these conditions hold, then a cooperative task can be accomplished *without* any communication between the contributing agents. However, for tasks that *do* contain local minima, the global solution can only be reached if at least one of the agents adapts its intention at the appropriate moments, and this can only be achieved by appropriately timed communication. In other words, it is hypothesised that in cooperative tasks, the function of communication is to coordinate actions in a complex search space that contains local minima. These principles have been verified in a computer-based simulation environment in which two independent one-dimensional agents are obliged to cooperate in order to solve a two-dimensional path-finding task.**

***Index Terms*—cooperation, communication, interaction, perceptual control theory, search, local minima**

I. INTRODUCTION

An important open question in human-robot interaction (HRI) is precisely *when* an agent should decide to communicate [1]. Unfortunately, research in human-human interaction has been obsessed with 'turn-taking' as the underlying mechanism [2]–[5], somewhat overlooking the observation that conversation can overlap as well as interleave [6], [7], as well as ignoring the question as to *why* agents should communicate in the first place [8]. Clearly, communication supports information exchange [9], [10] and learning [11], but more importantly it facilitates collaborative problem solving [12] and goal sharing [13], i.e. *cooperation*. However, little research has been conducted into what conditions the timing and structure of communication in continuous cooperative interaction [14].

This paper addresses these issues from the perspective of Perceptual Control Theory (PCT) [15]. Results are presented from a PCT-based simulation of a cooperative task, and it is shown how appropriately timed communication between agents can overcome local minima in a joint problem space.

II. COMMUNICATION IN COOPERATION

Perceptual Control Theory (PCT) is founded on the mantra "*behaviour is the control of perception*", and agents are modelled as a hierarchy of negative-feedback control loops. Solidly grounded in the tradition of 'cybernetics' [16], PCT has been shown to be capable of accounting for a wide range of 'intelligent' phenomena based on a parsimonious architecture of replicated closed-loop structures [17]. In particular, PCT tells us that agents are able to cooperate on a joint task simply by sharing the same reference signal, i.e. by having the same *intention* [18]. The consequence is that the effort required to complete a task may be distributed among the agents involved.

However, it is claimed here that successful convergence towards a solution of a joint task is based on three assumptions:

- the goal is observable (that is, each agent has an appropriate input function),
- the combined actions are sufficient to complete the task (that is, the agents possess complimentary output functions), and
- the goal is accessible (that is, there are no *local minima* in the search space).

If these three conditions are met, then a cooperative task can be accomplished *without* any communication between the contributing agents.

This means that, for tasks that *do* have local minima, the global solution can only be reached if at least one of the agents adapts its intention at the appropriate moment(s). That is, an agent may need to abandon its original goal in favour of a temporary alternative that facilitates an escape from a local minimum. Such behaviour requires timely coordination between the agents, and this can only be achieved by appropriately-timed *communication*. In other words, it is hypothesised that, in cooperative tasks, one function of communication is to coordinate actions in a complex search space that contains one or more local minima. From a PCT perspective, this implies that a perceived signal from one agent should trigger a change in a reference signal for another agent.

This hypothesis has been verified in a computer-based simulation in which two independent one-dimensional agents are obliged to communicate (that is, actively cooperate) in order to solve a two-dimensional path-finding task [19].

III. Simulation Environment

The simulation environment – implemented in the Pure Data (Pd) dataflow programming language [20], [21] – is illustrated in Fig. 1. Two 1D agents control the X and Y positions of a ‘vehicle’ in a 2D space, the task being to steer the vehicle towards a ‘target’ location. Each agent can only ‘see’ the target in its single dimension, hence cooperation *may* be required to solve the joint 2D problem. The difficulty of the task is scaled by the introduction of various forms of obstruction (as illustrated in Fig. 1), and the ‘solution time’ (ST) for each successful run was measured.

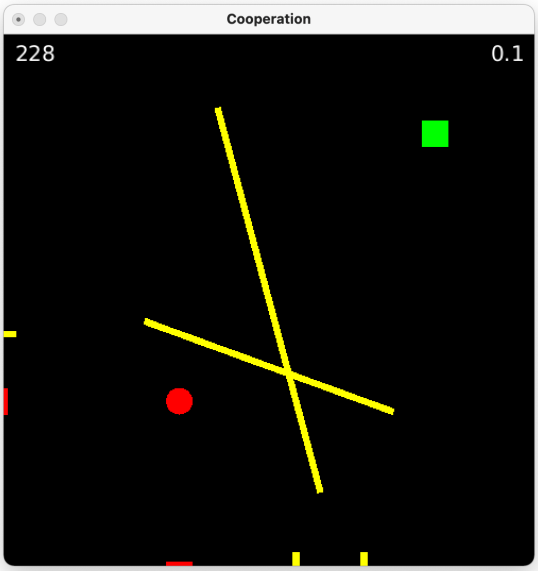


Fig. 1. Screenshot of the Pd-based simulation environment showing the target (green square), the vehicle (red circle) and two barriers (yellow lines). The X and Y axes depict the 1D projection of the vehicle, target and barriers (if visible from the agent’s perspective). The number in the top-right corner indicates the elapsed time in the current run, and the number in the top-left corner shows the number of runs completed.

There are many configurations in which each controlling X and Y agent can move the vehicle towards the target by reducing their individual ‘error’ in a monotonic fashion (that is, by gradient descent), even if there are barriers present. For example, Fig. 2 shows a configuration with three barriers but *no* local minimum. Also, some barrier configurations create situations in which it is impossible for the vehicle to reach the target at all – see Fig. 3.

However, some configurations (such as the one shown in Fig. 1) create situations which require agents to *increase* their error momentarily in order for the vehicle to eventually reach the target. For example, if one agent has reached its target (in 1D), but the other is stuck behind a barrier, then the first needs to be requested to abandon its target temporarily in an attempt to free the second agent. Hence, the presence/absence of timely communications is critical in determining whether a run is ultimately successful or not.

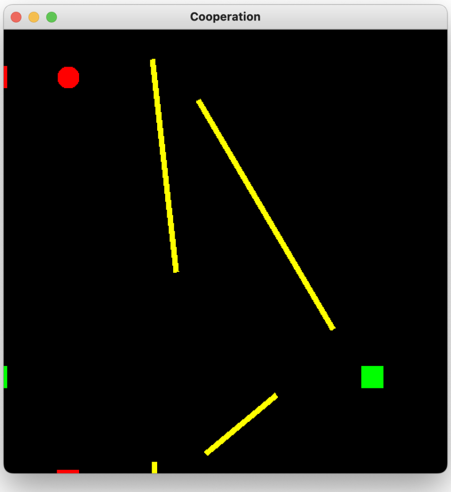


Fig. 2. Screenshot of the Pd-based simulation environment showing a configuration with three barriers but *no* local minimum. This means that the vehicle can reach the target without getting ‘stuck’.

Since not all configurations are solvable, the simulation environment was set up such that any experimental run lasting longer than 30 seconds was terminated and marked as ‘did not finish’ (DNF). In such cases, the solution time was ignored in subsequent data analysis.

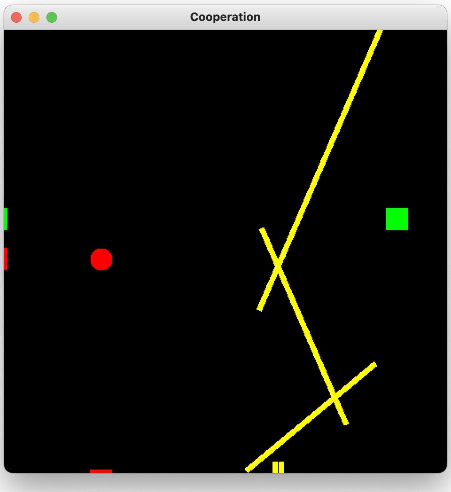


Fig. 3. Screenshot of the Pd-based simulation environment showing a configuration with three barriers in which it is impossible for the vehicle to reach the target.

Table I lists the variables instantiated in the simulations. Overall, five cooperation modes were implemented *per agent* (listed at the bottom of Table I), and different combinations were able to be specified by means of an agent-specific 4-bit binary code. This meant that there was a total of 16 possible levels of cooperation available for each agent. Two of these involved no communication at all, but distinguished between just stopping at an obstruction (i.e. no active cooperation) versus moving randomly (i.e. potential cooperation *without* communicating).

TABLE I
SIMULATION ENVIRONMENT SETUP

Agent Variables	
Action	Forward, Reverse, Stop
Status	Collided with edge Collided with barrier* ("*stuck*") Target known Target accessible* ("*access*") Arrived at target ("*arrived*")
Experiment Variables	
Environment	Initial target position [XY] Initial vehicle position [XY] Number of barriers [0–3] Location, orientation & size of barrier(s) [XYRL] Number of runs [100–1000]
Agent	Target view [true/false] Cooperation level [0000–1111][a–e] Loop gain [0.01] Back-off time [1000 msec]
Measures	Time to target (ST) [secs.] Did not finish (DNF) [#] Time spent communicating [%]

[a] [0000] if self = target known ⇒ approach target; else stop
[b] [1000] if self = target known ⇒ approach target; else move randomly
[c] [0100] if self = "*arrived*" + other = "*stuck*" ⇒ back-off
[d] [0010] if self = "*stuck*" + other = "*stuck*" ⇒ back-off
[e] [0001] if self = "*access*" + other = "*access*" ⇒ approach target

Of particular interest are each agent's 'status' parameters that were available to be communicated *for a given level of cooperation*. These are marked with a * in Table I.

The first parameter – "*stuck*" – relates to the identification of a potential local minimum. Such a condition arises when one agent has collided with a barrier and the other has arrived at the target, or when both agents have collided with barriers. Crucially, it was realised that just one agent being stuck at a barrier is not sufficient evidence for a local minimum, as the other agent may be making progress which could resolve the problem.

The second parameter – "*access*" – relates to whether the target was accessible, i.e. there was no barrier between the agent and the target. However, it is important to appreciate that such a condition does not guarantee a successful approach, as the target may subsequently become inaccessible for one agent due to the activities of the other agent.

IV. EXPERIMENTS & RESULTS

A number of experiments have been conducted, each using multiple simulation runs to investigate different configurations of obstacles and levels of cooperation. For example, Fig. 4 shows the distribution of solution times resulting from 1000 runs in an environment containing two fixed barriers (configured as shown in Fig. 1.) for four incremental levels of cooperation. As expected, enabling explicit communication between the agents had a measurable effect in speeding up solution times. However, it was also noted that the low solution times for [1000] was due to the high number of runs that did not finish (DNF). In particular, the results revealed that [1000] gave rise to 64% DNFs, whereas [1100] had 13% DNFs, [1110] had 1% DNFs, and [1111] had only 0.6% DNFs.

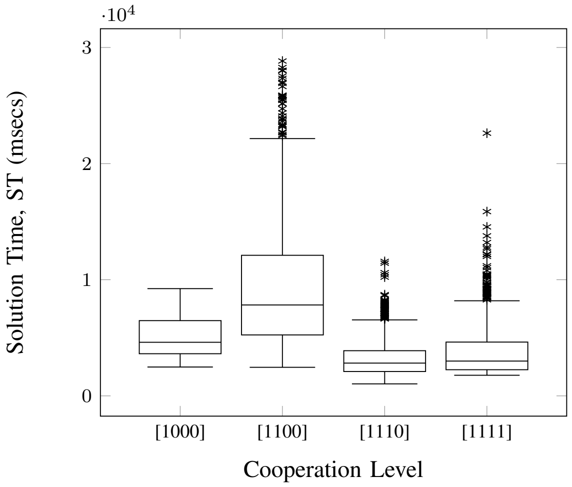


Fig. 4. Distributions of solution times for different levels of cooperation in an environment containing two fixed barriers.

In attempting to analyse the results of the more complex cooperation experiments, it became clear that an overall 'goodness measure' was needed in order to resolve the compromise between fast solution times and the numbers of runs that did not finish. This was necessary because, as seen above, a high number of DNFs tends to give rise to a low mean solution time because the runs that succeed have less challenging barrier configurations. Likewise, a low number of DNFs may be associated with higher mean solution times as a consequence of the cooperating agents taking longer to solve more challenging barrier configurations.

Hence, an appropriate 'goodness measure' was defined as:

$$GM = log(ST^{1+\frac{DNF}{nruns}}), \tag{1}$$

where GM is the goodness measure (low is good), ST is the mean solution time for a run, DNF is the number of times a run did not finish, and $nruns$ is the number of runs.

Fig. 5 shows the combined results from the solution times shown in Fig. 4 and the corresponding number of DNFs plotted using the goodness measure. This representation clearly shows that, as expected, increasing the level of cooperation between the agents leads to significant improvements in their ability to solve the designated task.

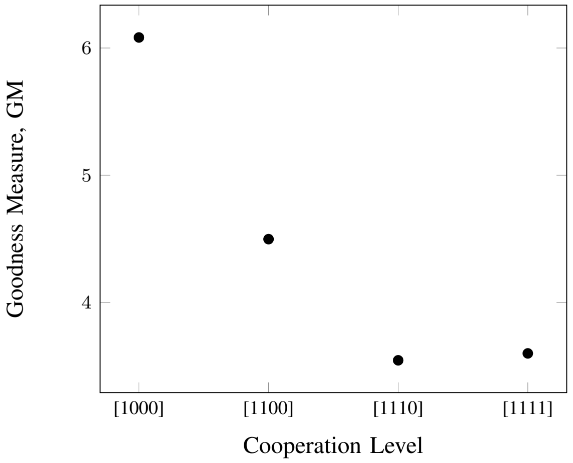


Fig. 5. Relationship between the cooperation level and the ‘goodness measure’ (low is good) in an environment containing two fixed barriers.

A. Matched Agents

As mentioned above, the simulation environment allowed the cooperation level to be set for each agent independently. However, due to the combinatorics, the majority of experiments were conducted with *matched* agents. For example, Fig. 6 shows the impact of all sixteen levels of cooperation ranked by the ‘goodness’ of the outcome for matched agents in an environment containing three randomly placed barriers with random lengths and orientations.

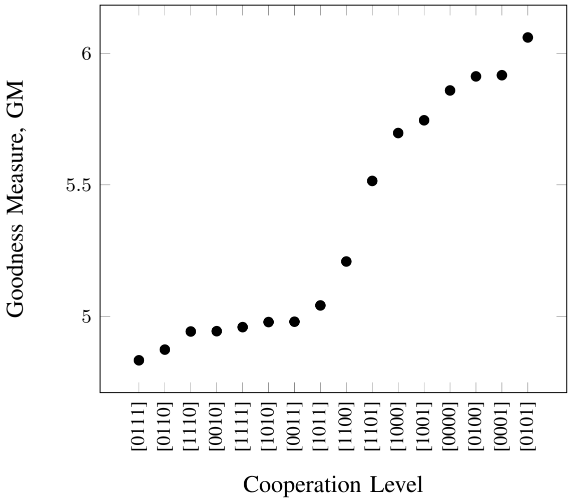


Fig. 6. Relationship between different cooperation combinations ordered by their ‘goodness’ (low is good) for matched agents in an environment containing three randomly placed barriers with random lengths and orientations.

As can be seen in Fig. 6, the relationship between different combinations of cooperation and the goodness measure reveals that the cooperation combination [0111] “*arrived*”+“*stuck*”, “*stuck*”+“*stuck*” and “*access*”+“*access*” gives rise to the best overall performance. The second-best is [0110] “*arrived*”+“*stuck*” and “*stuck*”+“*stuck*”. Next is [1110] “*random movements*”, “*arrived*”+“*stuck*”, and “*stuck*”+“*stuck*”, and fourth is [0010] “*stuck*”+“*stuck*”. The following three combinations [1010], [0011] and [1011] also have relatively high ‘goodness’, and confirm that the top eight all have [0010] ‘*stuck*”+“*stuck*” enabled, and performance drops significantly without it.

The highest number of DNFs was 854/1000 (for [0001]), the lowest was 284/1000 (for [0111]), and there were 769/1000 DNFs for no cooperation at all ([0000]). These results imply that up to 28% of barrier configurations were unsolvable and ∼23% were solvable *without* cooperation, which means that ∼49% were able to be solved *with* cooperation. The fact that [0001] resulted in a higher number of DNFs than [0000] implies that enabling the “*access*”+“*access*” strategy was actually detrimental to performance.

With regard to the proportion of time agents spent communicating, the results shown in Fig. 7 reveal that there is a clear relationship between the goodness of the cooperation combinations and the proportion of time the agents spent communicating. As noted above, this is a function of whether [0010] “*stuck*”+“*stuck*” is enabled or disabled, and it clearly reflects the frequency with which situations containing local minima arise given the three random barriers. It is also interesting to note that the highest levels of communication occurred for the top two cooperation combinations.

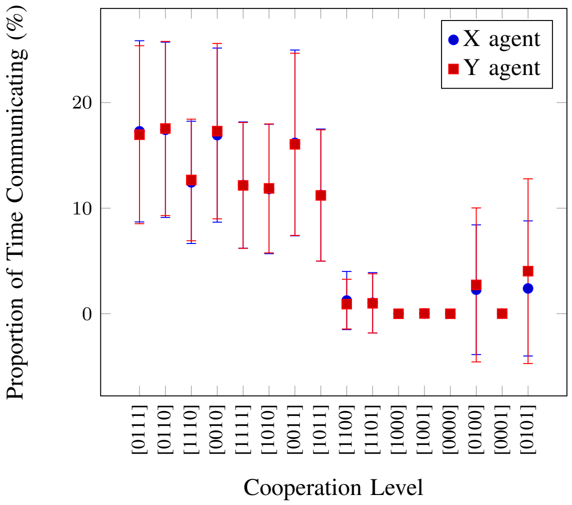


Fig. 7. Relationship between different combinations of cooperation and the % of time the X and Y agents were communicating for matched agents in an environment containing three randomly placed barriers with random lengths and orientations. The results are ordered by ‘goodness’ from left to right.

Finally, the correlation between the mean % time spent communicating and mean solution times was consistently ∼0.4 for the best eight cooperation combinations. This shows that harder barrier configurations required proportionally more inter-agent communications, as well as taking longer to solve.

B. Mismatched Agents

As an example of the consequences of allowing the cooperation level to be set independently for each agent, Fig. 8 shows results for all combinations of matched and *mismatched* agents in an environment containing three randomly placed barriers with random lengths and orientations.

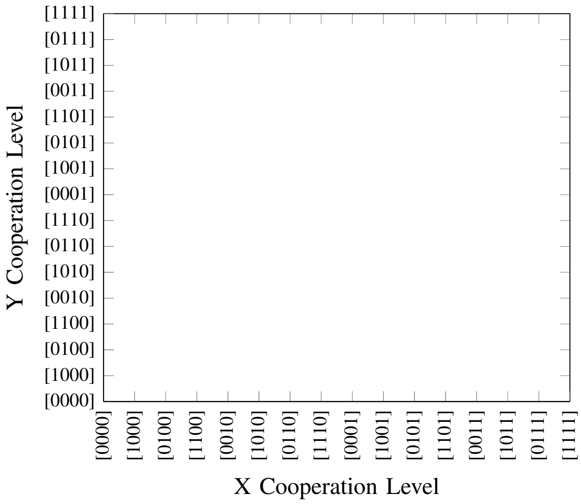


Fig. 8. Heat map of the goodness measure for all combinations of X and Y cooperation for mismatched agents in an environment containing three randomly placed barriers with random lengths and orientations. Blue corresponds to the best, and red to the worst.

The results of this experiment showed that up to 26% of barrier configurations were unsolvable and ∼30% were solvable *without* cooperation. This meant that ∼44% were able to be solved *with* cooperation.

What is particularly interesting in the results shown in Fig. 8 is that the good solutions are not confined to the diagonal, i.e. not restricted to the matched agents conditions. In fact the outcomes resulting from the best mismatched agents are comparable to those for the best matched agents. For example, the best performance over all cooperation combinations was obtained for [1111]+[0010] (i.e. where one agent had all cooperation modes enabled, and the other agent was only responding to "*stuck*"), and this result was slightly better than the best matched agents at [0110]+[0110].

Further investigation into the consequences of allowing the cooperation level to be set independently for each agent was made by comparing the performance of the best matched and mismatched combinations mentioned above (i.e. [1111]+[0010] versus [0110]+[0110]) with varying numbers of barriers. The results (shown in Fig. 9) reveal that the matched agents performed slightly better than the mismatched agents. However, as can be seen in Fig. 10, the mismatched agents communicated less than the matched agents, with the difference being proportionally larger for the more difficult barrier configurations.

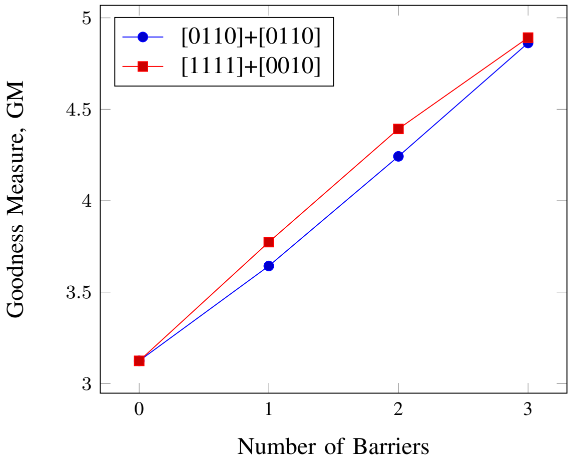


Fig. 9. Relationship between the goodness measure and the number of random barriers for matched and mismatched agents.

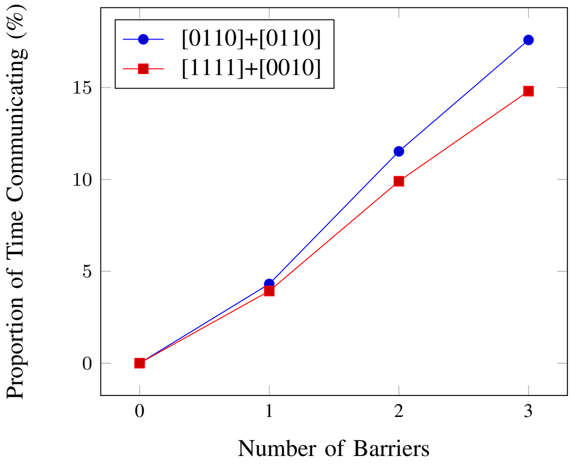


Fig. 10. Relationship between the number of random barriers and the total % of time the X and Y agents were communicating for matched and mismatched agents.

V. DISCUSSION

Clearly, the task posed here is related to finding an optimal route on a map. As such, two scenarios are possible: (i) the distance between the vehicle's current position and the target is known, but gradient descent may lead to a local minimum, or (ii) the distance between the vehicle's current position and the target is unknown due to an obstruction. The first of these may be solved by *planning* (assuming that the map is known), or by recognising that a local minimum has occurred and trying to jump out stochastically. In the second scenario, only random search is possible.

However, this map-based analysis is based on the privileged perspective of a 2D agent. In the task posed in this paper, the agents were purposefully designed to be 1D, precisely so that they did *not* have access to a 2D map. This meant that planning was not possible, and the recognition of arriving at a local minimum (or of simply not being able to see the target) required message-passing between the agents, i.e. explicit cooperation by communication.

Another insight to emerge from this work is the realisation that communication may be achieved by signalling (i.e. a 'push' from a sending agent) or by observation (i.e. a 'pull' from a receiving agent). Clearly, the latter is less efficient due to the need for continuous monitoring. Hence it can be said that, while an attention mechanism may be important, raising alerts in a timely manner are critical to success in a cooperative task.

It is also interesting to note that the overall paradigm is not specifically concerned with explicit message passing. Given that the *meanings* of the particular messages have implications for the subsequent behaviour, the scenario may also be viewed as one agent needing to appreciate the other's situation. In other words, timely communications to overcome local minima in a cooperative problem space may be viewed as instantiating a primitive 'theory-of-mind' [22].

VI. SUMMARY AND CONCLUSION

This paper has addressed the question as to what conditions the timing and structure of communication in continuous cooperative interaction. Experiments have been conducted using a PCT-based simulation of a cooperative task in which two independent one-dimensional agents are obliged to communicate in order to solve a two-dimensional path-finding problem.

Results from a number of simulation experiments have confirmed the hypothesis that appropriately timed communication between agents can overcome local minima in a joint problem space. It has also been shown that asymmetric levels of cooperative communication can be as effective as equally matched partners, and can even reduce the level of communications required to achieve the same level of performance.

Finally, although this study was aimed at *extrinsic* communication between multiple agents, it is interesting to note that the results also apply to *intrinsic* communications within a single agent.

REFERENCES

[1] G. Skantze, "Turn-taking in conversational systems and human-robot interaction: a review," *Computer Speech & Language*, vol. 67, p. 101178, 2021.

[2] H. Sacks, E. A. Schegloff, and G. Jefferson, "A simplest systematics for the organization of turn-taking for conversation," *Language*, vol. 50, no. 4, pp. 696–735, 1974.

[3] T. P. Wilson, J. M. Wiemann, and D. H. Zimmerman, "Models of turn taking in conversational interaction," *Journal of Language and Social Psychology*, vol. 3, no. 3, pp. 159–183, 1984.

[4] A. Gravano and J. Hirschberg, "Turn-taking cues in task-oriented dialogue," *Computer Speech & Language*, vol. 25, no. 3, pp. 601–634, 2011.

[5] S. C. Levinson, "Turn-taking in human communication – origins and implications for language processing," *Trends in Cognitive Sciences*, vol. 20, no. 1, pp. 6–14, 2015.

[6] E. Schegloff, "Overlapping talk and the organization of turn-taking for conversation," *Language in Society*, vol. 29, pp. 1–63, 2000.

[7] M. Heldner and J. Edlund, "Pauses, gaps and overlaps in conversations," *Journal of Phonetics*, vol. 38, no. 4, pp. 555–568, 2010.

[8] S. J. Cowley, "Of timing, turn-taking, and conversations," *Journal of Psycholinguistic Research*, vol. 27, no. 5, pp. 541–571, 1998.

[9] C. V. Goldman and S. Zilberstein, "Optimizing information exchange in cooperative multi-agent systems," in *AAMAS '03: Proceedings of the Second International Joint Conference on Autonomous Agents and Multiagent Systems*, 2003, pp. 137–144.

[10] M. C. Buehler and T. H. Weisswange, "Theory of mind based communication for human agent cooperation," in *IEEE International Conference on Human-Machine Systems (ICHMS)*, 2020.

[11] A. Das, T. Gervet, J. Romoff, D. Batra, D. Parikh, M. Rabbat, and J. Pineau, "TarMAC: Targeted Multi-Agent Communication," in *36th International Conference on Machine Learning, PMLR*, 2019, pp. 1538–1546.

[12] A. Lazaridou and M. Baroni, "Emergent multi-agent communication in the deep learning era," *arXiv*, 2020.

[13] P. Xuan, V. Lesser, and S. Zilberstein, "Communication decisions in multi-agent cooperation: model and experiments," in *AGENTS '01: Proceedings of the Fifth International Conference on Autonomous Agents*, 2001, pp. 616–623.

[14] C. Vesper, R. P. R. D. van der Wel, G. Knoblich, and N. Sebanz, "Are you ready to jump? Predictive mechanisms in interpersonal coordination," *Journal of Experimental Psychology: Human Perception and Performance*, vol. 39, no. 1, pp. 48–61, 2013.

[15] W. T. Powers, *Behavior: The Control of Perception*. NY: Aldine: Hawthorne, 1973.

[16] N. Wiener, *Cybernetics: or Control and Communication in the Animal and the Machine*, 2nd ed. Cambridge, Mass.: The MIT Press, 1965.

[17] W. Mansell, *The Interdisciplinary Handbook of Perceptual Control Theory: Living Control Systems IV*, W. Mansell, Ed. Academic Press, 2020.

[18] P. S. E. Farrell, "Perceptual control theory multiple agent interaction: a classical control theory treatment," in *The Interdisciplinary Handbook of Perceptual Control Theory, Volume II*, W. Mansell, E. de Hullu, V. Huddy, and T. Scholte, Eds. Academic Press, 2023.

[19] R. K. Moore, "Communication in cooperation: a PCT perspective," in *32nd Conference of the International Association for Perceptual Control Theory (IAPCT)*, 2022.

[20] M. Puckette, "Pure Data (Pd) open source visual programming language for multimedia," https://puredata.info.

[21] R. K. Moore, "On the use of the 'Pure Data' programming language as a real-time computational environment for exploring PCT," in *31st Conference of the International Association for Perceptual Control Theory (IAPCT)*, 2021.

[22] D. Premack and G. Woodruff, "Does the chimpanzee have a theory of mind?" *Behavioral and Brain Sciences*, vol. 1, no. 04, pp. 515–526, 1978.

Exploring the Impact of External Factors on Ride-Hailing Demand : A Predictive Modelling Approach

Anand Sriram
Department of Information Technology
National Institute of Technology Karnataka, Surathkal
Mangalore, India
anandsriram.191it105@nitk.edu.in

Ananthanarayana V. S.
Department of Information Technology
National Institute of Technology Karnataka, Surathkal
Mangalore, India
anvs@nitk.edu.in

***Abstract*—This paper presents a comprehensive study on the usage of Uber in different markets, with a focus on understanding the impact of demographic factors, public transit proximity, weather and extreme events on the demand for Uber ride-hailing services. This study involves application of Explainable AI techniques for feature selection among multiple data sources to model external factors on the Uber ride usage. Furthermore, factors such as weather and local events are used for ride usage forecasting using spatiotemporal aspects and extreme event analysis. The results of this study showed that certain factors like demography, proximity of public transit play a role in shaping the usage patterns of Uber. Also, extreme events, such as weather conditions and local events, were found to have a significant impact on the demand for Uber services. This study provides valuable insights for Uber, similar ride-hailing services and policymakers for optimal resource allocation, and lays the foundation for further research on the relationship between transportation services and various contextual factors.**

***Index Terms*—Explainable AI, Feature selection, Contextual Dependency, Extreme event analysis, Spatiotemporal analysis, Ride forecasting, Regression analysis, Ride-hailing services**

I. INTRODUCTION

The transportation industry has undergone a significant transformation in recent years, with the rise of ride-sharing platforms like Uber. The ride-hailing services has disrupted the traditional taxi industry and has become a popular mode of transportation for millions of people worldwide. The success of such ride-hailing services can be attributed to its ability to provide convenient, reliable, and cost-effective rides to customers. It is important for various ride-hailing services to understand the impact of various exogenous factors and contextual dependencies on the ride demand.

In this study, we aim to address these challenges by utilizing advanced data analytics techniques to gain insights into the factors that influence ride usage by analyzing the impact of contextual factors such as demography, existing public transit solutions, weather and public events.

We propose a novel, Explainable AI driven feature selection technique to identify the most important features that influence ride usage across multiple data sources. We then perform regression analysis to understand the impact of demographic factors and proximity to the nearest public transit hub on ride usage in different areas of a city. To identify the influence of weather and public events on ride demand, we use time series forecasting with a focus on extreme event analysis.

This study provides valuable insights into the contextual dependencies that influence ride usage, and can inform the decision-making process for ride-hailing services for optimal resource allocation and improved profitability. The results of this study can also be used by other ride-sharing platforms, transportation companies, and policymakers to gain a better understanding of the various exogenous factors that influence ride usage.

II. LITERATURE REVIEW

A. Background and Related Works

The analysis of the impact by various external factors on the trip fare of ride sharing services have attracted the attention of many researchers pertaining to different case studies on the same area. S. Guo et al. 2018 [1] performs dynamic price prediction of ride-on-demand (RoD) services based on multi-source urban data. The price prediction methodology is based on features extracted from multiple sources such as the RoD service, taxi service, public transportation, weather, the map of a city, etc. A neural network is trained using different combinations of feature sets to perform the required prediction and are appropriately evaluated. The first drawback of this methodology is the absence of a well-defined feature selection strategy. The paper makes use of different combinations of multiple feature sets to illustrate the improvement in accuracy and related metrics. However, this method is not scalable with addition of more external factors. The second drawback is the absence of a feature selection strategy within a given feature set. As a result, the neural network would have likely been trained with low impact features that could have affected the performance of the model.

Chao Chen et al. 2018 [2] makes use of different data sources pertaining to urban data such as weather, traffic, duration, etc. The solution involves incorporating these data sources to train a simple linear regression model with high-dimensional composite features to perform the prediction. The composite features are obtained by combining simple basic features based on certain properties. By using linear regression for prediction, the model fails to take into account the possible non-linear relationships between the input parameters and target variables. Furthermore, the basic features are combined into composite features by intuition and brute force, without any definite algorithm for the same.

Le Chen et al. 2015 [3] tries to evaluate Uber's dynamic pricing mechanism by treating it as a black-box, and predicts future prices based on a predicted relationship between the historical trip fare and the supply and demand. The prediction is inaccurate, due to the lack of real service data and the absence of including additional external factors for predicting the relationship.

M. Battifaranoa et al. 2019 [4] proposes a general data-driven framework is developed for predicting short-term surge multipliers. The solution utilises multiple data sources from traffic, built environment and weather. A log-linear model is trained using features on surges, patterns, events and traffic speeds.

Y. Liu et al. 2017 [5] presents a preliminary study on predicting price multipliers. The solution makes use of only the RoD service data and weather data to predict the hourly average price multiplier in specific areas. Different predictors such as Markov-chain and neural network predictors are used to predict the surge prices at different localities within the city. The limited factors used do not present accurate results.

J. Hall et al. 2015 [6] examines a case study of Uber prices surging during well-attended events. Due to the effect, supply rises to meet demand and completion rate remains high. In comparison, when the surge pricing system experienced an outage on New Year's Eve, completion rate fell dramatically. The solution demonstrates that events impact the surge price and the corresponding impact on the supply and demand aspect.

B. *Outcome of Literature Review*

It has been observed that there is limited literature available [7][8][9][10] that demonstrates the use of multiple contextual factors like weather and traffic to predict the trip fare of ride sharing companies. The papers that explore the impact of multi-source urban data on the trip fare do not make use of any specially designed feature selection strategy for their respective prediction models. A good feature selection technique is even more necessary in such cases due to the fact that incorporating features from different data sources would bring about a considerable amount of noise in the dataset.

After identifying the gaps in the existing literature, this paper aims to address each of the drawbacks by (i) incorporating more features from the external data sources, (ii) designing a novel feature selection strategy, considering the ride sharing scenario as the relevant case study, and (iii) presenting the combined results in order to analyse the impact of contextual features driving the model's predictions.

III. METHODOLOGY

The methodology of this study follows the typical workflow of a data science solution. Fig. 1 presents the complete architecture diagram, along with interaction among the different components. More description pertaining to each component is presented in the further subsections. The primary goal of this paper is to apply various analytical techniques and machine learning models, which would aid in deducing the impact of various exogenous factors on the usage of ride-hailing services. The city of New York is considered for the case study in this paper.

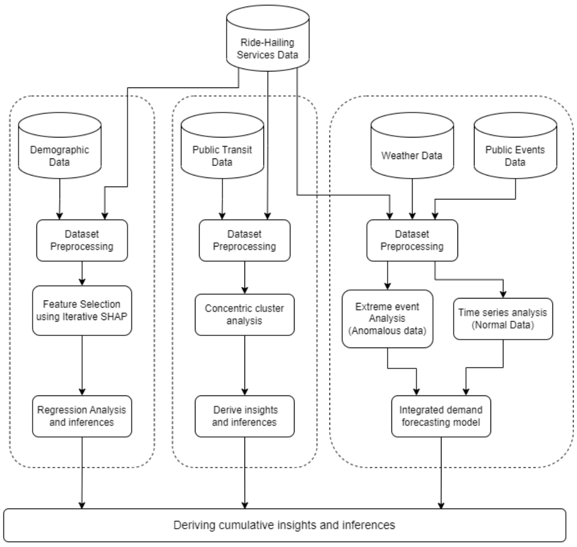

Ride-Hailing Services Data
Demographic Data
Public Transit Data
Weather Data
Public Events Data
Dataset Preprocessing
Dataset Preprocessing
Dataset Preprocessing
Feature Selection using Iterative SHAP
Concentric cluster analysis
Extreme event Analysis (Anomalous data)
Time series analysis (Normal Data)
Regression Analysis and inferences
Derive insights and inferences
Integrated demand forecasting model
Deriving cumulative insights and inferences

Fig. 1. Architecture Diagram

A. *Dataset Description*

There are a good number of contextual dependencies to be considered to understand their impact on the usage of ride sharing services. Data sources such as weather data, proximity to public transportation hubs, public events and demographic data are some feature sets to be considered. Due to the unique nature of each feature set, they are to be sourced from different data sources and collated together in a standard form to ensure seamless processing. To represent the data for ride-hailing services, Uber data is sourced from a popular Kaggle dataset. Weather data is sourced from weather.com. Public events data is source from Timeout. Public transportation data are sourced from NYC Open Data. Demographic data is sourced from NYC Census data. The Uber rides dataset consists of over 4.5 million data points spanning across the months of April to September 2014, with 4 attributes denoting the timestamp, latitude and longitude of the ride pickup, and the spatial coordinates of the same. The weather dataset consists of hourly weather data for the same timescale, with 14 attributes denoting the spatiotemporal features along with different weather conditions such as temperature, precipitation, humidity, etc. the public events dataset consists of 38 most popular events in New York that occurred during the same 6-month timescale of 2014. The attributes pertain to the event name, location coordinates and the date-time field. The public transit data consists of more than 2200 bus stops with several features, the key ones being the location coordinates and details of each bus stop. The demographics census data consists of 30 socio-economic features across five broad categories : Economic status, Employment status, Employment sector, Mode of Commute and Ethnicity. There are over 2000 data points, providing census information for each census tract of New York. The common attribute among all these datasets are timestamps. Based on timestamps and datetime-related attributes, the datasets are merged and grouped as required by each of the components.

B. Dataset Preprocessing

The collated data from the previous step are appropriately transformed to ensure consistency for further processing of the data. This data is cleaned using two techniques, normalization and outlier removal.

Data normalization is a crucial step in the preprocessing of data that involves transforming data into a common scale to prevent any disproportionate feature values from influencing the results of the analysis. The Min-Max scaling is used to normalize in this regard. In this technique, the data is scaled to a fixed range of between 0 and 1. This is done by subtracting the minimum value in the feature column from each value and then dividing the result by the difference between the maximum and minimum values in the column. Min-Max scaling preserves the original distribution of the data, but scales the data to a common range.

Outliers are data points that lie far from the majority of the data and may have a disproportionate impact on the analysis. Outliers can be detected and removed using the Z-score method. In this technique, the Z-score of each data point is calculated by subtracting the mean of the feature from the data point and dividing the result by the standard deviation of the feature. Data points that have a Z-score greater than a certain threshold are considered outliers and can be removed from the dataset.

C. Analyzing The Impact of Exogenous Factors

1) Novel Feature Selection - Demographic Variables: Feature selection is a very important aspect of any data science workflow. A good feature selection strategy has the potential to boost the model performance manifold, while a bad feature selection approach can hamper the performance of an otherwise fine-tuned model. To address this challenge, a novel feature selection technique based on Explainable AI is developed to rank the importance of each feature.

The novel feature selection technique is termed as “Iterative SHAP”, which makes use of SHAP analysis for feature selection. SHAP analysis is a popular technique used in the Explainable AI domain, where it is used to measure the degree of impact by each feature on the prediction of a particular ML model. In this scenario, the SHAP analysis would be used for feature selection as a wrapper-based technique wherein the best subset of features are iterative selected through SHAP scores for the maximum model performance. The workflow of this technique is depicted in Fig. 2.

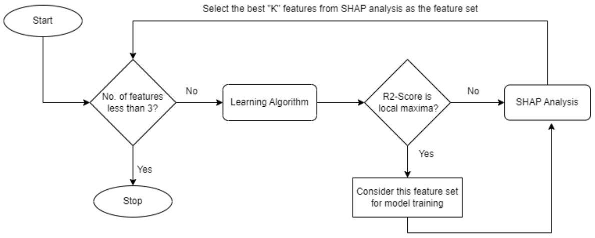

Start
Select the best "K" features from SHAP analysis as the feature set
No. of features less than 3?
No
Learning Algorithm
R2-Score is local maxima?
No
SHAP Analysis
Yes
Stop
Yes
Consider this feature set for model training

Fig. 2. Iterative SHAP Analysis For Feature Selection

The novel “Iterative SHAP” feature selection technique is used to determine the relationship between ride usage and a multitude of different demographic variables. These demographic factors are meaningfully categorized into different feature buckets such as economic status, ethnicity and employment sector to name a few. Within each feature bucket, there are a number of attributes pertaining to the respective category. Considering the subject-oriented nature of the attributes and the feature buckets, this novel feature selection technique is used to identify the most appropriate features among multiple attribute buckets of the demographic variables. The integration of the best features from the different buckets would be used for further analysis. The results of this analysis provide insights into the demographic characteristics influencing ride usage.

The resultant R2-scores from this technique are compared with the standard baseline feature selection techniques such as Recursive Feature Elimination (RFE), correlation analysis, etc. To maintain consistency in the experiment and avoid skewness of results, the selected features from each technique would be used to train multiple ML models for regression analysis ranging from Linear Regression to XGBoost model.

2) Analysis of Public Transit Proximity: The study also aimed to determine the impact of public transit proximity on the ride usage. It is a known fact that ride-hailing services and public transportation systems are contrasting modes of transit. As a result, one can infer that areas closer to public transit stations would have lesser density of ride pickups as compared to areas farther away from the same stations.

To infer this empirically, the spatial distribution of the ride pickups was analyzed in proximity to the respective nearest bus stations in New York City. The spatial analysis involves visualizing the distribution of the rides using concentric clusters centered around public transit stations. The radii of each of these concentric clusters are spaced out in intervals of the Fibonacci sequence. The rationale for specifically opting for the Fibonacci sequence as the radii intervals is due to the fact that the Golden Ratio and the Fibonacci sequence appear in different areas, forms and patterns in the world. Further analysis involves analyzing the distribution of the rides within each of these concentric clusters in absolute magnitude and normalized quantity to empirically deduce how the ride density varies as the distance from the nearest public transit station increases.

The results of this analysis show the relationship between usage of ride-hailing services and the proximity to the nearest public transit mode, which would help ride-hailing services to optimize its resource allocation and operations in different areas of a city.

3) Extreme Event Analysis : Impact By Weather And Public Events: The study aims to identify the influence of exogenous factors such as weather and public events on usage of ride-hailing services. Intuitively, we would assume that the time series distribution of the rides is a periodic distribution on a weekly basis. The rationale behind this intuition is the fact that ride share during weekdays are dominated by the working population and during weekends, the ride share is dominated by leisure seekers. This implies that the general ride pattern is expected to be periodic on a weekly basis.

However on certain occasions like public events, there would be a sudden spike in the ride demand. While on other instances such as extreme weather conditions would not make it conducive for the people to venture outdoors, which may lead to a dip in the number of rides.

In order to address the challenge presented by such extreme events on the time series data, this study proposes a hybrid forecasting solution involving rare event analysis and a two-step time series analysis. The rare event analysis would be used to identify the days when the number of rides vary significantly from the median value. Experimentally, this value has been found to be a standard deviation of 0.5 from the median number of rides. Using this analysis, the time series dataset of the rides would be split into normal days and rare event days for further experimentation. Based on the split of the rides time series dataset, the time series datasets of weather and public events would also be split correspondingly to maintain consistency of input parameters.

After the precomputation based on extreme event analysis, the two-step time series analysis is implemented to formulate an integrated forecasting model that accounts for the normal days and the anomalous days. The first step is used to model the time series distribution for weather and public events. The second step involves modelling the rides time series distribution for the anomalous days. The time series modeling involves combining the integrated time series model of exogenous factors (from the first step) with the standalone time series model of the rides. This solution is primarily used to model the time series distribution of rides for the anomalous days. The integrated time series model obtained in the second step is then merged with the standalone time series model of rides for the normal days. The resultant time series model is used as a singular solution to forecast the ride demand of ride-hailing services, irrespective of the days being normal or anomalous.

D. Deriving Cumulative Insights And Inferences

The complete model architecture consists of three standalone components that analyze the impact of various contextual dependencies on the usage of ride-hailing services. Hence, it is vital to tie together the various models and derive cumulative insights and inferences from them. These insights and inferences involves churning out the hidden patterns from the results of the model analyses, which in turn would provide a human-understandable justification for why the model behaves in a certain way for any given set of input parameters. In models with lesser number of features, it is possible to come up such explanations based on simple statistics. However this model would be trained on several, unrelated feature sets. Explaining this model based on simple statistics would be lossy and inefficient. Thus, it becomes essential to present valid insights from the results that would otherwise go unnoticed by statistical models.

After assessing the empirical results of the model performance, it would be logical to make suitable inferences pertaining to the impact of different contextual feature sets on the ride demand. Comparison of results based on different feature selection strategy would also help in assessing the best feature selection strategy to be used in such scenarios. Furthermore, a cumulative analysis of the impact by various exogenous factors on the ride usage would aid in decision making for the stakeholders, thereby improving profitability for the ride-hailing services.

IV. RESULTS AND ANALYSIS

A. Feature Selection using Iterative SHAP

Feature selection involves the selection of the best subset of features among all attributes in the dataset for the purpose of training the model. This is one of the most vital operations in the machine learning workflow and has a major impact on the model performance.

The Iterative SHAP feature selection is a wrapper-based technique wherein the best subset of features are iteratively selected through SHAP analysis for the maximum model accuracy score. In this experiment, the multiple regression models are used to compute the R2-Score which would be used for relative comparison of the results.

The dataset consists of 30 features spanning across 5 different demographic categories namely, Ethnicity, Employment Status, Economic Status, Mode of Commute and Employment Sector.

TABLE I
R2-SCORE VS. NUMBER OF FEATURES (MODEL : LINEAR REGRESSION)

Number of Features	R2-Score
30	0.548
21	0.628
14	0.802
12	**0.826**
8	0.761
3	0.485

TABLE II
Comparison of R2-Score by Different Feature Selection Techniques

Technique	Type	Lin. Reg.	R. Forest	G. Boost	XGBoost
Iterative SHAP	**Wrapper-Based**	**0.826**	**0.894**	**0.876**	**0.908**
FFS	Wrapper-Based	0.795	0.871	0.798	0.873
Correlation	Filter-Based	0.732	0.634	0.593	0.656
PCA	Filter-Based	0.738	0.729	0.726	0.722
RFE	Wrapper-Based	0.764	0.762	0.703	0.744
Chi-Square	Filter-Based	0.474	0.489	0.374	0.538
Lasso	Wrapper-Based	0.646	0.689	0.677	0.775
Info. Gain	Filter-Based	0.587	0.543	0.692	0.667
BFS	Wrapper-Based	0.751	0.774	0.726	0.728

Initially, the SHAP values are computed for the dataset with all 30 features using the linear regression model. The corresponding SHAP summary plot is generated and based on the distribution of data points with respect to each feature, we determine the impact of all the features towards driving the model's prediction. Upon a manual selection of a suitable threshold, the number of features are brought down to 21 of the most impactful features until this point.

This process is iteratively continued until the attainment of a predefined terminating condition. In this case, the terminating condition is less than 3 features for further computation. An alternative possible terminating condition is the attainment of a point of inflection which would imply a local or global maximum of the R2-score.

Table 1 tabulates the variation of R2-Score of the regression model with respect to the number of features used in the model. We observe that R2-Score attains a global maximum when the feature count is 12. Hence, we can conclude that these 12 features are the most impactful amongst all features and would lead to better model performance.

Table 2 tabulates the R2-Scores obtained from the selected regression models, based on multiple feature selection techniques including the proposed Iterative SHAP technique. The result obtained by the Iterative SHAP feature selection method is compared with the currently popular feature selection techniques. In comparison to these two standard existing techniques, the proposed Iterative SHAP technique has the highest R2-Score across all ML models. The reason for the better performance is due to the inclusion of Explainable AI in feature selection, which clearly defines the impact of individual features on the model. Other wrapper-based methods tend to use heuristic measures which do not attain a global maxima. Filter-based method tend to no factor in the inter-dependence of features and hence, they provide inferior results.

B. Proximity to Public Transit Hubs

The goal of this analysis is to understand how the distribution of the ride pickups varies with respect to the distance from a public transit hub (bus stop).

Fig. 3 presents a line distribution of how the number of ride pickups varies by the distance from all the bus stops in absolute quantity. The individual line distributions of all bus stops are plotted with a translucent opacity. To summarize the plot, the mean and median of the distribution are also plotted.

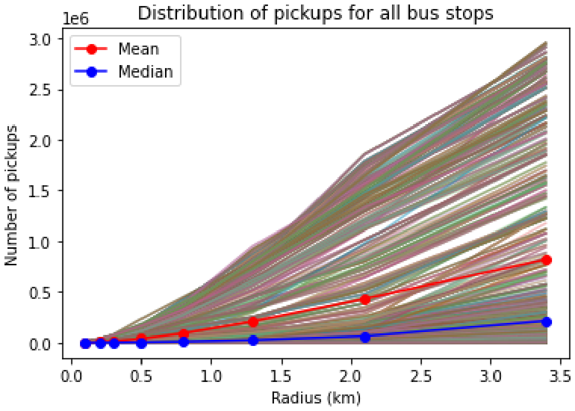


Fig. 3. Distribution of ride pickups from bus stops (Absolute)

In Fig. 3, we observe that as the distance from the bus stop increases, the number of rides tends to increase exponentially. Furthermore, we observe the mean is much higher than the median, signifying a skew towards bus stops with higher density of pickups. From this plot, we can infer that people do not use ride-hailing services closer to the public transit hubs, probably because of the convenience offered by the close proximity to public transportation. However, as the distance from the bus stop increases, the people are likely to use ride-hailing services for commuting due to the other option being farther away.

Fig. 4 presents a line distribution of how the number of ride pickups varies by the distance from all the bus stops, in normalized quantity with respect to the distance. The individual line distributions of all bus stops are plotted with a translucent opacity. To summarize the plot, the mean and median of the distribution are also plotted. In Fig. 4, we observe the distribution of ride pickups with respect to the proximity from the bus stop that has been normalized by the distance from the said transit hub. In this plot, we observe that after the 2 km mark, the normalized number of pickups tends to reach a saturation point. The saturation value implies that the distribution of the ride pickups is going to be similar for

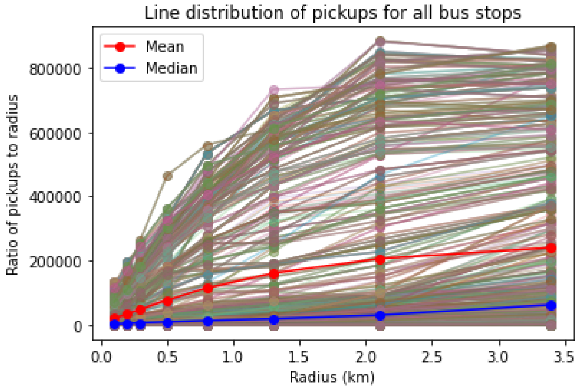


Fig. 4. Distribution of ride pickups from bus stops (Normalized)

even farther distances. From this plot, we infer that the public transit hub has an influence on the user's commute choice within a radius of 2 km from the respective bus stop. Beyond this region of influence, the ride choice of the people would incline more towards ride-hailing services rather than public transit options.

C. *Time Series Forecasting with Extreme Event Analysis*

The goal of this component is to forecast ride demand for ride-hailing services, driven by external factors such as weather conditions and public events. It had been discussed that on certain occasions like public events, there would be a sudden spike in the ride demand, and on other instances such as extreme weather conditions would not make it conducive for the people to venture outdoors, which may lead to a dip in the number of rides. Fig. 5 presents a time series plot of the number of rides on each Saturday for a total 26 weeks. The mean and median of the distribution are also plotted.

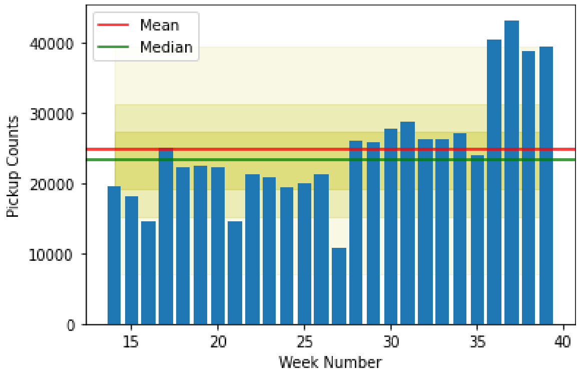


Fig. 5. Weekly Distribution of rides for Saturday

The shaded regions in the figure represent the values within standard deviations of 0.5, 1 and 2 from the median. The darkest region represents 0.5 times the standard deviation and the lighter regions represent 1 and 2, respectively. In Fig. 5, we observe a predominantly constant distribution of rides for all weeks. This is consistent with the assumption that the number of rides are periodic on a weekly basis. However, on certain weeks there is a surge from the usual pattern (weeks 36-39). And on certain weeks, there is a dip in the ride demand (weeks 17, 21, 27). It can be safely assumed that the reason for this surge and dip would be certain extreme events, like weather and public events.

In order to identify these "anomalous" days, we consider the data points that lie beyond 0.5 times the standard deviation from the median as anomaly. And the points that lie within the zone as normal days, that adhere to the periodic pattern. The data points are then split between normal and anomalous days, for separate processing.

Based on the weather-related data and public events data, the time series forecasting model is generated which is driven by these extreme events. These time series forecasting models are implemented using standard ML algorithms, through a 70:30 train-test split of the data points. In order to perform an appropriate comparison, the ML algorithms are used to generate the forecasting on all the rides data using lag features consisting of the past 3 weeks, without any input of the exogenous variables. This comparison would be useful to understand the importance of extreme event analysis in the case of the ride demand forecasting. Table 3 tabulates the results in terms of standard error values generated by the four ML algorithms for the standard approach and our proposed approach. The standard error conveys the fact that by how much the generated forecasts deviate from the ground truth.

From the values tabulated in Table 3, we observe that the proposed approach involving extreme event analysis is better than the naive baseline approach for all the standard ML algorithms used. This is due to the fact that the standard approach does not consider the influence of extreme events on the time series data. As a result, our proposed approach performs better than the standard approach.

TABLE III
COMPARISON OF STD. ERROR

ML Algorithm	**Our Approach**	Standard Approach
Linear Regression	**0.514**	0.742
Random Forest	**0.814**	0.852
Gradient Boost	**0.385**	0.637
XGBoost	**0.458**	0.656

D. *Cumulative Insights And Inferences*

The first component examines how demographic factors impact Uber ride usage. Economic status and employment sector were found to have a greater influence on ride patterns, while ethnicity had little impact. Riders' transportation choices varied based on economic status, with some opting for public transit, ride-hailing services, or personal vehicles. Professional sector also played a role, with regular workers preferring ride-hailing services in order to beat the public transit rush during peak hours. From this perspective, ride-hailing services can use demographic variables to optimally allocate vehicles to certain neighborhoods at specific times.

The second component analyzes how proximity to public transit stations influences users' commute choices. From Fig. 3 and Fig. 4, we observe that as distance from the station increases, the number of ride pickups also increases, suggesting that the presence of the public transit station influences the user's choice for their mode of commute within a 2 km radius. The reason is due to the fact that public transit is considered more economical than ride-hailing services, and users tend to make a tradeoff between time and money. If a public transit mode is in close proximity to the users, they are more likely to prefer the public transit mode given that it is a more economical transportation mode than any ride-hailing service. From this perspective, ride-hailing services can allocate more vehicles to areas farther from public transit hubs to improve resource allocation and profitability.

The third component is to factor in the impact of extreme events such as weather and public events on the ride usage. The initial assumption was that the time series distribution of Uber rides is a periodic distribution on a weekly basis, due to repeated weekday usage by the working population and weekend usage by leisure seekers. Also, the sudden spike and dip in the ride demand was speculated to be due to extreme weather conditions and occurrences of public events. Based on the results tabulated in Table 3, we can conclude that the aforementioned extreme events influenced the ride demand as assumed. From this perspective, the ride-hailing services can anticipate ride demand, irrespective of the days being normal or with extreme events. Without this analysis, the ride-hailing services can potentially over-allocate on days when a dip in demand is certain, leading to increased operational costs. And on days when a surge is likely, ride-hailing services may fail to allocate additional fleet, thus missing out on the increased revenue streams. Thus, the extreme event analysis aids in reducing costs, improving revenue and thereby, maximizing the profitability for ride-hailing services.

V. CONCLUSION AND FUTURE WORK

The study aimed to understand the contextual dependencies affecting usage of ride-hailing services, which will aid in forecasting ride demand. The results of the study showed that demographic factors such as economic status and employment sector play a crucial role in determining usage patterns of ride-hailing services. The proximity of public transit was also found to have a significant influence on the rider's choice, between ride-hailing services and public transit modes. Extreme events such as weather conditions and events also had a noticeable impact on ride demand, which boosted the performance of the ride demand forecasting model.

Future work could include incorporating more data sources, such as traffic patterns and road conditions, to improve the accuracy of the demand forecast. Another area of future work could be to study the impact of government policies and regulations on Uber usage. This would provide valuable insights into how government policies could be used to promote sustainable transportation solutions and reduce the negative impact of ride-sharing services on the environment.

Overall, this study highlights the importance of considering multitude of contextual dependencies and exogenous variables when analyzing the demand for ride-hailing services. The findings of this study can be used to inform future policies and strategies aimed at promoting sustainable transportation solutions.

REFERENCES

[1] Suiming Guo, Chao Chen, Jingyuan Wang, Yaxiao Liu, Ke Xu, and Dah Ming Chiu. 2018. Dynamic price prediction in ride-on-demand service with multi-source urban data. In *Proceedings of EAI International Conference on Mobile and Ubiquitous Systems: Computing, Networking and Services (MobiQuitous '18)* New York, USA: ACM

[2] Chao Chen, Suiming Guo, Jingyuan Wang, Yaxiao Liu, Ke Xu, and Dah Ming Chiu. 2018. A simple but quantifiable approach to dynamic price prediction in ride-on-demand services leveraging multi-source urban data. In *Proceedings of ACM Interact. Mob. Wearable Ubiquitous Technol* ACM

[3] Le Chen, Alan Mislove, and Christo Wilson. 2015. Peeking beneath the hood of uber. In *Proceedings of ACM Conference on Internet Measurement Conference (IMC '15)* New York, USA : ACM

[4] Matthew Battifaranoa and Zhen (Sean) Qian. 2019. Predicting real-time surge pricing of ride-sourcing companies. In *Transportation Research Part C: Emerging Technologies Volume 107*

[5] Yaxiao Liu, Ke Xu, Suiming Guo, Chao Chen, and Dah Ming Chiu. 2017. It can be cheaper: Using price prediction to obtain better prices from dynamic pricing in ride-on-demand service. In *Proceedings of the 14th EAI International Conference on Mobile and Ubiquitous Systems: Computing, Networking and Services (MobiQuitous '17)* New York, USA : ACM

[6] J. Hall, C. Kendrick, and C Nosko. 2015. The effects of uber's surge pricing: A case study. In *Uber Under The Hood*

[7] Elizabeth Rani G, Sakthimohan. M, Revanth Raj. R, Sri Ganesh. M, and Shyam Sunder. Rand Karthigadevi. 2022. An automated cost prediction in uber/call taxi using machine learning algorithm. In *Proceedings of the 2nd International Conference on Advance Computing and Innovative Technologies in Engineering (ICACITE)* : IEEE

[8] Y.W. Bai and E. W. Wang. 2012. Design of taxi routing and fare estimation program with reprediction methods for a smart phone. In *Proceedings of International Instrumentation and Measurement Technology Conference* :IEE.

[9] T. Prem Jacob, A. Pravin, K. Mohana Prasad, G. T. Judgi, and R. Rajakumar. 2022. Real time prediction of cab fare using machine learning. In *Proceedings of International Conference on Electronics and Renewable Systems (ICEARS)* : IEEE.

[10] K. Zhao, D. Khryashchev, and H. Vo. 2021. Predicting taxi and Uber demand in cities: Approaching the limit of predictability. In *Proceedings of EEE Transactions on Knowledge and Data Engineering* vol. 33, no. 6. : IEEE.

Evolving Time-Dependent Cognitive Models

Peter Lane[1*], Noman Javed[2], Angelo Pirrone[2], Laura Bartlett[2] and Fernand Gobet[2]

[1]School of Physics, Engineering and Computer Science, University of Hertfordshire, Hatfield, AL10 9AB UK.
[2]Centre for Philosophy of Natural and Social Science, London School of Economics, Houghton Street, London, WC2A 2AE UK.

*Corresponding author(s). E-mail(s): p.c.lane@herts.ac.uk;
Contributing authors: n.javed3@lse.ac.uk; a.pirrone@lse.ac.uk; l.bartlett@lse.ac.uk; f.gobet@lse.ac.uk;

Developing, understanding and verifying the behaviour of cognitive models is a non-trivial task. A good cognitive model will provide some explanation of how a human performs in a particular experimental setting, and even provide predictions for new settings. In many cases, cognitive models are based around computer programs which need to be designed and written for the given experiment and behaviour: ideally, models will be developed to cover a wide range of possible explanations, but often time constraints or natural bias (oversights) lead to models written by human programmers being constrained to particular groups of solutions. This paper will describe, discuss and illustrate a proposed system for helping to automate the construction of programs to form high quality cognitive models.

Our proposed system is based on Genetic Programming [4], a technique which searches a large space of programs for candidate solutions to a given fitness criterion. We have applied our methodology to various tasks, including variants of the Delayed Match To Sample task (DMTS) and a Decision Making task – in this paper, we discuss only the DMTS task. Unique aspects of our system include a *phased-evolution system*, which aids in finding models with both behaviour and time fitness requirements, and *extensive post-processing steps*, which reduce the large number of models output by the system to a smaller, more understandable subset, with graphical and text representations. Our approach [2, 5, 6] using GP appears unique in developing cognitive models which focus on symbolic, information-processing [8] explanations of human cognition. This contrasts with many current approaches in artificial intelligence which rely on connectionist (statistical) explanations based on large datasets: a recent study in this area is that of [7].

As an example of program synthesis, our system can be conveniently divided into three parts [3]: the task definition (*user intent*), to express what makes a good program; a *search space* of candidate programs; and a *search technique*, to explore the given search space for good programs.

The task studied in this paper is the DMTS task [1], a typical neuroscientific experiment, popular for studies of short-term memory, which tests the accuracy and reaction time for subjects to recognise images. In this experiment a picture is presented for 1 second in the center of the screen. Then, after a delay of 0.5 seconds, two pictures are presented for 2 seconds, one on the left and the other on the right of the screen. The participant has to select which of those two pictures is the same as the first picture.

Although this task is an example of "programming-by-example", where the model must reproduce the example input-output behaviour, the overall quality of the model is not judged on the number of correct input-output pairs. As reported in [1], across the complete set of presentations, human subjects only score 95.7% accuracy, with an average response time of 767ms: the model's accuracy and simulated response times are judged against these values. Such time-dependent models require special attention during the evolution process and, in particular, we have created a novel phased-evolution system which gradually introduces elements of the fitness function over time.

Each individual model is defined by a control program to be interpreted within a simple cognitive architecture: the space of possible control programs is the search space for our system. This architecture has some task-specific input/output components: a set of *inputs* and a *response*. It also has some task-independent components: a fixed-size short-term memory (STM), and a working memory. Finally, each model has a *clock*, to record its current in-task time. The model's control program is composed from a set of operators defining a simple imperative programming language. The model's current working value, STM and clock values can all be manipulated, inputs read and a response prepared: the current response is "made" when the program ends.

The GP search technique is used to generate multiple candidate models: each run generates many 'good' models (based on a fitness value threshold). A unique aspect of our approach is the amount of post-processing performed on the candidate models, which are otherwise too numerous to analyse and understand. For example (see [6] for details), a typical output of six runs of the GP system produced 1164 distinct models with a good fitness value. By removing bloat, these were reduced to 248 distinct models. We then rewrote semantically equivalent programs to further reduce the number to 11 distinct models. Application of a clustering algorithm divides this space of models into two basic groups – studying individual members of those groups enables a scientist to develop an explanation of behaviour in this domain.

References

[1] Chao L, Haxby J, Martin A (1999) Attribute-based neural substrates in temporal cortex for perceiving and knowing about objects. Nature Neuroscience 2:913–20

[2] Frias-Martinez E, Gobet F (2007) Automatic generation of cognitive theories using genetic programming. Minds and Machines 17:287–309

[3] Gulwani S (2010) Dimensions in program synthesis. In: Proceedings of the 12th International ACM SIGPLAN Symposium on Principles and Practice of Declarative Programming, pp 13–24

[4] Koza JR (1992) Genetic Programming: On the programming of computers by means of natural selection. MIT Press, Cambridge, MA

[5] Lane PCR, Sozou PD, Gobet F, et al (2016) Analysing psychological data by evolving computational models. In: Wilhelm A, Kestler H (eds) Analysis of Large and Complex Data. Studies in Classification, Data Analysis, and Knowledge Organization. Springer, Cham, pp 587–97, https://doi.org/10.1007/978-3-319-25226-1_50

[6] Lane PCR, Bartlett L, Javed N, et al (2022) Evolving understandable cognitive models. In: Proceedings of the 20th International Conference on Cognitive Modeling

[7] Peterson JC, Bourgin DD, Agrawal M, et al (2021) Using large-scale experiments and machine learning to discover theories of human decision-making. Science 372:1209 – 1214

[8] Simon HA (1981) Information-processing models of cognition. Annual Review of Psychology 30:363–96

Part III.

AI & Games

Switchable Lightweight Anti-Symmetric Processing (SLAP) with CNN Outspeeds Data Augmentation by Smaller Sample – Application in Gomoku Reinforcement Learning

Chi-Hang Suen
City, University of London
London, UK
chi.suen@city.ac.uk

Eduardo Alonso
City, University of London
London, UK
e.alonso@city.ac.uk

***Abstract*—To replace data augmentation, this paper proposed a method called SLAP to intensify experience to speed up machine learning and reduce the sample size. SLAP is a model-independent protocol/function to produce the same output given different transformation variants. SLAP improved the convergence speed of convolutional neural network learning by 83% in the experiments with Gomoku game states, with only one eighth of the sample size compared with data augmentation. In reinforcement learning for Gomoku, using AlphaGo Zero/AlphaZero algorithm with data augmentation as baseline, SLAP reduced the number of training samples by a factor of 8 and achieved similar winning rate against the same evaluator, but it was not yet evident that it could speed up reinforcement learning. The benefits should at least apply to domains that are invariant to symmetry or certain transformations. As future work, SLAP may aid more explainable learning and transfer learning for domains that are not invariant to symmetry, as a small step towards artificial general intelligence.**

Keywords*—*data augmentation, convolutional neural network, symmetry invariant, group transformation, data preprocessing, SLAP, reinforcement learning

I. INTRODUCTION

A. Problem

Convolutional neural network (CNN) is now the mainstream family of models for computer vision, thanks to its weight sharing mechanism to efficiently share learning across the same plane by so-called kernels, achieving local translational invariance. But CNN is not reflection and rotation invariant. Typically, it can be addressed by data augmentation to inputs by reflection and rotation if necessary, but the sample size would increase substantially. [1] criticised CNN that it could not learn spatial relationships such as orientation, position and hierarchy and advocated their novel capsule to replace CNN. [2] improved capsule using routing by agreement mechanism and outperformed CNN at recognising overlapping images, but they also admitted that it tended to account for everything in the structure. This implies capsule is too heavy in computation. Inspired by the idea of capturing orientation information in capsule network [2], this paper proposed a novel method called Switchable Lightweight Anti-symmetric Process (SLAP), a protocol to produce the same output given different transformation variants, with the main research question: can symmetry variants be exploited directly by SLAP to improve and combine with CNN for machine learning?

Very often, we know in advance if a certain machine learning task is invariant to certain types of transformation, such as rotation and reflection. E.g. in Gomoku, the state is rotation (perpendicularly) and reflection (horizontally and vertically) invariant in terms of winning probability, and "partially" translation invariant. Symmetry is often exploited by data augmentation for deep learning. But this greatly increases the dataset size if all symmetry variants are included – e.g. there are 8 such variants for each Gomoku state. SLAP was invented in this paper to avoid such expansion (see I.B).

On the other hand, reinforcement learning is notorious for lengthy training time and large sample size required. Data augmentation may help improve performance in reinforcement learning, but it would increase the sample size. This research tried to kill two birds by one stone, SLAP, by applying with CNN in reinforcement learning (of Gomoku), challenging the widely used practice of data augmentation, aiming at reducing the sample size and improving the learning speed.

B. Switchable Lightweight Anti-symmetric Process (SLAP)

SLAP is a model-independent protocol and function to always produce or choose the same variant regardless of which transformation variant (by specified symmetry) is given, and if required also output the corresponding transformation. It can be used upon any function or model to produce outputs that are invariant regarding specified symmetric properties of the inputs. If some types of the outputs are not invariant but follow the same transformation, the corresponding transformation information from SLAP may be used to transform these outputs back. It can be viewed as standardization of symmetry, as opposed to standardization of scale. After processing, symmetric variants are filtered out – that's why it is named 'anti-symmetric process'. Ironically, with this anti-symmetric process, the function or model (e.g. CNN) to be fed would look as if it is symmetric with regard to whichever the symmetry variant is the input, and the same output is produced. It is a novel method to exploit symmetry variants in machine learning without increasing the number of training samples by data augmentation. The motivation is to concentrate experience to speed up learning, without enlarging the sample size by data augmentation. See details in III.A.

C. Gomoku

Gomoku, or Five in a Row, is a 2-player board game, traditionally played with Go pieces (black and white stones) on a Go board (19x19), nowadays on 15x15 board. For experiments in this research, mini board 8x8 was used instead to save computation, and the rule of freestyle version was adopted:

- Black (first) and white place stones of his colour alternately at an unoccupied intersection point.
- Winner: first one to connect 5 stones of his colour in a straight line (horizonal, vertical or diagonal).
- Draw happens if the board is full without a winner.

Gomoku was chosen to demonstrate the benefit of SLAP because:

- Gomoku has huge number of state representations ($3^{225}\sim= 2\text{x}10^{107}$), justifying the use of neural network.
- Gomoku is rotation and reflection invariant, but only "partially" translation invariant, so ideal to test different transformations.
- Gomoku is Markov Decision Process, meeting a basic assumption of reinforcement learning.
- [3] and [4] showed a general effective reinforcement learning algorithm for board games and Gomoku is simple to implement.

II. BACKGROUND

A. CNN

CNN (convolutional neural network) has been widely used for computer vision, but it is known that CNN is weak to deal with changes by rotation or orientation unless with much larger sample size by data augmentation. To address this problem, [1] proposed that neural network should make use of their then novel capsule, learning to recognize an implicitly defined visual entity and output probability of its existence and instantiation parameters such as pose; they showed that a transforming auto-encoder could be learnt to force the output (which is a vector instead of scalar) of a capsule to represent an image property that one might want to manipulate. [2] showed that a discriminatively trained, multi-layer capsule system achieves state-of-the-art performance on MNIST and was considerably better than CNN at recognizing highly overlapping digits, using the so-called routing by agreement mechanism, and yet [2] admitted that one drawback was the tendency of capsule to account for everything in an image. It implies that the capsule might be too "heavy" for computation and so a lightweight method is required. On lightweight capsule, DSC-CapsNet was proposed as lightweight capsule network, which focused on computing efficiency and reducing number of parameters [5]; [6] proposed dense capsule network with fewer parameters – neither had structure similar to SLAP. The capsule network with routing by agreement algorithm was proved by [7] not to be a universal approximator, i.e. not fit to all kinds of problems. As such, this research did not attempt to replace CNN by capsule, but simply created SLAP to combine with CNN. Instead of forcing the output to represent certain transformation information (e.g. orientation angle), SLAP forces the input of different variants (e.g. different rotation angle) to give the same output variant (and output the transformation information e.g. angle, if needed). Nevertheless, the invention of SLAP was inspired by [1] & [2] trying to address the weakness of CNN. On symmetric CNN, [8] proposed to impose symmetry in neural network parameters by repeating some parameters and achieved 25% reduction in number of parameters with only 0.2% loss in accuracy using ResNet-101, a type of CNN; but unlike SLAP, symmetry was not imposed in the inputs.

B. Groupoid in Gomoku

There are different Gomoku states of the same groupoid (see Fig. 1), which means having local symmetry but not necessarily global symmetry of the whole structure [9]. Groupoid is more challenging than symmetry or group, as some groupoids may not have the same status, e.g. see Fig. 1. But the potential for learning is huge as there are many more variants, e.g. 156 variants just by translation in Fig. 1.

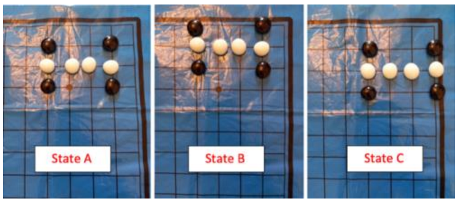


Fig. 1. Gomoku groupoid. Black can stop white win in C, but not in A or B.

C. AlphaGo Zero / AlphaZero

For reinforcement learning of Gomoku in this research, the baseline algorithm was chosen to follow that of AlphaGo Zero [3] and Alpha Zero [4] because domain knowledge was not required. The algorithm was concisely summarized by [10] as follows:

1) Neural Network

The neural network feature extractor is a type of CNN. It takes state s_t as input and yields value of state $v_\theta(s_t) \in [-1,1]$ and policy $\overrightarrow{p_\theta}(s_t)$ as probability vector over all possible actions. It has the following loss function (excl. regularization terms):

$$\text{loss} = \sum_t (v_\theta(s_t) - z_t)^2 - \vec{\pi}_t \,.\, \log(\overrightarrow{p_\theta}(s_t)) \tag{1}$$

where z_t, $\vec{\pi}_t$ are final outcome {-1,0,1} and estimate (to be discussed below) of policy from state s_t respectively, with 1, 0, -1 representing win, draw, lose respectively for current player.

2) Monte Carlo Tree Search (MCTS) as policy improvement operator

At each node, action is chosen by maximizing U(s, a), the upper confidence bound of Q-value Q(s, a):

$$U(s, a) = Q(s, a) + C * P(s, a) * \frac{\sqrt{\sum_b N(s,b)}}{1+N(s,a)} \tag{2}$$

where N(s, a) is no. of times taking action a from state s in MCTS simulation, P(s, .) is $\overrightarrow{p_\theta}(s)$, and the policy estimate of probability is improved by using (3):

$$\vec{\pi}_t = N(s, .) \,/ \sum N(s,b) \tag{3}$$

When a new node (not visited before from parent node) is reached, instead of rollout, the value of new node is obtained from neural network and propagated up the search path. Unless the new node is terminal, the new node is expanded to have child nodes.

3) *Self-play training as policy evaluation operator*

In each turn, a fixed number of MCTS simulations are conducted from the state s_t, and action is selected by sampling from the policy estimate of probabilities improved by MCTS, thus generating training sample data. At the end of an iteration, the neural network is updated by learning from the training sample data.

The evaluation metric would be based on winning and drawing percentages of the AI against an independent evaluation agent. There are differences among AlphaGo Zero [3] and Alpha Zero [4], see Table I:

TABLE I. DIFFERENCES BETWEEN ALPHAGO ZERO AND ALPHAZERO

	AlphaGo Zero	**AlphaZero**
Pitting models	Yes, model with new weights plays against previous one; new weights are adopted only if it wins 55% or above	No, always use new weights after each iteration of neural network learning
Symmetry	Data augmentation by rotation and reflection to increase sample size by 8 times for training; transform to one of 8 variants randomly in self-play for inference	Not exploited, as it is intended for generalization
Action in self-play	Sampled proportional to visit count in MCTS in first 30 moves, then selected greedily by max visit count (asymptotically with highest winning chance) in MCTS	Sampled proportional to visit count in MCTS
Outcome prediction	Assume binary win/loss, estimate & optimise winning probability	Also consider draw or other outcomes, estimate & optimise expected outcome

D. *Other Related Works, Symmetry and AGI*

[11] incorporated symmetry into neural network by creating symmetry (of specific type) invariant features, but no implementation or idea similar to SLAP was used. Studies have shown rotation-based augmentation performed better than many other augmentation techniques [12]. The type of data augmentation used as baseline in this research was rotation and reflection based (also the type used by AlphaGo Zero[3]). The novelty lies in the fact that SLAP is opposite to the practice of data augmentation – decreasing the variety of variants in the data instead for machine learning, though also exploiting symmetry.

Symmetry is one of the natures of the real world. Animals can detect the same object or the same prey being moved (translated), or even rotated after being slapped (the novel method was deliberately abbreviated as SLAP). Recognising symmetry can also speed up learning patterns, a typical trick used for playing some board games. To facilitate research exploiting symmetry in machine learning, [13] connected symmetry transformations to vector representations by the formalism of group and representation theory to arrive at the first formal definition of disentangled representations, expected to benefit learning from separating out (disentangling) the underlying structure of the world into disjoint parts of its representation. Upon this work, [14] showed by theory and experiments that Symmetry-Based Disentangled Representation Learning (SBDRL) could not only be based on static observations: agents should interact with the environment to discover its symmetries. They emphasized that the representation should use transitions rather than still observations for SBDRL. This was taken into account for designing the Gomoku representation for reinforcement learning in this research.

One may expect that an artificial general intelligence (AGI) system, if invented, should be able to learn unknown symmetry. Researchers have worked on this, for example [15] proposed learning unknown symmetries by different principles of family of methods. But it is equally important to learn by exploiting symmetry more effectively. For example, if an AGI system can interpret the rules of Gomoku and realize from the rules that Gomoku is reflection and rotation invariant, it should directly exploit such symmetry instead of assuming symmetry is unknown. Ideally, such exploitation should be switched on easily if one wishes, and hence the term 'switchable' in SLAP, which can be used upon any function or model. If transfer learning in CNN is analogous to reusing a chair by cutting the legs and installing new legs to fit another, such 'switchable learning' in SLAP is analogous to turning the switch of an adjustable chair to fit certain symmetries. Such kind of 'switch' in design can also help AI be more explainable and transparent, and more easily reused or transferred, while an AGI system should be able to link and switch to different sub-systems easily to solve a problem. SLAP can also reduce memory required. For example, AlphaGo Zero used a transposition table [3], a cache of previously seen positions and associated evaluations. Had SLAP been used instead of data augmentation, such memory size could be reduced by a factor of 8, or alternatively 8 times more positions or states could be stored. Indeed memory plays an important role in reinforcement learning as well by episodic memory, an explicit record of past events to be taken as reference for making decisions, improving both sample efficiency and speed in reinforcement learning as experience can be used immediately for making decisions [16]. It is likely that an AGI system would, just like human, use memory to solve some problems rather than always resort to learning from scratch. And in the real world, a continuous space, there can be much more than 8 equivalent variants. Recently, [17] suggested that symmetry should be an important general framework that determines the structure of universe, constrains the nature of natural tasks and consequently shape both biological and artificial general intelligence; they argued that symmetry transformations should be a fundamental principle in search for a good representation in learning. Perhaps SLAP may contribute a tiny step towards AGI, by shaping input representations directly by symmetry transformation. Note that SLAP can be used upon any function or model, and even if some (types) of the outputs are not invariant but follow the same transformation, these may be broken down and use the transformation information output from SLAP to make appropriate transformation back later for these parts only. A little kid often mistakes *b* for *d* at the beginning of learning alphabets, and it appears that human learning types of objects by vision might naturally assume symmetry first and then learn non-symmetry later. If a machine learning problem is to be split into stages or parts by specified symmetry as a guide, SLAP might help by wrapping certain parts of a function or neural network model.

III. METHODS

A. *SLAP*

SLAP forces the input of different variants (e.g. different rotation angle) to give the same output variant (and output the transformation information e.g. angle, though not necessarily used). There can be multiple ways to achieve this. For rotation and reflection variants of Gomoku states, one way to implement this is simply flattening the pixels of 8 variants to 8 lists, compare the lists and always choose the largest. Below (Fig. 2) was the algorithm used for SLAP in dealing with rotation and reflection variants of Gomoku states, but the concept may be applied to other symmetries as well.

Algorithm SLAP

```
1: Generate symmetry variants of input, store required transformation
2: Convert each variant to a list
3: Compare each list and find the ‘largest’ list
4: return the ‘largest’ variant & required transformation of the variant
```

Fig. 2. SLAP algorithm. Positive large data cluster towards top left.

If the image/state has multiple input channels or planes in one sample, the first channel/plane is compared first by list comparison.

SLAP was implemented by numpy instead of torch tensor for faster speed, because numpy uses view for rotation and reflection. The output variant replaced the input state when SLAP was applied in training. During inference time, output action probabilities from neural network would be transformed back using the transformation information (rotation & reflection) from SLAP.

1) *Invariance*

Denote s, t = slap(x_i), where slap is SLAP function in pythonic style, s is the symmetry (of certain group G, with n symmetry variants for each state) variant and t is corresponding transformation information. Given property of slap, for all $i \in \mathbf{N}_{<=n}$,

$$s, t = \mathrm{slap}(x_1) = \mathrm{slap}(x_2) = \ldots = \mathrm{slap}(x_n) \tag{4}$$

Denote s = slap(x_i)[0], t = slap(x_i)[1], the pythonic expression to capture first and second return variables of a function respectively. Denote h(slap(x_i)[0]) as $h^{slap}(x_i)$ for any function h. Given an arbitrary function y = f(x),

$$y = f^{slap}(x_i)$$

By definition, $\Rightarrow$ y = f(slap(x_i)[0])

Using (4), $\Rightarrow$ y = f(s) for all i (5)

$\therefore$ $y = f^{slap}(x_i)$ is invariant with respect to i (i.e. symmetry of group G).

When f is the neural network, the composite function resulting from the neural network, f^{slap}, is invariant to symmetry (of group G).

2) *Differentiability*

SLAP was not applied to intermediate layers of neural networks for Gomoku, so its differentiability was not required in this research. Approximation would be required to make it differentiable.

3) *Groupoid and SLAP-CC*

As Gomoku is only ‘partially’ invariant to translation, it is also interesting to experiment with translation variants, which are considered to be groupoid instead of group as they are symmetric locally but not necessarily symmetric globally. There can be many more translation variants than rotation and reflection variants, see II.B. To save computation, another algorithm (crop and centre) was used to ‘standardize’ translation variants. It was denoted as SLAP-CC in the below to emphasize that it shared the same general idea as SLAP, but just different way for implementation. Denoted as *cc* in the code.

The algorithm of SLAP-CC, shown in Fig. 3, would concentrate experience around the centre, as input variant was centred to become output variant. If it could not be exactly centred, the algorithm would make it slightly lean to top left.

Algorithm SLAP-CC

```
1: Find non-empty min & max indices by row & column in input image
2: r_shift = (no. of rows –1 –min row index –max row index)//2
3: c_shift=(no. of columns –1 –min column index –max column index)//2
4: return numpy.roll(image, (r_shift, c_shift), axis=(-2, -1))
```

Fig. 3. SLAP-CC algorithm. Non-zero data cluster towards centre.

Note that since Gomoku is not completely invariant to translation, SLAP-CC was used to add information as additional planes instead, as opposed to replacing the input state when SLAP was applied. 2 planes representing stones of different colours (current and opponent players respectively) centred together by SLAP-CC, followed by 2 planes representing original indices of vertical and horizontal positions respectively (scaled linearly to [1, -1]) were added along with original 4 planes in Gomoku state representation (see III.B). The scaled position indices for whole plane were to give neural network a sense of original positioning.

B. *Representation of Gomoku*

In this research, the representation of Gomoku followed the style of AlphaGo Zero / AlphaZero, with simplification and taking [14] into account for representation design.

For each Gomoku state, there were 4 planes representing current player stones, opponent stones, last action and current colour respectively by one-hot-encoding. See Fig. 4 for a typical Gomoku state in this research, which used simplified board size 8x8 instead.

Current player	Opponent player	Last action	Current Colour
0 0 0 0 0 0 0 0	0 0 0 0 0 0 0 0	0 0 0 0 0 0 0 0	1 1 1 1 1 1 1 1
0 0 0 0 0 0 0 0	0 0 0 0 0 0 0 0	0 0 0 0 0 0 0 0	1 1 1 1 1 1 1 1
0 0 0 0 0 0 0 0	0 0 0 0 0 0 0 0	0 0 0 0 0 0 0 0	1 1 1 1 1 1 1 1
0 0 0 1 0 0 0 0	0 0 0 0 0 0 0 0	0 0 0 0 0 0 0 0	1 1 1 1 1 1 1 1
0 0 0 1 0 0 0 0	0 0 0 0 1 0 0 0	0 0 0 0 0 0 0 0	1 1 1 1 1 1 1 1
0 0 0 0 0 0 0 0	0 0 0 1 0 0 0 0	0 0 0 1 0 0 0 0	1 1 1 1 1 1 1 1
0 0 0 0 0 0 0 0	0 0 0 0 0 0 0 0	0 0 0 0 0 0 0 0	1 1 1 1 1 1 1 1
0 0 0 0 0 0 0 0	0 0 0 0 0 0 0 0	0 0 0 0 0 0 0 0	1 1 1 1 1 1 1 1

Fig. 4. Gomoku state representation example at time t = 4.

For labels, probabilities of a move over all positions were represented by 8x8 flattened vector. Final outcome (value) of current player was represented by 1, 0, -1 respectively for win, draw, lose.

C. *SLAP in Gomoku Reinforcement Learning*

SLAP was used to pre-process states for network training and inference. Transformation information from SLAP was only used in network inference to convert probabilities (not estimated outcome) back to corresponding game board positions for MCTS to improve probabilities of actions, which were used as sampling probabilities to make a move in self-play (but greedy in evaluation). See Fig. 5.

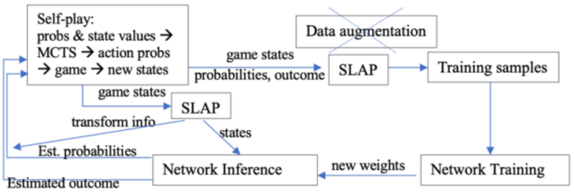


Fig. 5. SLAP used in Gomoku reinforcement learning.

For SLAP-CC, it was applied at the same place as SLAP in the above flow chart, but data augmentation was kept instead of being replaced and no transformation information was used to transform probabilities output of the network. See methods in III.A.3.

D. Testing Benefits for Neural Network Learning

To decouple from reinforcement learning dynamics, synthetic states of Gomoku were created for testing neural network learning with SLAP vs with typical data augmentation (by rotation and reflection), the latter of which had 8 times the number of training samples. Self-play was not involved in this testing.

Synthetic states were generated by first creating states each with only 5 stones connected in a straight line (i.e. win status) for all combinations for current black player, then removing one stone (to be repeated with another stone 5 times to create 5 different states) and randomly adding 4 opponent stones to become one about-to-win state. Together these were one set of 480 about-to-win states. Different sets could be created since white stones were merely random. Each set was mixed with 1000 purely random states, also with 4 stones for each player. 8 mixed sets were created, i.e. 11,840 states. 15%, i.e. 1,776 were reserved for validation test.

Labels were assigned as follows: if there were one or more choices to win immediately (include some purely random states, though the chance would be very remote), the value of state would be labelled as 1 and the wining position(s) would be labelled with probability of move = 1/no. of winning positions, while others were labelled 0; otherwise the value of state would be labelled as 0 and the probability of move for each available position would be random by uniform distribution, normalizing and summing to 1.

Neural networks (see Appendix) with SLAP vs with data augmentation would learn from training samples of states and labels to predict labels of validation data given the input states. Validation loss and its speed of convergence would be the key metrics.

First, at preliminary stage, for each set of hyperparameters the neural network ran 1000 iterations each with batch size 512 sampled from training samples of size 10,064 and 80,512 respectively for neural networks with SLAP and neural networks with data augmentation. Sampled with replacement, same as during reinforcement learning. There were 2400 combinations of hyperparameters by grid search, shown in Table II:

TABLE II. HYPERPARAMETERS TESTED AT PRELIMINARY STAGE OF CNN LEARNING

Hyperparameter	**Tested values**	**Remarks**
use_slap	True, False	False: data augmentation instead of SLAP
extra_act_fc	True, False	True: add extra layer (size 64) to action policy
L2	10^{-3}, 10^{-4}, 10^{-5}	weight decay of optimizer
Num_ResBlock	0, 5, 10, 20	no. of residual blocks
SGD	True, False	False: Adam optimizer
lr	10^{-1}, 10^{-2}, 10^{-3}, 10^{-4}, 10^{-5} (learning rate)	
dropout	0, 0.1, 0.2, 0.3, 0.4	

If Num_ResBlock > 0, the residual blocks replaced the common CNN layers and added a convolutional layer of 256 filters (3x3 kernel, stride 1, padding 1, no bias, ReLU activation) as the first layer. No autoclip [18] in optimizer, unlike reinforcement learning.

At stage 2, selected models from previous stage would run for 10,000 iterations instead of 1,000 iterations, with losses recorded every 10 iterations.

E. Testing Benefits for Reinforcement Learning

The baseline algorithm of Gomoku reinforcement learning followed AlphaGo Zero/AlphaZero (see II.C). Among their differences, the baseline algorithm in this research followed the better version, and thus followed AlphaZero except on symmetry exploitation. Like AlphaGo Zero, the baseline exploited symmetry by data augmentation to increase no. of training samples by 8 times, but random transformation was not done in self-play. Autoclip to gradients [18] was added in the optimizer for stable learning.

Reinforcement learning required much more computation than neural network learning, so to save computation, the same neural network will be used and the testing of hyperparameters would be based on best models in neural network learning by synthetic Gomoku states, with some deviations to be tested by grid search.

Stage 1: each of 240 models were trained by self-play of 250 games. Data buffer size: 1,250 and 10,000 for SLAP and non-SLAP models respectively, both roughly equivalent to storing latest 60 games.

Stage 2: selected models were trained by self-play of 5000 games. With more games arranged for training, larger data buffer size could be used. So, data buffer size was increased to 5000 and 4000 respectively for SLAP and non-SLAP models, roughly equivalent to storing latest 250 games. To align with stage 1 testing initially, the initial data buffer size was kept same as stage 1 for first 1000 games. This also got rid of initial poor-quality game state data quickly. Learning rate multiplier was used to adaptively decrease learning rate by half if validation loss increased beyond 3-sigma limit, measured every 100 games.

Evaluation: Independent agent(s), also called evaluation agent or evaluator, was built by pure Monte Carlo Tree Search (MCTS) with random policy to play against the trained AI. The strength of a pure MCTS agent depends on no. of playouts (aka. simulations) in each move. To facilitate observation of growing strength, multi-tier evaluation was built by playing 10 games against each of 3 pure MCTS agents (30 games total), each with 1000, 3000, 5000 playouts respectively. Overall winning rate (tie counted as half win) against them would be the key metrics for reinforcement learning. It was often either a win or loss, and seldom a tie. Assuming that a tie could be neglected, especially after counting tie as half win, it simplified as Bernoulli distribution with standard deviation approximated by $\sqrt{p(1-p)/30}$ to calculate confidence interval, where 30 is the number of trials in each evaluation.

F. Code Implementation

The part regarding AlphaZero was upgraded from [19]. Details of implementation and code repository:
https://github.com/chihangs

IV. RESULTS

A. *Impact on Neural Network Learning*

1) SLAP vs baseline (data augmentation)

The best few SLAP and baseline models converged to loss around 2.81 (difference < 0.01), all without residual blocks. 3 SLAP models (denoted as s0_...) and 3 baseline models (denoted as n0_...) were selected and their losses were plotted in Fig. 6, where each model had Adam optimizer, same learning rate 0.001, no dropout, no residual blocks, but different values of L2 (10^{-3}, 10^{-4}, 10^{-5}).

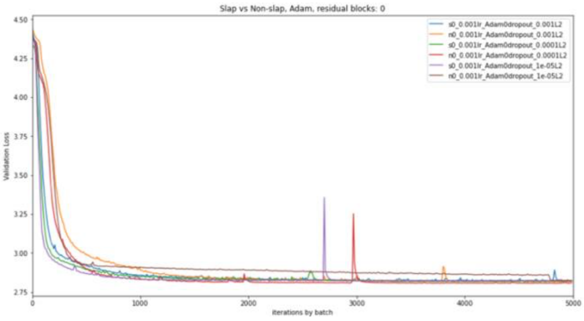


Fig. 6. Validation losses of SLAP and baseline models.

Above 6 models were repeated 3 more times to calculate average time (by no. of iterations) for convergence. SLAP speeded up the convergence by 95.1% and 71.2% measured by validation loss reaching 3.0 and 2.9 respectively, 83.2% in average.

2) Testing sample size

Holding validation dataset unchanged, the training data sample size was reduced by holding out some samples to match required size, using models with L2=10^{-4} from Fig. 6. SLAP models converged when sample size was 5032 or above, but they were more vulnerable to decreasing no. of training samples and failed to converge when the sample size decreased to 2516 or below, while their baseline counterpart models (8 times the sample size) still converged.

3) SLAP-CC vs baseline (data augmentation)

SLAP-CC (see III.A.3) was added to the 3 best baseline models from Fig. 6. Validation losses of SLAP-CC converged to around 2.8 for all 3 values of L2, like its baseline counterparts. Experiments were repeated 3 more times to calculate average time (by no. of iterations) to converge. The time for validation loss to reach 3.0 and 2.9 both worsened by 30.7% in average for SLAP-CC.

B. *Impact on Neural Network Learning*

1) SLAP vs baseline (data augmentation)

The best SLAP model had highest winning rate 86.7%, equivalent to winning 26 games out of 30. 95% confidence interval was 86.7% +/- 12.2%, i.e. (74.5%, 98.9%). The best baseline model had highest winning rate 93.3%, equivalent to winning 28 games out of 30, with 95% confidence interval 93.3% +/- 8.9%, i.e. (84.4%, 100%).

The best SLAP and baseline models had similar winning rates, by confidence intervals. If winning rate of two thirds (66.6%) is used as benchmark for this three-tier evaluation, both took 1000 games to achieve or surpass this. However, non-SLAP took 1250 games only to first achieve winning rate of 86.6%, while SLAP took 3000 games.

SLAP spent 0.761 second per move in self-play, 10.8% more time than baseline (only 5% more in a separate speed-optimizing version). SLAP tended to decrease learning rate multiplier more frequently, implying more frequent significant increase of validation loss.

2) Testing buffer size

Best models of SLAP and non-SLAP were repeated but with smaller data buffer size of only 1,250 and 10,000 respectively throughout whole reinforcement learning. Similar to stage 2, above models were trained by 5000 games. With fewer data in buffer, the highest winning rate achieved for SLAP model was only 73.3%, below the corresponding confidence interval. The highest winning rate achieved for non-SLAP model was only 83.3%, below the corresponding confidence interval. So, it harmed reinforcement learning when data buffer was too small, and it was good decision to use larger data buffer at stage 2.

3) SLAP-CC vs baseline (data augmentation)

SLAP-CC was tested by same configurations as best baseline model from IV.B.1 but adding information from SLAP-CC and scaled position indices as extra input feature planes. The new model also ran for 5000 games. See methods in III.A.3 and III.C. Learning rate multiplier did not change throughout training. The best winning rate achieved for SLAP-CC model was 96.7%, slightly higher than the baseline, but within the confidence interval.

V. DISCUSSION

Despite the widely use of data augmentation to increase the variety of transformation variants in samples to improve machine learning, we proved that using SLAP to decrease the variety could achieve the same performance of typical data augmentation with sample size reduced by 87.5% and faster by 83.2% in CNN network learning, and statistically the same performance for reinforcement learning with sample size reduced by 87.5%. The success could be explained by concentrating learning experience to certain regions when different variants were transformed by SLAP, implicitly sharing weights among variants. The proof of invariance (see III.A.1) after applying SLAP did not require the network to be CNN and it could be an arbitrary function, so the applicability of SLAP should not be restricted to CNN. While SLAP exploited only reflection and rotation symmetries in learning Gomoku, the general concept of SLAP and the proof of invariance could apply to other symmetries. As no domain specific features or knowledges (except symmetry) were used in SLAP, the benefits shown in the experiments should apply generally for domains that are symmetry invariant.

Shortcomings: in Gomoku reinforcement learning, SLAP tended to decrease learning rate multiplier more frequently, implying more frequent significant increase of validation loss. This instability could be caused by faster neural network learning. Note that AlphaGo Zero only dropped learning rate twice over 1,000,000 training steps in their planned schedule [3]. It might imply that SLAP would need quite different hyperparameters in reinforcement learning (as opposed to sharing the same hyperparameters of baseline models in the neural network learning experiment), and more or better searches of hyperparameters for reinforcement learning would be required, though it was constrained by computation resources. Another plausible explanation for not speeding up reinforcement learning was the insignificant portion (~1%-2%) of neural network learning in the whole reinforcement training, implying that the time saved in neural network learning would be insignificant for the whole reinforcement learning in our chosen setting (which used a relatively simple CNN), and enough neural network learning iterations would have been allowed if hyperparameters were optimal.

Limitations: the results only applied to symmetry-invariant domain, and SLAP could be more vulnerable if the sample was too small (see IV.A.2). SLAP required 10.8% more time for self-play in IV.B.1, but the overhead would be insignificant if the simple CNN were replaced by a deep one. It was not yet proved to speed up reinforcement learning. Neither was it proved to be able to exploit groupoid patterns.

VI. CONCLUSION AND FUTURE WORK

SLAP could improve the convergence speed of neural network (CNN in the experiment) learning synthetic states in Gomoku by 83.2%, with only one eighth of training sample size of baseline model (data augmentation). Since no domain specific features or knowledges were used in SLAP, it should also benefit neural network learning generally for domains that are symmetry invariant, especially for reflection and rotation symmetries. As SLAP is model-independent, the benefits should apply to models beyond CNN. But it was not yet proved to speed up reinforcement learning, though it could achieve similar performance with smaller training sample size. Neither was it proved to exploit groupoid variants effectively.

As future work, SLAP may be applied in domains that are not fully symmetry invariant, by breaking down the neural network layers into two parts – first learning as if it were fully symmetry invariant. Or even split into stages by type of symmetries. Although SLAP is not directly differentiable, one workaround would be similar to that in transforming Gomoku action probabilities: given the transformation information as another input, transform the learned output back to corresponding original position, and then carry out necessary subsequent computations forward. This helps create more explainable stages and transfer learning. Another future work might be differentiable approximation of SLAP.

APPENDIX

Neural Network Architecture and Configurations

The architecture and configurations used (unless otherwise stated):

Architecture: consisted of 3 common convolutional layers (32, 64, 128 filters respectively) each with 3x3 kernel of stride 1 and padding 1 with ReLU activation, followed by 2 action policy players and in parallel 3 state value layers. The input was 8 x 8 x 4 image stack comprising of 4 binary feature planes. The action policy layers had one convolutional layer with 4 filters each with 1x1 kernel of stride 1 with ReLU activation, followed by a fully connected linear layer to output a vector of size 64 corresponding to logit probabilities for all intersection points of the board. The state value layers had one convolutional layer with 2 filters each with 1x1 kernel of stride 1 with ReLU activation, followed by fully connected linear layer to a hidden layer of size 64 with ReLU activation, finally fully connected to a scalar with tanh activation. Dropout would be applied to all action policy layers and state value layers except output layers; not applied to common layers.

Optimizer: Adam with autoclip [18]
Batch size per optimisation step: 512 (2048 in [3])
Data buffer size: 10,000 for baseline, 1,250 for SLAP
No. of network optimisation steps per policy iteration: 10
No. of self-play games per policy iteration: 1
No. of playouts: 400 (1600 in [3], 800 in [4])
C_{puct} (constant of upper confidence bound in MCTS) : 5
Temperature parameter: 1 (same as in [4])
Dirichlet alpha of noise: 0.3 (same as chess in [4])

Smaller batch size and no. of playouts per move in MCTS were used because Gomoku is less complex than Go. Dirichlet alpha was initially set at 0.3 because mini Gomoku (8x8 board) has same board size as chess and similar no. of available action choices per move.

REFERENCES

[1] G. E. Hinton, A. Krizhevsky and S. D. Wang, 'Transforming auto-encoders' in *International Conference on Artificial Neural Networks (ICANN)*, 2011.

[2] S. Sabour, N. Frosst and G. E. Hinton, 'Dynamic routing between capsules' in *31st Conference on Neural Information Processing Systems (NIPS 2017)*, Long Beach, CA, USA, Oct 2017.

[3] D. Silver et al., 'Mastering the game of go without human knowledge', *Nature,* 550, pp. 354– 359, Oct 2017.

[4] D. Silver et al., 'Mastering chess and Shogi by self-play with a general reinforcement learning algorithm', *Science*, Vol 362, Issue 6419, pp. 1140-1144, Dec 2018.

[5] S. Dan, X. Hu, Y. Zhou and S. Duan, 'Lightweight multi-dimensional memristive CapsNet' in *International Joint Conference on Neural Networks (IJCNN),* Shenzhen, China, Jul 18-22 2021.

[6] K. Sun, X. Wen, L. Yuan and H. Xu, 'Dense capsule networks with fewer parameters', *Soft computing (Berlin, Germany)* (1432-7643), Vol 25 (Issue 10), pp. 6927, Apr 2021.

[7] D. Peer, S. Stabinger and A. Rodriguez-Sanchez, 'Limitations of capsule networks', *ScienceDirect (Pattern Recognition Letter)*, Vol 144, pp. 68-74, Apr 2021.

[8] X. S. Hu, S. Zagoruyko and N. Komodakis, 'Exploring weight symmetry in deep neural networks' 2018, *arXiv:1812.11027.*

[9] A. Vistoli, 'Groupoids: a local theory of symmetry', *Isonomia (Epistemologica)* 2011, 26, pp. 1–12, Sep 2011.

[10] S. Nair. 'A simple Alpha(Go) Zero tutorial' Stanford University. Accessed: 18 May 2022. [Online.] Available: http://web.stanford.edu/~surag/posts/alphazero.html

[11] D. Bergman, 'Symmetry constrained machine learning' *2019, arXiv:1811.07051v2.*

[12] K. Maharana, S. Mondal and B. Nemade, 'A review: data pre-processing and data augmentation techniques', *Global Transitions Proceedings*, vol. 3, Issue 1, pp. 91-99, Jun 2022.

[13] I. Higgins et al., 'Towards a definition of disentangled representations' *2018, arXiv:1812.02230.*

[14] H. Caselles-Dupré, M. Garcia-Ortiz and D. Filliat, 'Symmetry-based disentangled representation learning requires interaction with environments' in *33rd Conference on Neural Information Processing Systems (NeurIPS 2019), Vancouver, Canada, 2019.*

[15] F. Anselmi, G. Evangelopoulos, L. Rosasco and T. Poggio, 'Symmetry regularization', *The Centre for Brain, Mind and Machines (CBMM) Memo No. 63*, May 2017.

[16] M. Botvinick et al., 'Reinforcement learning, fast and slow' *Trends in Cognitive Sciences*, Vol 23, Issue 5, pp. 408-422, Apr 2019.

[17] I. Higgins, S. Racanière and D. Rezende, 'Symmetry-based representations for artificial and biological intelligence', *Frontiers in Computational Neuroscience*, vol. 16, 2022, doi: 10.3389/fncom.2022.836498.

[18] P. Seetharaman, G. Wichern, B. Pardo and J. L. Roux, 'AutoClip: adaptive gradient clipping for source separation networks' in *2020 IEEE 30th International Workshop on Machine Learning for Signal Processing (MLSP)*, Espoo, Finland, Sep 21-24 2020, doi: 10.1109/MLSP49062.2020.9231926.

[19] J. Song. 'An implementation of the AlphaZero algorithm for Gomoku (also called Gobang or Five in a Row)' github.com. Accessed: Jun-Sep 2022. [Online.] Available: https://github.com/junxiaosong/AlphaZero_Gomoku

Part IV.

AI Ethics

Neighbor Migrating Generator: Finding the closest possible neighbor with different classes

Hassan Eshkiki
Department of Computer Science
Swansea University
Swansea, Wales
H.G.Eshkiki@swansea.ac.uk

Benjamin Mora
Department of Computer Science
Swansea University
Swansea, Wales
B.Mora@swansea.ac.uk

***Abstract*—The Neighbor Migrating Generator (NMG) is a simple and efficient approach for identifying the closest potential neighbor(s) with a different label for a given instance without the need to calibrate any kernel settings. This allows for the determination and explanation of the most important features that influence an AI model.**

The NMG technique can be used to migrate a specific sample to the class decision boundary of the original model within a close neighborhood of that sample or to identify global features that help localize neighbor classes. The approach minimizes a loss function that is divided into two components, which are independently weighted according to four parameters: α, β, and ω, with α being self-adjusting. Results show that this approach outperforms past techniques when detecting the smallest changes in the feature space and can also highlight issues in models like overfitting.

***Index Terms*—Explainable AI, Interpretable AI, AI, Machine learning, Counterfactual models,**

I. INTRODUCTION

Many end-users who are not machine learning experts often request a model's description and the reason behind its decision. Understanding why a given decision was made can be highly beneficial, especially in professions where risk assessment and analysis are critical, such as finance and medicine. To meet this demand, Explainable AI has recently gained attention.

One of the current approaches to Explainable AI involves analyzing the effect of small changes to an input and how they impact the output of a black box model. This enables the identification of the most influential features in the input space. Various open-source frameworks, such as [1], [2], and [3], have implemented this approach, making AI more accessible to non-machine-learning experts. However, these approaches may have limitations, as results can vary with each execution.

Another approach to analyzing a model involves treating it as a black box and identifying the smallest input adjustments that result in a different output. This technique can be useful, for example, when a loan application is denied, and suggestions are needed to influence the bank's decision.

In this paper, we propose a technique that improves on past research [4] by estimating the nearest location to the boundary from the loss function used by our model to locate the decision border. We demonstrate the utility of our Neighbor Migrating Generator (NMG) in both a global and local context for three datasets, including a synthetic mathematical model, the Iris dataset, and the Pima Indians Diabetes dataset in the result section.

II. LITERATURE REVIEW

Individual predictions made by experts rely on the fact that humans can learn enough about what parameters have the biggest influence in the feature space or why a model inferred a particular decision. Many ways to explain individual predictions utilize local interpretation tools, such as Individual Conditional Expectation (ICE) [5], Counterfactual Explanations [4] [6], Local Interpretable Model-agnostic Explanations (LIME) [7], and SHAP methods [8], which assign weights to each dimension of inputs to explain a model.

ICE [5] is a global method for plotting the global effect of modifying each feature individually. ICE can modify one feature while retaining the same value for the rest of the feature space. It then displays the changes in predictions according to the value of the target feature. However, comprehending the differences between individual lines in the plot can be difficult at times. This can be resolved by centering the curves at a point and comparing the predictions to this point, which is known as Centered ICE (c-ICE). However, ICE cannot allow one to see and detect any association between features [9], which is a significant limitation.

LIME [7] generates a linear approximation of the local decision boundary by considering a trained model as a block box since the reasons behind a prediction for a given input are not understandable by users. It develops local surrogate models to explain the provided input and the decisions of the trained model. The explaining process consists of three steps. First, LIME alters the feature space values (creating a new training set) and, by using Gaussian kernels, associates a weight to the new sample points based on the distance to a given input whenever a new instance is passed to the LIME model. The new samples are then sent to the trained model to approximate the local decision boundary, which would be used to map the inputs and outputs together (to extract explanations). LIME explains a model in a feature space locally. Laugel et al. have shown [10] that LIME is highly dependent on some kernel settings because it relies on the correct definition of the local neighborhood. This problem

Samples / Class migration	Number of Pregnancies	Glucose mg/dl ≤ 140	Diastolic Blood Press., mm Hg ≤ 80	Skin Thickness mm	Insulin $\mu U/ml$ $16 - 166$	BMI kg/m^2 ≤ 25	Diabetes Pedigree Function	Age year
$1 : 1 \rightarrow 0$	2 **+3.5** (+175%)	100	66	20	90	32.9	0.867 **-0.1** (-12%)	28
$2 : 0 \rightarrow 1$	4	129 +18.8 (15%)	86	20	270	35.1	0.231	23
$3 : 1 \rightarrow 0$	3	169	74	19 **-0.1** (-1%)	125 **-21.9** (-18%)	29.9	0.268	31
$4 : 0 \rightarrow 1$	2	101 **+0.9** (+1%)	58	35 **+0.2** (+1%)	90 **-88.4 (-98%)**	21.8 **+2.2** (+10%)	0.155	22 **+1.1** (+5%)
$5 : 0 \rightarrow 1$	3 **+1.3** (+43%)	96	56	34 **+0.6** (+2%)	115	24.7 **+1.3** (+5%)	0.944	39
$6 : 0 \rightarrow 1$	1	118	58	36	94 **-23.2** (-25%)	33.3	0.261	23
$7 : 0 \rightarrow 1$	0	114 **+46.3** (+41%)	80 **-30.1** (-38%)	34 **-19.9** (-59%)	285 **-186** (-65%)	44.2 **-22.4** (-51%)	0.167 **+0.1** (+60%)	27 **+37.9** (+140%)
$8 : 0 \rightarrow 1$	1	87	68	34	77	37.6	0.401	24 **+4.4** (+18%)
$9 : 0 \rightarrow 1$	5 **-4** (-80%)	96 **+11** (+11%)	74 **-6.1** (-8%)	18 **+1** (+6%)	67 **-54 (+81%)**	33.6 **-6.5 (-19%)**	0.997 **-0.8** (-80%)	43 **-18.9** (-44%)
$10 : 1 \rightarrow 0$	3	107 **-5** (-5%)	62	13	48	22.9	0.678	23
$11 : 1 \rightarrow 0$	9 **-4.7** (-52%)	156 **-45.4** (-29%)	86	28	155	34.3 **+0.9** (+3%)	1.189 **-0.5** (-42%)	42
$12 : 1 \rightarrow 0$	1 **-0.1** (-10%)	128 **-9.7** (-8%)	48 **+37.9 (+79%)**	45 **-5.7** (-13%)	194 **+25.9 (+13%)**	40.5 **+5 (+12%)**	0.613 **+0.2** (+33%)	24 **+1.8** (+8%)

Table I

Variation among features when changing instances (local variant of the algorithm) for the Pima Indian Diabetes Dataset for 12 randomly selected samples. This dataset contains several characteristics for patients undergoing an oral glucose tolerance test, including insulin and glucose levels after two hours and the type 2 diabetes status (0 being non-diabetic). Units and standard ranges for glucose, blood pressure, insulin and Body Mass Index are given at the top, and samples outside the range are highlighted in red. The second value in cells, when present, indicates the variation generated by the migration model on some of the features. Unchanged values are not shown.

is not only a parameter or sampling distribution issue but has a significant impact on the relevance and quality of the local black-box decision boundary approximation, and thus on the meaning and accuracy of the given explanation. Guidotti et al. have looked into the development of new local train sets [11]. However, there are certain difficulties with their solutions, such as losing sight of a crucial notion like the tangent or inspecting a small neighborhood without taking linearity into account in the ML function.

Based on cooperative game theory, SHapley Additive exPlanations (SHAP) [8] assigns an importance value to each feature of a model. The aim of cooperative game theory is to determine how to fairly compensate all participants based on their contribution. SHAP generates the Shapley score by replacing a particular feature value with new values and then applying the trained model's predictions to the new dataset. However, analyzing all input dimensions can be expensive, so we only need to investigate a subset of them, resulting in slightly different outcomes each time. Another limitation of SHAP is that the order of features for a given instance may affect the prediction.

Our approach is similar to others mentioned above in that we treat the model as a black box with no access to the internal structure or weights. The main difference is that we aim to find the closest available neighbor that belongs to a different class.

In contrast to most existing techniques in the area, Counterfactual Explanations [4] and [6] are introduced to highlight the smallest modifications needed for a sample to change the prediction made by the model. This is different from the vast majority of approaches, which differ in how they define the loss function. For example, Wachter et al. [4] minimize the loss function as shown in equation 1.

$$arg \min_{x'} \max_{\lambda} \lambda(\hat{f}(x') - y')^2 + d(x, x') \tag{1}$$

There, the approach adds the Manhattan distance (weighted with the inverse median absolute deviation) between the original input and the modified one to the square of the L^2 norm of the differences between the predictions of these inputs. The weight λ indicates how influential the input instance can be, where a higher value indicates that we do not wish to alter values frequently. However, categorical features pose a problem with this approach. To address this issue, Dandl et al. [6] proposed a new loss function that minimizes a four-section loss and counts the number of altered features using the Gower metric and a counter.

$$\begin{split} L(x, x', y', x^{obs}) = o_1(\hat{f}(x'), y'), o_2(x, x'), \\ o_3(x, x'), o_4(x', X^{obs})) \end{split} \tag{2}$$

The Manhattan distance between $\hat{f}(x')$ and y' is denoted as o_1. The Gower distance function (o_2) calculates the distance between the original sample x and the close target sample x' on the decision boundary. The number of features that

Item	Descriptions
Dataset	Implicit Heart Function Dataset
Model F	Input(2),FC(256),RelU,FC(256),RelU,FC(2),Soft-Max
Model G	Input(2),FC(256),RelU,FC(256),RelU,FC(256),RelU,FC(2)
Dataset	PIMA
Model F	Input(8),FC(256),RelU,FC(256),RelU,FC(2),Soft-Max
Model G	Input(8),FC(256),RelU,FC(256),RelU,FC(256),RelU,FC(8)
Dataset	IRIS
Model F	Input(4),FC(128),RelU,FC(128),RelU,FC(3),Soft-Max
Model G	Input(4),FC(1024),RelU,FC(1024),RelU,FC(1024),RelU,FC(4)

Table II

A DESCRIPTION OF THE ARCHITECTURES OF THE MODELS USED IN THE THREE DATASETS. THE F MODEL REPRESENTS THE ORIGINAL MODEL THAT WE AIM TO STUDY, WHILE G REFERS TO THE NMG MODEL THAT IS UTILIZED TO MIGRATE TO DECISION BOUNDARIES

Feature Name	Description	Range
Pregnancies	Number of pregnancies	[0..17]
Glucose	Oral glucose tolerance test, measured after 2 hours.	[0..199]
BloodPressure	Diastolic blood pressure	[0..122]
SkinThickness	Triceps skin fold thickness	[0..99]
Insulin	Insulin level after 2 hours	[0..846]
BMI	Body Mass Index	[0..67.1]
DiabetesPedigreeFunction	Diabetes history in relatives	[0.08..2.42]
Age	Age in Years	[21..81]

Table III

DESCRIPTION OF FEATURES IN THE PIMA DIABETES DATASETS.

have been altered is calculated by o_3, while o_4 calculates the distance between the estimated sample and the nearest observed sample in the training set.

Our work is similar to the two approaches described above, where a model is trained for each case to determine the minimum modification required to change the class label for prediction. However, our approach differs in how we define a loss function that allows the model to freely change along with the feature space. Furthermore, we have introduced a new way of identifying the most influential features for a given model by training its NMG model representative of the decision boundaries with all the training set samples (instead of a single sample) and analyzing the modifications. Our proposed strategy addresses the challenge of freezing features, particularly when modifying some features is undesirable.

III. THE NEIGHBOR MIGRATING GENERATOR MODEL

In this section, we will introduce our new model and its two variations, as well as how to train them. The two variations differ depending on whether we want to provide local or global explanations. For a local explanation, the model is trained to focus on the nearest decision boundary to a specific sample, while the global explanation will identify the most influential features for each class. Both variations use the same architecture, but they are trained using either a single sample (for local explanations) or all samples (for global explanations). We associate a new model, denoted as G, with an original model, F, that we aim to explain.

A. Local and Global Variants

In the local explanation variant, the NM generator employs common machine learning techniques to identify a neighbor. A single input is fed through the fully connected network G, and we minimize the loss function explained later to approach a point that is close to the original sample but has a different target label, which is specified by the user. Consequently, the NMG strives to recognize and alter only the primary features that contribute to that class, indicating the most influential factors for a given sample.

Another potential use of our model is to observe feature migration in the global space, which provides insight into the most significant factors for a given class. In this variant, instead of specifying the target label for migration, we generate labels based on the predictions of a trained model F. We use the second most likely (or predicted) class label as the target label and train our Neighbor Migrating Generator G on the entire training set. It is important to note that for the global variant, an NMG uses the same model for all samples, which limits the impact of outliers and produces more generic results. In contrast, for the local variant, we train the model separately for each input, generating different weights each time. However, with the global variant, we may not be able to establish a minimum loss cost for outliers, but this would not affect the final results.

The global approach can be formalized as follows. Let d be the number of classes, n be the number of samples, and F be the original network that maps an input sample x_i to a vector of inferred probabilities for each class $F(x_i)$, with each component represented as $F_j(x_i)$. We define the best prediction of our original network $H_F^1(x_i)$ as:

$$H_F^1(x_i) = \arg\max_{j \in d} F(x_i) = \{j \mid F_j(x_i) \leq F_k(x_i) \, \forall k \in d\} \tag{3}$$

and the second best class prediction $H_F^2(x_i)$ as:

$$\begin{gathered} H_F^2(x_i) = \arg\max_{j \in d} F(x_{i,j}) = \\ \{j \mid F_j(x_i) \leq F_k(x_i) \, \forall k \in d, k \neq H_F^1(x_i)\} \end{gathered} \tag{4}$$

We can now define our new NMG model G trained from a set of samples $\mathcal{D}$ such that:

$$\begin{gathered} G : \mathcal{D} \to \mathcal{D} \\ x'_i = G(x_i) \mid H_F^2(x_i) = H_F^1(x'_i) \wedge \\ \forall x'' \in \mathcal{D}, \, H_F^2(x_i) = H_F^1(x'') \wedge \\ \| x_i - G(x_i) \|_l \leq \| x_i - x'' \|_l \end{gathered} \tag{5}$$

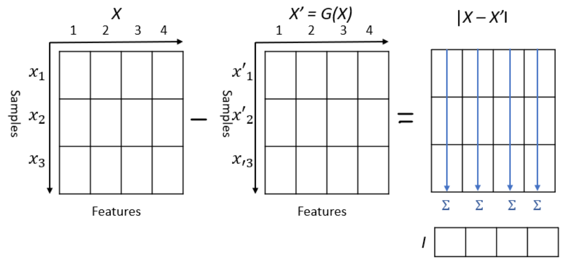
with l indicating the norm used.

After the training process, we calculate, for each input feature, the sum of differences between the input sample x_i and the nearest corresponding sample with a different label, defined as $x'_i = G(x_i)$. x'_i is essentially an imaginary sample that we want located on the class decision boundary in the feature space and that intends to be as close as possible to the initial sample x_i. For a given feature f, this is equal to $I_f = \sum_{i=1}^{n} |x_i(f) - x'_i(f)|$, as shown in Fig. 1. Larger values

Description	Age	Sex	TSH	T3	TT4	T4U	FTI	TBG	Referral source
Class 1	58	0	1.3	2.3	77	0	0	0	1
Migrated Class 0	58	0	**10.69**	2.30	**77.01**	-0.01	0.01	0	1
Difference	-5.38E-04	1.97E-04	**9.39**	1.20E-03	**0.012**	-6.81E-03	1.04E-02	0	9.14E-05
Class 0	63	1	8.20	2.10	80	1.02	78	0	1
Migrated Class 1	63	1	**2.37**	2.10	**80.22**	1.02	78.01	0	1
Difference	6.48E-05	3.49E-05	**-5.83**	-1.88E-03	**0.22**	-4.66E-03	1.14E-02	0	8.45E-05
Class 0	60	1	8.10	1.80	59	0.96	61	0	1
Migrated Class 1	60	1	**1.24**	1.79	**59.57**	0.96	61.01	0	1
Difference	-1.53E-05	5.58E-05	**-6.86**	-7.10E-03	**5.67E-01**	-3.45E-03	1.24E-02	0	1.29E-05
Class 1	29	1	0	0	0	0	0	0	1
Migrated Class 0	29	1	**10.30**	0.03	**-1.31**	0.01	0	0	1
Difference	-2.48E-03	9.05E-05	**10.3**	2.89E-02	**-1.31**	1.29E-02	-3.61E-03	0	3.18E-05
Class 1	59	1	0	0	0	0	0	0	1
Migrated Class 0	59.01	1	**10.30**	0.02	-1.05	0.01	-0.01	0	1
Difference	9.94E-03	9.15E-05	**10.3**	2.05E-02	-1.05	9.11E-03	-1.42E-02	0	2.38E-07
Class 1	71	1	0	0	0	0	0	0	1
Migrated Class 0	70.98	1	**10.20**	0.02	**-1.04**	0.01	0.04	0	1
Difference	-1.55E-02	1.68E-04	**10.2**	2.11E-02	**-1.04**	9.31E-03	4.33E-02	0	2.05E-05

Table IV
THE CORRESPONDING VALUES OF SIX RANDOMLY SELECTED SAMPLES IN THE THYROID DATASET AFTER NMG TRANSFERRED THEM TO ANOTHER CLASS. THE DIFFERENCES BETWEEN THE TWO VALUES ARE SHOWN BELOW. THE RESULTS INDICATE THAT WHILE NMG MODIFIES THE CLASS LABELS, THE ω VECTOR CAN PRESERVE SOME FIELDS UNTOUCHED, SUCH AS AGE, SEX, FTI, TBG, AND REFERRAL SOURCE. THE TOP TWO MOST CHANGED FIELDS ARE HIGHLIGHTED IN BOLD, WHICH INDICATES THE PATTERNS LEARNED BY THE MODEL. SPECIFICALLY, TSH AND TT4 ARE THE FEATURES THAT ALLOW MODIFYING THE CLASS LABELS.

Figure 1. The calculation of the Feature Importance weights I returns an explanation of the global impact of features for the different classes of the original model.



of I indicate that a feature has a larger impact and is therefore a major contributor in explaining how the model distinguishes between classes. Algorithm 2 shows a global method approach, and an application of this will be presented in section IV-D.

Algorithm 1: Algorithm for calculating the loss function at each iteration

```
$\lambda \leftarrow 0$ /*Importance of the inner loss*/;
$\epsilon \leftarrow 0.05$  /*A constant for the the thickness of the
    decision boundary*/ ;
for each iteration do
    if ($y == y'$) then
        $\lambda \leftarrow clip(\lambda + 1.0, 100.0)$;
    else
        $\lambda \leftarrow clip(\lambda - 1.0, 0.0)$;
    end
    $in\_loss \leftarrow (\lambda/100.0) * (\sum((x'_i - x_i) * \omega))$ ;
    $\beta_{y,y'} \leftarrow |CrossEntropy(1, y == y')$
            $-\epsilon|$;
    $out\_loss \leftarrow CrossEntropy(y, y')$;
    $loss = in\_loss + out\_loss + \beta_{y,y'}$
end
```

B. Loss function

Algorithm 1 shows how the loss function for G is calculated during each iteration. The loss function that is to be minimized is the same for both variants, as given in equation 6.

$$L(x, x', y, y') = CrossEntropy(y, y') + \lambda \,.\, L^1(x - x').\omega + \beta_{y,y'} \tag{6}$$

Where y is the class label for x, and y' is the predicted label for x'. λ is a weight that is recalculated after each training step, and $\beta_{y,y'}$ is a cross-entropy function as discussed in the next paragraph. ω is a user-defined weight vector that determines which input dimensions are more desirable to change with the default value of one. Lower values suggest a greater likelihood of change.

The equation for the loss function is composed of three parts. The first part, $CrossEntropy$, measures the difference in information between the current and the new labels. The second part, L^1, determines how close the new input should

Algorithm 2: Algorithm for a global method to choose a model's most influential dimensions.

```
$\Theta(X)$ Identifies the argument of X's second-largest value;
$G$ is a global NMG model ;
    , and $F$ shows a trained model;
$Y' \leftarrow \Theta(F(X))$;
$G \leftarrow$ train $G$ on $F$, and $(X, Y')$;
$X' \leftarrow G(X)$; /* a single value for every dimension*/
$differences \leftarrow Sum(X - X', 1)$;
$first\_important\_dimension \leftarrow max\_arg(differences)$;
$second\_important\_dimension \leftarrow \Theta(differences)$;
```

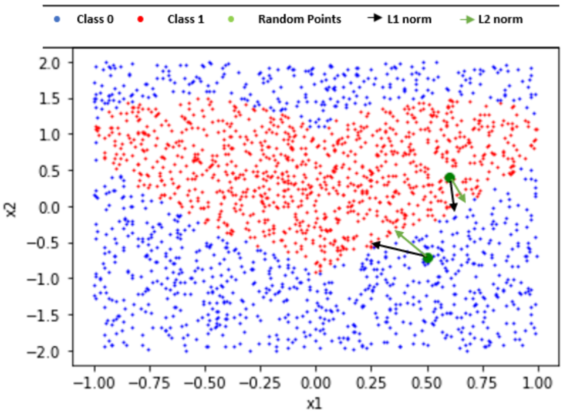


Figure 2. An example of the migration obtained with our NMG model G for some random samples on an implicit 2D heart function [12] is shown in the figure. Blue dots represent class 0 samples, while red dots represent class 1. The green dots are chosen as random samples that are transformed with our NM Generator. The results show that the NMG model successfully migrates samples to the decision boundary.

be to the original input. We use the Manhattan distance instead of the Euclidean distance to minimize the number of features that are changed. Although the L^2 distance can also be used, the L^1 norm has the advantage of producing outputs with a small number of features with significant changes. On the other hand, the L^2 norm tends to distribute the changes over more features, which is not ideal if we want to identify the most important features only.

The last component of the loss function is the $\beta_{y,y'}$ parameter, which is calculated as follows:

$$\beta_{y,y'} = |CrossEntropy(1, y == y') - \epsilon|,$$
$$\text{With } \epsilon = 0.34.$$

The last component of the loss function, $\beta_{y,y'}$, serves to prevent the first part of the loss function from having a value too close to zero, which would allow the second part to have a greater probability of changing. Additionally, λ is a parameter in the range of $[0, 1]$ that emphasizes the importance of being close to the input value. This value is initialized to zero in the first step, allowing the model to adjust the input as much as needed to change the label. The λ value for any x' increases gradually after each correct label, and decreases if the label is guessed wrongly. One advantage of this strategy is that parameters do not need to be adjusted manually, and the model can converge to the appropriate value of x'.

IV. EXPERIMENTAL RESULTS

We experimented with our new approach on four different datasets: a synthetic 2D heart function dataset, the Pima Indians Diabetes database, the Thyroid-disease dataset, and the IRIS dataset.

All tests were performed on an Intel 4770K processor running at 3.9 GHz, coupled with 32 GB of RAM and an Nvidia GeForce GTX 3070TI Graphics card. Tensorflow 2.0 was used as the AI framework and ran on a Linux distribution.

Each feature in all datasets was standardized to the range $[-1, 1]$, with a mean value of 0 to ensure that each dimension had an equal probability of change, as both the L^1 and L^2 norms are sensitive to the units used for features.

For training, we used an Adam optimizer with a learning rate set to 10^{-4}. Every layer in the models had a Relu activation function, except for the last layer which had a linear activation function. Our models were trained for 200 epochs, and the best model was saved. A description of the architecture of the three original models and their associated NMG models can be found in Table II.

A. *Implicit Heart Function Dataset*

To demonstrate the impact of the loss function on the decision of our local variant model, we created an artificial dataset by taking a two-dimensional slice of a heart-shaped function described in [12] (eq. 7). Any point inside the heart is labeled as one, and zero otherwise. We regard the original model as a black box once it is trained, and we will use it only to see how far our new predictions are from the decision boundary.

$$f(x_1, x_2) = (x_1^2 + x_2^2 - 1) - x_1^2 x_2^3$$
$$y = \begin{cases} 1 & \text{if } f(x_1, x_2) < 0 \\ 0 & \text{otherwise} \end{cases} \tag{7}$$

To identify the distribution of decision boundaries using our NMG, we utilized a fully connected network with three layers, each consisting of 256 neurons, and a final layer with only two neurons, as the output should have the same number of features as the input. The distribution of the function is depicted in Figure 2. We assigned two random points inside and outside the heart and drew the output position given by our NMG. The results show that the model was able to locate the decision boundary well with both L^1 and L^2 norms.

B. *Pima Indians Diabetes Database*

The Pima Indians Diabetes Database [13] includes seven continuous but sometimes incomplete features and a binary categorical feature that indicates whether or not a patient has type 2 diabetes. It contains 768 individuals who are at least 21 years old, female, and of Pima Indian heritage. Table III lists the features present in the dataset.

In this example, we apply our local NMG variant to a set of 13 random samples to see what minimal changes can be made to a sample to switch its class. Here again, our NMG migrates to the decision boundary by minimizing the loss function of an input sample. The very first sample we tested actually demonstrated that something was not right with the original model. In this case, we see that the NMG prioritizes a change in blood pressure. While diabetes patients often have high blood pressure, diabetes is mainly detected from glucose or insulin levels in the blood, which can be tested two hours after ingesting 75mg of glucose (Oral Glucose Tolerance Test - OGTT). While the NMG migrated the sample to a nearby point very similar to the original one, but still clearly in the diabetic range, the original model was inferring it as non-diabetic. A study of 12 more samples (Table I) indicated that this was not an exception (e.g. samples 1,3,6), while other samples looked more logical (e.g. sample 2 migration to a diabetic status logically increases parameters like glycemia and insulin but also BMI and blood pressure parameters, which are frequently raised for diabetic patients).

The only explanation for small, non-logical changes being able to change sample classes was that the original model was overfitting. This was confirmed later by a) noticing that the model had far more parameters than samples (Table II), and b) observing that training accuracy was 100%, but testing accuracy was only 73%, which is typical of overfitting. Therefore, in this example, our NMG was able to indicate to a domain expert that overfitting was occurring in the original model, even before details of the training were disclosed.

C. Thyroid-disease dataset

The original Thyroid-disease datasets from the UCI machine learning repository contain 2800 training instances and 972 test instances, each with three classes. Each instance has 21 attributes, consisting of 15 categorical and six continuous numbers. For the ease of displaying the experiment result, we narrowed the classes down to two and kept records based on the following criteria:

$$\begin{aligned} On\,thyroxine = Query\,on\,thyroxine\,\&\\ On\,antithyroid\,medication:\ False\,\&\\ Thyroid\,surgery:\ False\,\&\\ I131\,treatment:\ False\,\&\\ Tumor:\ False\&\\ Hypopituitary:\ False \end{aligned}$$

After filtering, there were 3065 records left, including 2298 training samples and 767 test instances. We removed most of the categories from the dataset except for three columns: 'FTI', 'TBG', and 'Referral source', and did not utilize them to train the original model F or the NMG model G. We employed the same setup as in our experiment in section IV-A (Local variant with L^1 norm) and normalized the dataset. However, we altered the weight vector ω to make some input dimensions less variable. The dimensions and weights connected with them are shown in table V.

Feature Name	Weight	Description
Age	99	Continuous
Sex	99	Categorical
TSH	1	Continuous
T3	1	Continuous
TT4	1	Continuous
T4U	1	Continuous
FTI	99	Continuous
TBG	99	Continuous
Referral source	99	Categorical

Table V

DESCRIPTION OF FEATURES IN THE THYROID-DISEASE DATASETS.

Original	Migrated	Sepal L	Sepal W	Petal L	Petal W
Virginica	Versicolor	0.0662	0.0207	**0.5398**	0.0965
Versicolor	Virginica	0.0340	0.0081	**0.3372**	0.0033
Setosa	Versicolor	**0.3106**	**0.4010**	**0.9289**	**0.7361**

Table VI

THE TABLE DISPLAYS THE VARIATION AMONG FEATURES WHEN CHANGING INSTANCES. THE LEFT TWO COLUMNS SHOW THE ORIGINAL CLASS AND THE CLOSEST CLASS PREDICTED BY THE GLOBAL NMG MODEL. THE RESULTS ALIGN WITH WHAT CAN BE OBSERVED IN FIG. 3. THE RIGHT-HAND SIDE PRESENTS THE IMPACT VALUES RETURNED BY THE NMG MODEL FOR EACH FEATURE. FOR THE FIRST TWO MIGRATIONS, THE RESULTS INDICATE THAT PETAL LENGTH IS THE ONLY SIGNIFICANT FEATURE IN CLASS MIGRATION. FOR THE SETOSA CLASS, ALL PARAMETERS HAVE SOME INFLUENCE, BUT PETAL LENGTH AND WIDTH REMAIN THE MOST IMPORTANT FACTORS.

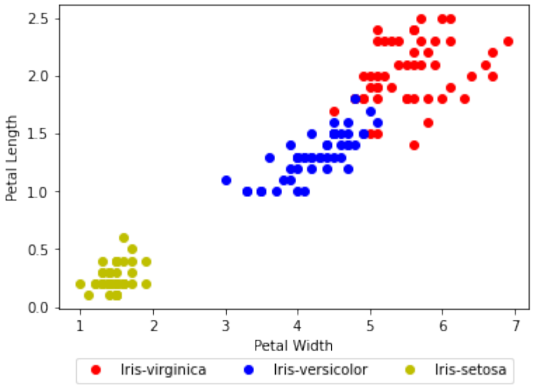
According to the results, the ω vector prevents features from changing by adding the cost of modifying them in the loss function. A detailed examination of the findings reveals that two input features influence migrants from one class to another. The final results are listed in table IV.

D. IRIS

The IRIS dataset [14] is well-known for its taxonomic difficulties in incorporating many measures. The original dataset contains 50 samples from each of the three Iris species ($Iris, Setosa, Iris, Virginica$, and $Iris, Versicolor$), resulting in a total of 150 items. Each sample has four attributes measured in centimeters, including the length and width of both the sepals and petals. Fig. 3 shows that two of the four features are sufficient to provide a good separation of the three classes. This figure also provides insight into the neighboring relationships between the classes. We apply the global variant of our Neighbor Migrating Generator to this dataset to determine if this neighboring relationship can be captured by our technique. With this global approach, the model learns how to go to the nearest border rather than specifying which class border it should learn.

Our global NMG model is defined by three fully connected layers of 1024 neurons with Relu activation. New labels in the training set are defined by the second best predicted label of

Figure 3. 2D plotting the IRIS dataset. The graph indicates that using only two features ($Petal\,Length$ and $Petal\,Width$) out of four is enough to separate most samples of this dataset. Note that $Iris\,Versicolor$ appears to be in between the two other classes.



the original model. Results in Table VI show that the migration of the three different classes corresponds pretty much to what can be observed in figure 3, with for instance the $Setosa$ and $Virginica$ classes migrating the $Versicolor$ class and the latter one being migrated to the closest of the two, i.e., $Virginica$.

V. Conclusion

In this work, we introduced the Neighbor Migrating Generator (NMG), which is a method for finding the closest neighbor with a different class label. In other words, the NMG searches for decision boundaries that allow us to change a given input with the least amount of effort. Our NMG can be used to determine the smallest modifications and the influential dimensions in the decision space that distinguish the classes. In three experiments, we demonstrated the capabilities of our generator to: a) find decision boundaries (e.g., heart function) in the original model; b) detect overfitting in results (e.g., Pima Indians Diabetes Database); and c) provide close neighboring clues between classes (Iris dataset and Thyroid-disease dataset). Another benefit is that, unlike previous approaches [4], our method can learn the boundary without having to change any kernel parameters, resulting in more reliable outcomes.

Future work could extend this technique to the global neighboring algorithm variant and establish a visualization tool of the decision boundaries that could help understand the impact features have on the decision made by the original model.

References

[1] Marcin Kapcia, Hassan Eshkiki, Jamie Duell, Xiuyi Fan, Shangming Zhou, and Benjamin Mora. Exmed: An ai tool for experimenting explainable ai techniques on medical data analytics. In *2021 IEEE 33rd International Conference on Tools with Artificial Intelligence (ICTAI)*, pages 841–845, 2021.

[2] Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H Witten. The weka data mining software: an update. *ACM SIGKDD explorations newsletter*, 11(1):10–18, 2009.

[3] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[4] Sandra Wachter, Brent D. Mittelstadt, and Chris Russell. Counterfactual explanations without opening the black box: Automated decisions and the GDPR. *CoRR*, abs/1711.00399, 2017.

[5] Alex Goldstein, Adam Kapelner, Justin Bleich, and Emil Pitkin. Peeking inside the black box: Visualizing statistical learning with plots of individual conditional expectation, 2014.

[6] Susanne Dandl, Christoph Molnar, Martin Binder, and Bernd Bischl. Multi-objective counterfactual explanations. *Lecture Notes in Computer Science*, page 448–469, 2020.

[7] Marco Túlio Ribeiro, Sameer Singh, and Carlos Guestrin. ”why should I trust you?”: Explaining the predictions of any classifier. *CoRR*, abs/1602.04938, 2016.

[8] Scott Lundberg and Su-In Lee. A unified approach to interpreting model predictions, 2017.

[9] Ambreen Hanif, Xiuyun Zhang, and Steven Wood. A survey on explainable artificial intelligence techniques and challenges. In *2021 IEEE 25th International Enterprise Distributed Object Computing Workshop (EDOCW)*, pages 81–89, 2021.

[10] Thibault Laugel, Xavier Renard, Marie-Jeanne Lesot, Christophe Marsala, and Marcin Detyniecki. Defining locality for surrogates in post-hoc interpretablity. *CoRR*, abs/1806.07498, 2018.

[11] Riccardo Guidotti, Anna Monreale, Salvatore Ruggieri, Dino Pedreschi, Franco Turini, and Fosca Giannotti. Local rule-based explanations of black box decision systems, 2018.

[12] Eric W. Weisstein. Heart curve – from wolfram MathWorld.

[13] Jack Smith, J. Everhart, W. Dickson, W. Knowler, and Richard Johannes. Using the adap learning algorithm to forcast the onset of diabetes mellitus. *Proceedings - Annual Symposium on Computer Applications in Medical Care*, 10, 11 1988.

[14] R. A. Fisher. The use of multiple measurements in taxonomic problems. *Annals of Eugenics*, 7(2):179–188, 1936.

Raising User Awareness of Bias-Leakage via Proxies in AI Models to Improve Fairness in Decision-making

Peter Daish
dept. Computer Science
Computational Foundry
Swansea University
Swansea, Wales
SA1 8EN
peter.daish@swansea.ac.uk

Matt Roach
dept. Computer Science
Computational Foundry
Swansea University
Swansea, Wales
SA1 8EN
m.j.roach@swansea.ac.uk

Alan Dix
dept. Computer Science
Computational Foundry
Swansea University
Swansea, Wales
SA1 8EN
a.j.dix@swansea.ac.uk

***Abstract*—Artificial Intelligence systems are becoming more common in decision-making, both for facilitating automated decisions or in tandem with human decision-makers as decision-support systems. AI-assisted DSS are typically employed to make data-driven recommendations to human decision-makers in an effort to improve efficiency and accuracy. In addition, the AI used to power DSS are typically blackbox in nature, meaning that human decision-makers are unaware of exactly how these systems are coming to their conclusions. This is problematic since research in algorithmic fairness already shows that data-driven AI systems can be influenced by social biases present in training data, to reinforce systemic biases and perpetuate unfairness towards minority social groups. When used in high-stakes decision-making, such systems risk protracting systemic biases and further driving social division. An area of research is emerging acknowledging that unfairness can leak through 'proxy' features, causing an implicit-bias effect. In this work-in-progress paper, we propose explaining fairness properties of AI systems and their downstream social impacts to decision-makers - by visualising bias-leakage through proxies - for improved fairness. Finally, we are currently in the process of conducting a study to empirically assess how visualising proxy-biases in AI-assisted DSS can affect decision-making and improve fairness.**

***Index Terms*—machine learning, fairness, proxy, bias-leakage, decision support system, downstream unfairness, machine-bias, bias-visualisation**

I. INTRODUCTION

Automated decision-making (ADM) and decision-support systems (DSS) have been hot-topics of research for many years. Originating from a desire to improve the efficiency and effectiveness of the decision-making process [1]–[4], many examples of decision-systems have already been deployed in society, with use-cases spanning from recreational through to safety-critical. Examples of such systems include artificial intelligence (AI) powered decision-making in: complex board games [5], video-games [6], [7], self-driving cars [8] and clinical settings [9].

ADMs and DSS make use of AI and Machine Learning (ML) technologies to automate decision-making and provide automated recommendations to human decision-makers, respectively. These systems, as a result of their 'autonomous' nature, often leave little room for human intervention or supervision and even when they do, it is mostly in the capacity of acknowledging the machine's result rather than understanding what led to the decision being made [10]. Furthermore, many of these systems make use of 'blackbox algorithms' to make decisions: blackbox algorithms are of a type of algorithm whereby the users and developers often have little-to-no knowledge as to how the system is actually making decisions. This can occur either because the system is complex, or because the code used to develop the algorithm is a trade secret [11]. As such, these systems are labelled 'opaque' and stand in contrast to whitebox systems which provide transparency by design. The use of a system whose mechanics eludes the users, can be problematic: without knowing what causes the system to make its decision, it may perform unexpectedly [12] and it may also be difficult to assess its fairness [13]. Such a system may well perpetuate and reinforce systemic social biases unbeknownst to its stakeholders [14], [15].

Adding to concerns that deployed systems are perpetuating unfairness, there are worries that legal systems do not provide sufficient protections for citizens who are recipients of automated decisions or profiling. Wachter (et al., 2016) [16] stresses that the language in the European Union's (EU) General Data Protection Regulation (GDPR) - drafted to protect EU citizens' digital rights - is not sufficient to guarantee citizens a "right to explanation" for automated decisions - explanations are expected, but do not need to cover all the factors that led to the decision being made. Meanwhile, Kaltheuner and Bietti (2018) [17] state that despite this, citizens ought at least be afforded the right to object to receiving an automated decision, given their interpretation of EU GDPR Article 22, titled: "Automated individual decision-making, including profiling" [18]. Wachter's argument pro-

---

The first author of this paper is funded through a 1+3 UKRI ESRC PhD scholarship.

poses a challenge in itself, in that generating useful and easily interpretible explanations of blackbox systems is non-trivial [19].

The lack of transparency inherent in blackbox algorithms' designs have given rise to a wealth of research dedicated to exploring the fairness properties of ML systems and their robustness when deployed in society. For example, systems trained on historical data have been discovered to perpetuate gender biases [20], perform in a racially biased manner [21], [22] and for unfairness to persist through proxies [23].

Proxies are features which exhibit a relation to protected features (i.e., 'race', 'gender', 'religion', etc.) and are responsible for the Fairness Through Unawareness phenomenon, whereby models are found to discriminate according to designated protected features, even when they have been omitted from the training data of the model [24]. The existence of proxies in blackbox systems perpetuating social biases were by no means unforeseen, with references to their influence on decision-making recorded early on in Alan Dix's (1992) paper on *Human Issues in the use of Pattern Recognition Techniques*.

In light of unfairness persevering through "unawareness", AI systems may opt to include protected features in their training data, but implement fairness constraints (FC) [25] which 'enforce' computational fairness definitions such as demographic parity [26], equality of opportunity or equality of odds [27]. Whilst these methods ensure that similar groups of people are treated similarly, they are not infallible and can result in poor performance when protected features and non-protected features are highly related [28]. Poor, but fair, algorithmic performance presents an issue if used in decision-support contexts, in that it could result in individuals receiving recommendations arbitrarily, as long as fairness is met eventually.

Rather than sacrifice performance for the sake of fairness, we propose for human-decision makers to re-gain control and responsibility of high-stakes decision-making, by providing them a means of understanding the fairness properties of the underlying AI systems used to support their decisions. In so doing, decision-makers will become aware of potential biases being propagated through the system and can take them into consideration when deliberating over the correct and fair decision.

Throughout the rest of this paper, we refer to the phenomenon by which proxies exist as 'bias-leakage' since it succinctly refers to the predictive power that non-protected features exhibit, that is related to their protected feature counterparts. An example of this is the non-protected feature "geolocation" acting as a proxy for the protected feature "race" [24]. In such a circumstance, we can determine that race exhibits 'bias-leakage' through to geolocation, which is acting as its proxy.

The relationships of features involved in bias-leakage might be linear, or might have complex non-linear relationships. We can think of bias-leakage as being similar in nature to human 'implicit' or 'unconscious' biases and when dealing with DSS in this paper, we refer to these types of features performing as 'proxy-biases' for protected features, to make the distinction between explicit and implicit machine-biases clear.

II. PLANNED RESEARCH AND RESEARCH QUESTIONS

Whilst various methods for bias visualisation have been proposed [29], [30], they often require expert domain knowledge and substantial training to use. As a result, they are not best suited for use in DSS, unless the end-users are technical domain experts.

An area of literature that has gained little momentum - but is now becoming ever more relevant - is in the field of developing understanding of AI fairness properties by non-technical domain expert decision-makers.

In our full paper, we will propose a technique for fairness-explained DSS, which will be empirically evaluated in a lab study designed to assess its impact on human decision-making. Within the study, research participants will preside over a number of parole sentencing cases in the following three interventions: no assistance, assistance provided by an AI-assisted DSS and finally, fairness-explained assistance. The study conforms to a within-groups design as a means of providing participants lived experience with all interventions, to ascertain how each has impacted their decision-making.

In addition to an empirical evaluation of a novel fairness-explanation technique we will contribute the algorithmic approach to generating fairness explanations. Furthermore, we shall contribute a novel dataset of parole cases derived from the COMPAS dataset [15], realistically matched with FairFace images based on demographic data [31]. To the best of our knowledge, this study will be the first of its kind to provide empirical evidence on how human understanding of machine-bias at decision time can affect the decisions they make.

Our research aims to answer the following questions, by way of a deployed lab-study and follow-up interviews:

> ***RQ 1****: How do proxy-bias-leakage visualisations, impact decision-making?*
>
> ***RQ 2****: What is the perceived usefulness of visualising proxy-bias-leakage to non-experts when making high-stakes decisions?*
>
> ***RQ 3****: How does visualising proxy-bias-leakage affect Human decision-maker confidence in combined decision-making?*
>
> ***RQ 4****: How does visualising proxy-bias-leakage affect Human decision-maker automation bias?*

By answering these questions, we hope to provide conclusive evidence for future work looking into making proxy-bias-leakage known to human decision-makers when using AI-assisted DSS to improve decision-making.

III. RESEARCH OUTCOMES AND SOCIAL IMPACT

The successful outcome of this project will contribute to current fairness and explainable AI literature by opening an avenue for research into user-centered fairness-explained AI systems. We plan to achieve this by empirically evidencing how decision-maker understanding of bias-leakage through proxies affects the decision-making process.

References

[1] S. Liu, A. H. B. Duffy, R. I. Whitfield, and I. M. Boyle, "Integration of decision support systems to improve decision support performance," *Knowledge and Information Systems*, vol. 22, no. 3, pp. 261–286, Mar. 2010. [Online]. Available: http://link.springer.com/10.1007/s10115-009-0192-4

[2] R. H. Bonczek, C. W. Holsapple, and A. B. Whinston, *Foundations of decision support systems*, ser. Operations research and industrial engineering. New York: Academic Press, 1981.

[3] U. R. Averweg, "Historical Overview of Decision Support Systems (DSS):," in *Encyclopedia of Information Science and Technology, Second Edition*, M. Khosrow-Pour, D.B.A., Ed. IGI Global, 2009, pp. 1753–1758. [Online]. Available: http://services.igi-global.com/resolvedoi/resolve.aspx?doi=10.4018/978-1-60566-026-4.ch275

[4] A. E. Waldman, "Power, Process, and Automated Decision-Making," *Fordham Law Review*, vol. 88, p. 613, 2019. [Online]. Available: http://fordhamlawreview.org/wp-content/uploads/2019/11/Waldman_November_S_9.pdf

[5] P. Bory, "Deep new: The shifting narratives of artificial intelligence from Deep Blue to AlphaGo," *Convergence: The International Journal of Research into New Media Technologies*, vol. 25, no. 4, pp. 627–642, Aug. 2019. [Online]. Available: http://journals.sagepub.com/doi/10.1177/1354856519829679

[6] S. Risi and M. Preuss, "Behind DeepMind's AlphaStar AI that Reached Grandmaster Level in StarCraft II: Interview with Tom Schaul, Google DeepMind," *KI - Künstliche Intelligenz*, vol. 34, no. 1, pp. 85–86, Mar. 2020. [Online]. Available: http://link.springer.com/10.1007/s13218-020-00642-1

[7] O. Vinyals, I. Babuschkin, J. Chung, M. Mathieu, M. Jaderberg, W. Czarnecki, A. Dudzik, A. Huang, P. Georgiev, R. Powell, T. Ewalds, D. Horgan, M. Kroiss, I. Danihelka, J. Agapiou, J. Oh, V. Dalibard, D. Choi, L. Sifre, Y. Sulsky, S. Vezhnevets, J. Molloy, T. Cai, D. Budden, T. Paine, C. Gulcehre, Z. Wang, T. Pfaff, T. Pohlen, D. Yogatama, J. Cohen, K. McKinney, O. Smith, T. Schaul, T. Lillicrap, C. Apps, K. Kavukcuoglu, D. Hassabis, and D. Silver, "AlphaStar: Mastering the Real-Time Strategy Game StarCraft II," 2019. [Online]. Available: https://deepmind.com/blog/alphastar-mastering-real-time-strategy-game-starcraft-ii/

[8] C. Badue, R. Guidolini, R. V. Carneiro, P. Azevedo, V. B. Cardoso, A. Forechi, L. Jesus, R. Berriel, T. M. Paixão, F. Mutz, L. de Paula Veronese, T. Oliveira-Santos, and A. F. De Souza, "Self-driving cars: A survey," *Expert Systems with Applications*, vol. 165, p. 113816, Mar. 2021. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S095741742030628X

[9] R. T. Sutton, D. Pincock, D. C. Baumgart, D. C. Sadowski, R. N. Fedorak, and K. I. Kroeker, "An overview of clinical decision support systems: benefits, risks, and strategies for success," *npj Digital Medicine*, vol. 3, no. 1, p. 17, Feb. 2020. [Online]. Available: https://www.nature.com/articles/s41746-020-0221-y

[10] B. Wagner, "Liable, but Not in Control? Ensuring Meaningful Human Agency in Automated Decision-Making Systems: Human Agency in Decision-Making Systems," *Policy & Internet*, vol. 11, no. 1, pp. 104–122, Mar. 2019. [Online]. Available: https://onlinelibrary.wiley.com/doi/10.1002/poi3.198

[11] M. Perel (Filmar) and N. Elkin-Koren, "BLACK BOX TINKERING: Beyond Transparency in Algorithmic Enforcement," *SSRN Electronic Journal*, 2016. [Online]. Available: http://www.ssrn.com/abstract=2741513

[12] W. B. *, J. R. *, and M. Bethge, "Decision-based adversarial attacks: Reliable attacks against black-box machine learning models," in *International Conference on Learning Representations*, 2018. [Online]. Available: https://openreview.net/forum?id=SyZI0GWCZ

[13] M. Marabelli, S. Newell, and V. Handunge, "The lifecycle of algorithmic decision-making systems: Organizational choices and ethical challenges," *The Journal of Strategic Information Systems*, vol. 30, no. 3, p. 101683, Sep. 2021. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0963868721000305

[14] A. Chander, "The Racist Algorithm?" *Michigan Law Review*, no. 115.6, p. 1023, 2017. [Online]. Available: https://repository.law.umich.edu/mlr/vol115/iss6/13/

[15] J. Angwin, J. Larson, S. Mattu, L. Kirchner, and ProPublica, "Machine Bias There's software used across the country to predict future criminals. And it's biased against blacks." May 2016. [Online]. Available: https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

[16] S. Wachter, B. Mittelstadt, and L. Floridi, "Why a Right to Explanation of Automated Decision-Making Does Not Exist in the General Data Protection Regulation," *SSRN Electronic Journal*, 2016. [Online]. Available: https://www.ssrn.com/abstract=2903469

[17] F. Kaltheuner and E. Bietti, "Data is power: Towards additional guidance on profiling and automated decision-making in the GDPR," *Journal of Information Rights, Policy and Practice*, vol. 2, no. 2, Mar. 2018. [Online]. Available: https://jirpp.winchesteruniversitypress.org/article/10.21039/irpandp.v2i2.45/

[18] T. E. P. A. COUNCIL, "Art. 22 GDPR Automated individual decision-making, including profiling," p. 46, Apr. 2016. [Online]. Available: https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32016R0679

[19] A. Adadi and M. Berrada, "Peeking inside the black-box: A survey on explainable artificial intelligence (xai)," *IEEE Access*, vol. 6, pp. 52 138–52 160, 2018.

[20] D. K. Citron and F. Pasquale, "The scored society: Due process for automated predictions," *Washington Law Review*, vol. 89, no. 1, p. 1 – 33, 2014, cited by: 359. [Online]. Available: https://ssrn.com/abstract=2376209

[21] T. Simonite, "When it comes to gorillas, google photos remains blind," Jan 2018. [Online]. Available: https://www.wired.com/story/when-it-comes-to-gorillas-google-photos-remains-blind/

[22] A. Rose, "Are face-detection cameras racist?" Jan 2010. [Online]. Available: https://content.time.com/time/business/article/0,8599,1954643,00.html

[23] B. E. Harcourt, "Risk as a Proxy for Race," *Federal Sentencing Reporter*, vol. 27, no. 4, pp. 237–243, Apr. 2015. [Online]. Available: https://online.ucpress.edu/fsr/article/27/4/237/43360/Risk-as-a-Proxy-for-RaceThe-Dangers-of-Risk

[24] A. Castelnovo, R. Crupi, G. Greco, D. Regoli, I. G. Penco, and A. C. Cosentini, "A clarification of the nuances in the fairness metrics landscape," *Scientific Reports*, vol. 12, no. 1, p. 4209, Mar. 2022. [Online]. Available: https://www.nature.com/articles/s41598-022-07939-1

[25] M. B. Zafar, I. Valera, M. G. Rodriguez, and K. P. Gummadi, "Fairness constraints: Mechanisms for fair classification," 2015. [Online]. Available: https://arxiv.org/abs/1507.05259

[26] R. Zemel, Y. Wu, K. Swersky, T. Pitassi, and C. Dwork, "Learning fair representations," in *Proceedings of the 30th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, S. Dasgupta and D. McAllester, Eds., vol. 28, no. 3. Atlanta, Georgia, USA: PMLR, 17–19 Jun 2013, pp. 325–333. [Online]. Available: https://proceedings.mlr.press/v28/zemel13.html

[27] M. Hardt, E. Price, and N. Srebro, "Equality of opportunity in supervised learning," *CoRR*, vol. abs/1610.02413, 2016. [Online]. Available: http://arxiv.org/abs/1610.02413

[28] M. B. Zafar, I. Valera, M. Gomez-Rodriguez, and K. P. Gummadi, "Fairness constraints: A flexible approach for fair classification," *J. Mach. Learn. Res.*, vol. 20, no. 1, p. 2737–2778, jan 2019.

[29] R. K. E. Bellamy, K. Dey, M. Hind, S. C. Hoffman, S. Houde, K. Kannan, P. Lohia, J. Martino, S. Mehta, A. Mojsilovic, S. Nagar, K. N. Ramamurthy, J. Richards, D. Saha, P. Sattigeri, M. Singh, K. R. Varshney, and Y. Zhang, "AI Fairness 360: An extensible toolkit for detecting, understanding, and mitigating unwanted algorithmic bias," Oct. 2018. [Online]. Available: https://arxiv.org/abs/1810.01943

[30] M. Hardt, X. Chen, X. Cheng, M. Donini, J. Gelman, S. Gollaprolu, J. He, P. Larroy, X. Liu, N. McCarthy, A. Rathi, S. Rees, A. Siva, E. Tsa, K. Vasist, P. Yilmaz, B. Zafar, S. Das, K. Haas, T. Hill, and K. Kenthapadi, "Amazon sagemaker clarify: Machine learning bias detection and explainability in the cloud," in *KDD 2021*, 2021. [Online]. Available: https://www.amazon.science/publications/amazon-sagemaker-clarify-machine-learning-bias-detection-and-explainability-in-the-cloud

[31] K. Karkkainen and J. Joo, "Fairface: Face attribute dataset for balanced race, gender, and age for bias measurement and mitigation," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2021, pp. 1548–1558.

A position on establishing effective explanations from human-centred counterfactuals for automated financial decisions

Alex Blandin
Swansea University
Bay Campus, UK
`a.j.blandin@swansea.ac.uk`

Matt Roach
Swansea University
Bay Campus, UK
`m.j.roach@swansea.ac.uk`

Daniele Doneddu
Swansea University
Bay Campus, UK
`d.doneddu@swansea.ac.uk`

Jen Pearson
Swansea University
Bay Campus, UK
`j.pearson@swansea.ac.uk`

Matt Jones
Swansea University
Bay Campus, UK
`matt.jones@swansea.ac.uk`

David Sullivan
Starling Bank
Cardiff, UK
`david.sullivan@starlingbank .com`

***Abstract*—There has been a historic focus among explainable artificial intelligence practitioners to increase user trust through the provision of explanation traces from algorithmic decisions, either by interpretable agents and expert systems or from transparent surrogate models. Beyond this deterministic causal reasoning, significant developments were made in probabilistic post-hoc explanations, often presented as percentage confidence or importance and contribution. However, simultaneous work in the social sciences revealed that typical users and experts both preferred and generated explanations that differed in conception with those often employed on artificial intelligence, with the existing deterministic traces and probabilistic reasoning being found less satisfactory, particularly when the best explanations for a given user were not the most likely.**

In this piece, we hold the position that incorporating an understanding of explanations as a model and process — inspired by social science research in human-centred explanations — will improve user satisfaction and trust in both the given decision and the overall model. We consider how practitioners may design explainable artificial intelligence that enables typical users to interface and interact with counterfactual explanations and therein develop more appropriate explanations for that specific user and decision. Specifically, we argue in favour of satisfying design desiderata for explanations that are causal, contrastive, contextual, and interactively selected with the user. This piece is based on ongoing work that demonstrates the practicability of interactive human-centred explanatory models and is inspired by previous work to uncover design characteristics for counterfactual explanations that enhance user trust and understanding in algorithmic decisions for automated financial decisions.

***Index Terms*—Explainable AI, Human-centred XAI, Human Computer Interaction, Counterfactual Explanations**

I. INTRODUCTION

The arguments presented here draw upon learnings from the social sciences, as highlighted by Miller [1], our own ongoing work, and the work and critique of others in the Explainable Artificial Intelligence (XAI) field. Commenting on the current state of XAI as a field, our techniques are still driven by either statistical arguments of correctness, such as the accuracy of a surrogate model, or by the intuition of the researcher regarding what constitutes good explanations.

We observe, across the field, an absence in the appreciation for the interaction problem between the human and the explanation system. Human-computer interaction (HCI) must become a central pillar of XAI, as we envisage XAI to be a tool that promotes trust in our human end users. Enshrining this will enable us to explore further into the explanatory space between the social sciences and artificial intelligence, becoming not simply HCI, but HHCI or HCHCI, exploring end users communicating with, say, an expert that is augmented by a computer, and end users with computers doing so.

Research in this space is ongoing, albeit currently outside the focus of many AI and XAI practitioners, under the domain of human-centred XAI (HCXAI). This includes new models for explanation being identified and validated in the social sciences, new ways to extract information from local and global behaviour in AI, and new ways to bridge these for emergent explanatory systems that further our goals of increased user trust and understanding. We still have much to learn from ourselves, such as in how we navigate around the computational infeasibility of explanation, and do so in a way that is appropriate and correct, and where we fail in this.

With this perspective, we identify that explanations, in order to improve understanding, must perform a critical task: identify and act towards a middle ground of understanding between the user and itself. In doing so, it will strive to bring the user closer towards that point, and thus closer to its own understanding. To achieve this, we argue that it will necessarily tailor its explanations to the user interactively, which entails from models of explanation found in the social sciences.

Alex Blandin is funded by the EPSRC Centre For Doctoral Training In Enhancing Human Interactions And Collaborations With Data And Intelligence Driven Systems (EP/S021892/1). We thank David Sullivan and his employer, Starling Bank, for sponsoring and supporting the project through this CDT.

This leaves open the critical question of how to build such an ongoing conversation with the user. We find answers to this are lacking in many areas of XAI at current, where the explanations often deployed are closed responses, intended for machine learning practitioners — whom make up the majority of practical XAI usage [2]. Where we do see user interaction, it is often solely with interface primitives, such as modifying a field in a dashboard, or altering axes on visualisations. This is not a conversation where a user can say "let's focus in on this feature" or ask "what's causing this?", and is therefore not an environment in which users can deepen their understanding as they would in interaction with other humans — as derived from the social sciences. Capitalising upon this has not been the main focus for many researchers. We feel this is a problem as, in one simplified phrasing, it is losing sight of a significant human element that remains at the heart of the matter.

Failing to engage with this element and develop human-centred XAI has lead to explanations that were hard to understand or unintentionally misleading [3], such as waterfall plots where users failed to understand causally how the contributions actually added up, ignoring the impact of negative contributions. Users had difficulty in mentally summing the effect of many small contributions, which was necessary in this explanation as it lacked a stacked presentation, such as a force plot [4]. This meant that contributions were not something users could easily read and understand the causal effects of, as that kind of visual processing is not a strength for many people, as recognised in HCI and visualisation [5].

This is a topical area, and our hope is that the increasing interest in XAI and human-centred explanations from both academia and industry is evidence of an inflection point that serves to normalise it. At the very least, we hope that it is normalised enough that users and practitioners understand and recognise its utility, even if it is not a universal feature.

Our position is specifically related to the interactive selection of explanations with counterfactual contrast, drawing from the social sciences, where "why was it P instead of Q" highlights the need for counterfactuals. This has received some traction in XAI, however other approaches garner the majority of human-centred exploration, which we identify as not yet fully exploring the transdisciplinary interaction of XAI with the social sciences.

Our position is that these interactive counterfactuals are not only effective explanations, but are also practicable by following principles that have been empirically validated by social sciences, HCI, and artificial intelligence research.

We argue that the principle challenge of XAI is understanding the given problem that an XAI is intended to solve. This seems to be an echo of the arguments put forward by thought-leaders in the field, such as Rudin [6]. We extend this to the specific area of explanations, where deeper understanding and harmonisation of both the problem domain and the explanation process — via explanatory models — gives rise to an approach that is practicable and can be realistically implemented to significant end user benefit. This extension supports the goals of HCXAI, providing grounding for XAI research.

Our ongoing work is a reflection of this position, currently in its early iterations, however draws from our previous efforts to co-design human-centred explanations and XAI. By empirically demonstrating the effectiveness and practicability of interactive explanatory models and human-centred perspectives, we should reveal the design desiderata of effective counterfactual explanations, and the viability of developing user understanding and trust in XAI with these techniques.

II. BACKGROUND

We feel it is appropriate to consider relevant and relative positions of other works in the field of AI and XAI. This serves to illuminate both the existing approaches and also the gaps that inspire our ongoing work. It is apparent that increasing the awareness of these developments among AI and XAI practitioners is of critical importance, particularly given the fulcrum that researchers and engineers occupy in what is, by all accounts, an accelerating arms-race between corporate entities, labs, and potentially even nation-state interests.

As epitomised by recent news [7], the industrial state-of-the-art continues to distance itself from the (performative) auspices of ethics and safety that were once held, all seemingly shed in anticipation of an escalation in AI technology that may bear highly lucrative new capabilities ripe for external investment and commercial opportunity. If we are to avoid overpromising and under-delivering as a field, and just as importantly avoid malpractice and preventable harm, we must re-establish the bounds of the research and deployment of AI as a human-centred effort that must be properly motivated and evaluated to ensure the just development and usage of this technology.

A consequence of this revitalised rush to develop and deploy AI has been a widespread misattribution of pattern recognition for reasoning, particularly with large language models (LLMs) and other generative AI technologies. Recent research has demonstrated that this often comes from improper or ungrounded experimental practices, which when corrected for provides evidence that these recent AI are far less capable than promised or as apparent from casual inspection [8, 9]. One experimental shortcoming that was ironically highly publicised is the performance on standardised human tests and benchmarks, which wrongly presupposes that the performance on these questions is meaningful and relevant to the capabilities of the AI; standardised tests rely on humans not knowing the contents of the test and oft interpret memorisation as a hallmark of "cheating". It is now publicly recognised [10] that LLMs repeatedly leak their answers from their training data, whilst failing (no better than random) on genuinely novel or modified questions, which implies those publicised results were instead evidence for pattern recognition (granted, how truly remarkable it is). It can be argued that LLMs should instead be evaluated as "non-human subjects", to borrow the term from cognitive psychology, for which there are known experimental methods [11, 12] that could be used to demonstrate LLMs capabilities. Following this line of enquiry, it is found that LLMs are not capable of non-trivial planning [13] nor capable of effective theory of mind [8, 9], indicating that initial reporting on the capacity for generalised understanding, reasoning, and planning in LLMs may be misrepresentative [14, 15].

Unfortunately, the "broad but shallow linguistic competence exhibited by LLMs" [16] can still prove harmful [17, 18], and their deployment may be unethical even with the current methods employed to make them "more polite". Indeed, this stretches beyond LLMs, as shown by Drage et al. [19, 20], touching on the failing to address the systemic injustices and predisposition to a status quo that perpetuates them. This shows how there are numerous AI systems being developed or deployed that fail to appropriately consider their impact on marginalised groups and communities and mitigate or avoid harms caused and propagated. With current and incoming legislation (GDPR [21], EU Guidelines on AI [22], and the EU's proposed "AI Act" [23]), it is emphatically imperative that progress be made on the problem of ethically sound AI. With the regulatory requirements of the GDPR, particularly in high-stakes scenarios like as recognised by the proposed AI Act, it is clear that Explainable AI and Interpretable Machine Learning (IML) are central in this as both a pragmatic and philosophical solution. This does not mean we necessarily lose the benefits of deep neural networks and our current state-of-the-art, merely that they should be made interpretable with appropriate modifications such that their decisions and behaviour can be understood [24].

To ensure soundness and correctness, it behoves us to tackle the epistemological challenges therein to the fullest extent of our collective capacity and wisdom as a field. Fortunately, there are bodies of research engaged with exactly this, and a growing transdisciplinary interest from those engaged in human-centred AI research. This includes applied examples of intersectional research on racism, sexism, and ableism in AI [19, 20, 25]. Helping resolve these systemic issues, rather than perpetuating them, should be upheld as a clear motivator for ethical, human-centred AI, which has demonstrably required transdisciplinary interaction with philosophy, HCI, and the social sciences, including gender studies, queer theory, disability justice, universal/accessible design, critical race theory, caste scholarship, etc. — efforts without this may entrench the very biases and behaviours they nominally mitigate [26].

Having motivated important areas in the wider AI landscape adjacent to our position, we wish to reaffirm that the goal in XAI is to create systems that are trustworthy via understandability. This awareness highlights the capability of holistic solutions to address the plenary of stakeholder needs, and signposts where the appropriate pivots are for research to deliver these advancements. We feel that the overarching lesson that is established by the background so far, that of fruitful interaction with intersecting disciplines to develop and mature the field of XAI, can be used to illuminate particular shortcomings and bottlenecks in our research and industrial efforts whilst simultaneously providing avenues to overcome them; much like mentors might provide. In this vein, we can now explore the specific terms of what we advocate for in this positional paper, with inspiration from Miller's survey [1].

Contextual Explanation

One of the truths that has been long-established by XAI research is that there are usually far more possible explanations than could reasonably be considered by a user. This motivated many technical approaches to evaluate and select the presented explanation based on statistical arguments. Unfortunately, the most likely explanation is not always the best, and this is approached in the social sciences by considering the multiplicity of available perspectives and selecting therein according to the given context. This is the product of separate internal and external frames of reference in a system for each agent involved (similar to the Lagrangian and Eulerian perspectives of particle motion in physics).

Explanatory models can encode relevant perspectives into a framework with layers of explanation, such as the Aristotelian "four causes", which identifies archetypical explanations in the material (physical phenomena or computation/data), formal (patterns, appearance, shape, categories), mechanistic (agents acting upon the object or data), and the functional (purpose, goal, cause). Other frameworks may give different explanatory layers that are relevant to user perspectives, such as for causal models [27, 28] and structural models [29, 30].

To demonstrate, we give explanations for why "someone was late" that use different perspectives of the same events: "because of a rail fault at Borough station" (material explanation), "because the Northern Line terminated early at London Bridge" (formal explanation), "because of increased pedestrian traffic across London Bridge" (mechanistic explanation), and "because they slept through their morning alarm, which is why this delay affected them" (functional explanation). Each layer of explanation is valid, so determining which to use is a key factor in the explanatory model. An explanatory model that involves these perspectives can then provide better explanations, such as a Tube map showing the outage for the material and formal explanations, or a geographic map visualising traffic delays for the mechanistic explanation.

It is known that users engage in different modes of abductive reasoning relative to their contextual cognition, which implies appropriate explanations be selected using psychological theories (such as transference and dependence [31]) relevant to contextually invoked layers in the explanatory model. For example, human evaluations of formal explanations [32, 33] were found to favour explanations consisting of intrinsic properties (particularly normative categorical associations) over those using extrinsic qualities (such as statistical correlations or provenance). This inherence bias does not imply that extrinsic factors should not be considered in formal explanations, but instead provides insight into human modes of cognition that should be considered in explanatory models for XAI.

Similarly, there is cognitive bias that is relevant to explanation when there is a broader context, such as a preference for functional explanations over mechanistic explanations when the explanation aligns with the intention, but a preference for mechanistic explanations when the intention diverged from the outcome (e.g., failing a goal) [31]. This is important as it indicates that explanatory models need not be able to give explanations at all levels simultaneously, and can be practicable within the context an XAI is operating within without this capability. However, if this is missing, then deployments of such XAI should still be made while cognisant of these levels, as they are indicative of questions that may be asked of explanations. Therefore, analysis within the framing of all levels of explanation should still be performed, to ascertain what levels are necessary without sacrifice to usability, which may be of benefit to technical implementations of XAI.

Causal Explanation

Perhaps the foremost form of explanation in XAI is causal explanation. This involves some model of reasoning which determines what factors and processes involved in a system were responsible for a given outcome. Notably, this may isolate and ignore that which is only statistically correlated with the outcome, and so some XAI techniques which solely argue based on correlation are excluded from this category. There are exceptions which integrate these statistical arguments as part of their reasoning, in particular, we highlight abductive reasoning as a family of sound methodologies that makes full use of available and inferred statistical arguments, as it has been shown to be significant in human reasoning and explanation [34, 35]. Abductive causal explanation has a long history in AI, with prior research in expert systems aiming at providing human-centred explanations [1]. One benefit of abductive reasoning in this role is that it can establish probable explanations in absence of given priors, or make full use of existing information (such as global model behaviour) to compute the likely explanations efficiently and reliably.

However, solely evaluating or maximising the probability of an explanation does not fully determine whether it is preferred by users for causal explanation, as a simultaneous and competing bias towards simplicity has been demonstrated [28]. Explanations with fewer causes that support more events (simple and general) are preferred over complex ones, although overgeneralisation is to be avoided [36, 35]. This makes it disproportionately difficult, but not impossible, for more likely but more complex explanations to overcome this bias towards simplicity. It follows that XAI that produce understandable explanations will typically focus on simpler explanations, even at cost of statistical likelihood. This stands in contrast to much of the existing XAI literature, though some methods have human-centric verification of their statistical arguments showing correspondence to known explanatory models [37].

The cognitive bias towards simplicity also comes up against another bias relevant to causal explanations; how obstructions elevate the importance of preconditions over the intention, such as not having enough time or money [38, 27]. This may be a special case of probability overcoming simplicity, as obstructions that lead to failed actions are typically events with near-certain probabilities that prevent an agent from fulfilling its goals, but we feel it intersects with the bias towards abnormal events, escalating the relevance of obstructions [38, 27]. In scenarios like this, the biases emphasise the impact of an abnormal obstacle instead of the intended outcome, which may be useful when establishing explanations that are specifically actionable or can contribute to planning. Based on the problem at hand, this may be an important consideration to ensure users are more likely to understand and trust an explanation, as these biases offset the increased complexity.

A frequent misunderstanding for AI practitioners arises in mistaking causal attribution for a complete causal explanation. Such causal attribution is typically in the form of causal chains or closed-form, static visualisation of local and/or global feature attributions. While a useful component in a causal explanation, these are incomplete, and have not been found to be interpretable by users, often resulting in misunderstanding and misattributing causal factors despite the given causal attribution [3]. Motivation for how we might develop XAI that provides more intuitive, interpretable causal explanations using attribution is given by Miller [1, 30].

Contrastive Explanation

A result from the social sciences that has received little adoption from the XAI field is the utility and feasibility of contrastive explanations. These are defined in terms of "why was it P instead of Q" questions, meaning that only two cases (an event and a counterfactual) and the difference between them needs to be understood at a time by an explanatory model [39, 38]; this has been found integral in causal reasoning and generalisation [34]. It is surprising that few approaches in XAI use this, as this seems much more technically tractable than causal attribution, where the causes of an event must be fully enumerated or reliably approximated — a taxing prospect.

Reasoning around the counterfactual in a contrastive explanation is well established in the social sciences, with humans engaging in a similar exercise to many XAI by (mentally) simulating the counterfactual via "what if" questions. A critical challenge lies in how to efficiently perform this with a large search space of possible counterfactuals, however humans often seem to handle this with aplomb, so the technical viability may improve by learning from humans here. Research in the social sciences has devised and demonstrated practical heuristics and biases that guide contrastive reasoning and explanation in humans, such as preference for abnormal events or selecting for high degrees of freedom [36, 1]. With an appropriate explanatory model, implementing and evaluating these biases may unlock significant improvements to both the latency and the understandability of explanations.

XAI with contrastive explanatory models can thereby establish user understanding as part of an explanatory process, as explanations will either be coherent with the existing mental model and understanding of a user, or they will be divergent. This may enable an XAI to actively aid a user that misunderstands an explanation or demonstrates a flawed mental model relative to the problem (say, due to insufficient financial literacy) by probing the user with specific pedagogical answers. Such an XAI can therein provide a learning experience to the user, enabling further development of trust in situations it may have otherwise been prevented in, typified by the annoying pattern of "talking past one another".

These contrastive explanations can also provide technical capabilities for XAI, as the heuristics seen in humans can similarly enable an XAI to consider many complex explanations despite the large search space of causal chains, which includes identifying when not to provide a counterfactual if there are simpler or more appropriate explanations without them.

To complete our motivation for contrast, we provide a worked example of this inspired by the abnormal condition model [40], where events outside the norm are given a much greater weighting when determining what counterfactual explanation to give. When asking "why was person A late while person B was on time?", we can compare "there was a disruption on A's journey" against "person B usually arrives five minutes early". The disruption is an example of an abnormal event, unlike the latter explanation, and is preferred under this model. Both are valid explanations, however it has been demonstrated that the abnormal event can be preferable in explanations, even with a lower prior chance [40, 41, 42].

Interactive Explanation

For any non-trivial system, there is no single, self-contained explanation that will suffice for every given user. This is often described as there being no "silver bullet" [43], with the suggestion that an ensemble of explanations should always be at hand (though perhaps not presented all at once). An important corollary of this is that XAI likely has need for interactivity to select the appropriate explanation. When viewed from the perspective of the social sciences, however, an explanatory model is a conversation [44], which requires some degree of interactivity (else it would be a monologue), and is therein distinct from causal attribution. From there, it follows that there are many explanations we can give, in ensemble, that reflect different questions asked of an explainer. This observation gives insight into how an explanation is necessarily constructed with the user by some recursive conversational model until the user reaches a satisfactory level of understanding [45].

Given that different users have distinct goals and reasons for interacting with an XAI, it is clear that XAI should have awareness of this, as it has been shown to affect the nature of the explanation required and expected [46]. It may also be pertinent to interactively provide specific levels of explanation according to the context, as there may be many possible perspectives. This might be achieved by iterative refinement and selection, which enables exploration for the explanatory model relative to the user (discovering their relevant context) and in reverse enables the user to explore relative to the model (to find alternative contrastive explanations). This represents a learning process, especially as combined with the prior explanatory qualities, which enables the user to learn generalisations [34, 35] external to the XAI.

One underexplored topic from the social sciences is the understanding for failure modes in human explanation. Models for this would be essential for analysing and correcting user misunderstanding, flawed explanatory models, and confusing explanatory presentations. Indeed, further to the capacity for user learning, we must know how users might not understand an explanation to avoid them diverging further, which must be made methodologically manifest so XAI designers know what to avoid and how to resolve misunderstandings. There is a pressing need for this, as merely the attempt at explanation may not be found sufficient when under legal or regulatory scrutiny, so further measures should be taken to ensure the soundness of explanations with respect to preventing the negative outcomes of confusion or outright misunderstanding, instead of only focusing on ensuring the positive outcome.

A central pillar of interactivity is the selection of explanations that are relevant to the questions asked of the explainer [1]. This recognises two forms of context, that being the global context of the use of an XAI, and the local context of a specific invocation of its use. Such a dichotomy implies that there is a cooperative dialogue to establish the (mental) models of both the user and the explainer in-situ, which is necessary to, for example, avoid explaining something the user already understands. With this, the user can actively learn with guidance from the explainer. Existing work has developed explanatory models capable of this, such as those by Byrne [47] and Miller [30]. Notably models based on argumentation [48] can provide this with a rich logical structure that does not rely on verbal or textual interaction, enabling integration with existing XAI and ML techniques, including visualisations; see Vassiliades et al. for a survey of argumentation in XAI [49].

Previous Work

Our prior work involved the co-design of XAI to explore what design elements were important for user trust [50]. A condensed version of this was presented in a workshop [51]. What emerged from participant responses was how natural language explanations were found actionable and presentable, and that their development had come about largely due to the human-centred co-design process.

Natural language has fallen out of favour in XAI with the rise of deep learning systems, compared to the prior state-of-the-art during the era of expert systems, which was much more amenable to this mode of explanation [1]. While natural language generation continues to develop [52], it has been noted that XAI faces many challenges in adopting it whilst symbolic representation remains underrepresented [53].

We found, however, that satisfactory explanations did not require that novel text be generated, enabling the technical floor of implementation to be lowered to simpler state-machine and templating techniques. This makes natural language explanation much more accessible and feasible, but left the problem of evaluating the performance of explanations unsolved. In this space, the work of Miller [1] and Lombrozo [36] provided clear arguments that this was not an open problem, but rather one of implementation and validation according to transdisciplinary learnings. While this stood in contrast to accepted wisdom in many parts of XAI, this inspired us to follow our findings and lead to our ongoing work.

III. DISCUSSION

The frequent inadequacies in the research and engineering of our field are, to once again echo Rudin [6], evidence that we have not put sufficient emphasis and effort into understanding the problems we are ostensibly tasking AI to solve, nor the orthogonal problem of explanation in XAI. A model reflects the understanding of a problem by operationalising its solution, so to exclude the learnings from the social sciences and HCI without evidence to the contrary is shown to be to the detriment of AI, and could be construed as negligence.

It was previously shown that the number of researchers engaging with social sciences and human behavioural experiments is dismally low [54]. While this has improved in recent years, particularly with the emergence of HCXAI, the gap between the mainstream of AI research and XAI has not demonstrably narrowed. Therefore, the cross-pollination of this human-centric theoretical and experimental rigour, such as argued for here and by other researchers [1], remains untapped and therein underdeveloped — especially relative to the state-of-the-art in research and industrial endeavours.

We reiterate that XAI, as is currently deployed, finds sole use with AI and ML practitioners in an ad-hoc capacity [2]. While there is significant utility in XAI in debugging models prior to and during deployment, we feel this belies a presumption that end users and lay-persons are unable to understand these models, or otherwise a reluctance to engage with such external stakeholders. Regardless, current regulations [21] and incoming legislation [23] requires that any data processing, including that performed by AI, be able to provide sufficient explanation. The limit of liability has yet to be tested out fully in court, and will likely take some time to fully resolve, however a proactive response necessitates that XAI and IML form an extensive part of research and development. In order to further the field, we must have awareness of the areas in which AI has already shown historical oversights, such as improperly accounting for racism and sexism [19, 20], and take measures accordingly to prevent similar failings.

We argue that, as motivated, there are valuable transdisciplinary theories and results whose learnings that we can apply to XAI. We do not presuppose that all benefits and capabilities transfer across without loss or error, but instead are in favour of further experimental research that demonstrates their integration into XAI has occurred within reasonable margins. This comes from an argument that, to ensure XAI is operationally and theoretically valid, it must be experimentally verified in a human-centric manner, to ensure that it is fit for purpose and actually engages users properly with understandable, transparent, and trustworthy models.

For XAI in particular, the problem of explanation, and the evaluation thereof, should be our principle quandary. As demonstrated by the social sciences, statistical arguments for the correctness of an explanation are rarely satisfying, but may be useful as a secondary factor. Unfortunately, much of XAI is motivated by these arguments, and it has lead to explanations that were unintentionally misleading [3], which may be preventable by building upon teachings from the social sciences and HCI that guide how to generate, select, and present explanations for users and lay-persons.

Unfortunately, when under scrutiny, LLMs have not followed through on the much lauded capabilities that are so engaging the AI community. Similarly, nothing has demonstrated the capacity for breadth and depth that humans possess, with the state-of-the-art being either narrow and deep or broad and shallow. If we are to avoid the rampant escalation of model size and the environmental and human consequences therein [55], then we must accept that there is inherent, intrinsic knowledge and structure that should be incorporated into AI — ironically, this fact has been apparent in how neural network "architectures" are central to the learning that enables much of our state-of-the-art, so really it just extends this to ask what other structure and knowledge should be encoded. Recognition of this has enabled state-of-the-art that uses an ensemble of task-specific architectures and ingrained knowledge that enables human-level performance [56], validating this argument. Therefore, to establish what knowledge and structure is appropriate for a given problem domain, and how best to integrate it, we should engage with the existing and ongoing research in relevant fields that establishes useful models and evaluations, enabling proper motivation and empirical demonstration, including the baseline we ultimately compare to: humanity. With this, AI can engage with humans, augmenting their capabilities.

Ancillary to this point, at least at current, will be the research about generalised interactions. Currently, AI-human interactions still requires significant development, however this HCI will eventually give way to complex group conversations and interactions. How best should an AI and human cooperate in order to communicate together with another human, such as support agents with customers or experts with regulators? Similarly, how can AI-human pairings interact with one another efficiently, as we might envisage with the promises of assistant-level AI? This will likely involve asymmetry as humans use different (perhaps tailored) AI to meet their needs and desires, including ensembles, making all interactions potentially many-to-many. We are starting to see AI that seek to engage with these spaces [15], but shockingly little research has been done in this field to accommodate and evaluate this.

Speaking to the variability of scenarios that will be encountered in deployment of XAI, it may be prudent to ensure that users are considered in all steps of development. The user is the one that will engage with an XAI, ask questions of it and its explanations, and is the avenue through which computer decisions and support are made manifest. We should be respectful of the available time the user has to digest explanations and interact with a system, the purposes for which they are using or recipient to the system, and the degrees of freedom available to the user. In the real world, an employee, a customer, an ML engineer, a regulator, and so forth all differ in these regards and more, and all have valid perspectives and uses for AI; it seems best that XAI are available to meet these disparate requirements, rather than assume one-size-fits-all, as the broadly accepted lack of "silver bullet" already suggested.

IV. ONGOING WORK

Following the findings from our previous work, and motivated by the insights brought to XAI by Miller [1], we are developing counterfactual HCXAI based on the position described here. It is still in an early phase, with an ongoing iterative process that seeks to improve on the methodology of our previous work in both breadth and depth. Our hope is that we demonstrate HCXAI, as informed by the social sciences, can be both practicable and a reliable approach to providing and presenting effective explanations. By evaluating HCXAI using empirical methodologies that include human-centred considerations, we aim to reveal how researchers and industry might design counterfactual explanations that can promote trust and understanding of XAI for users. We have focused on explanations and HCXAI in the financial and banking domain as this is an area that has many high-stakes decisions that users should have effective explanations for and a significant degree of trust in. Stakeholders in this area also have a forward-thinking proactivity to the challenges in AI and ML, and emphasise providing explanations and transparency on all levels for customers, support staff, and regulators alike. It also provides additional challenges that have not been explored extensively in XAI, such as regarding adversarial attacks, which has only recently been recognised in AI as an important part of model integrity. We feel these challenges will be fruitful signposts to future developments within the space of human-centred XAI and IML.

V. CONCLUSION

This paper argues that human-centred explanations are a central part of HCXAI, which forms a grounded approach to XAI. We expand prior arguments for XAI that understands the problem domain deeply by reflecting upon XAI and explanation itself, provoking conceptions of XAI that are informed by explanatory models verified by the social sciences.

We hold the position that human-centred AI is viable and practicable through HCXAI research, drawing from principles that have been empirically validated by social sciences, HCI, and AI research. At present, XAI is insufficiently motivated by these learnings, with approaches justified in absence of empirical verification and human-centred considerations [57, 58, 59, 60]. We argue that the discussed qualities of causal, contrastive, contextual, and interactive explanations contribute holistically to ensuring understandable and human-centred explanations, and that there are technical arguments that suggest implementation is a viable endeavour.

We specifically posit that interactive selection of counterfactual explanations that are causal, contrastive, and contextual will demonstrate a practicable XAI approach that gives effective explanations, as indicated in our previous and ongoing work and noted by the social sciences. Counterfactual XAI already corresponds closely to existing models for effective explanation, however their presentation and mode of explanation is often lacking in practice [2], so we suggest an emphasis on natural language counterfactuals.

To clarify, we broaden our intent towards natural language as including structured and semi-structured outputs (including mixed-media), such as bullet points and iconography, which may correspond to natural language idioms (e.g., punctuated lists) and conversational elements — at times non-verbal, such as gestures and indication, emotives, and imagery. This is not a stance that all explanation should be natural language and textual only, but rather that more XAI should consider its usage when presenting explanations, and that XAI relying on graphical techniques can still benefit from the conversational considerations that comes from these explanatory models.

Based on the arguments presented here, the breadth and experimental validity of the literature we draw upon, recent work by others in actualising similar positions [30, 61], and our ongoing work, we feel that our position is substantive and occupies an emergent space relative to prior efforts. There is a clear need for XAI that operates at the user level, therefore ensuring correctness and soundness of an XAI should involve human-centred considerations to provide appropriate explanations for these users. While our ongoing work is tailored to XAI in financial domains, we feel that the given arguments still hold in adjacent domains and in general, though continue to recommend that any transfer be validated experimentally. With the shifting research environment and pressing situation, we feel that HCXAI and transdisciplinary work is critical in high-stakes problem domains, and remains beneficially applicable beyond those domains. It is our hope that HCXAI contributes towards a diversity of solutions to meet these challenges.

REFERENCES

[1] T. Miller, “Explanation in artificial intelligence: Insights from the social sciences,” *Artificial Intelligence*, vol. 267, pp. 1–38, 2019.

[2] U. Bhatt *et al.*, “Explainable machine learning in deployment,” in *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 2020, pp. 648–657.

[3] Y. Zhang, Q. V. Liao, and R. K. E. Bellamy, “Effect of confidence and explanation on accuracy and trust calibration in AI-assisted decision-making,” in *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, Association for Computing Machinery, 2020, pp. 295–305.

[4] S. M. Lundberg *et al.*, “Explainable machine-learning predictions for the prevention of hypoxaemia during surgery,” *Nature Biomedical Engineering*, vol. 2, no. 10, pp. 749–760, 2018.

[5] T. Munzner, *Visualization Analysis & Design*. A K Peters / CRC Press, 2014.

[6] C. Rudin, “Stop explaining black-box machine learning models for high stakes decisions and use interpretable models instead,” *Nature Machine Intelligence*, vol. 1, no. 5, pp. 206–215, 2019.

[7] The Washington Post. “Tech companies are axing ‘ethical AI’ teams just as the tech explodes.” (2023), [Online]. Available: https://www.washingtonpost.com/technology/2023/03/30/tech-companies-cut-ai-ethics/.

[8] M. Sap, R. LeBras, D. Fried, and Y. Choi, “Neural theory-of-mind? On the limits of social intelligence in large LMs,” 2023. [Online]. Available: https://arxiv.org/abs/2210.13312.

[9] A. Lenci, “Understanding natural language understanding systems. A critical analysis,” 2023. [Online]. Available: https://arxiv.org/abs/2303.04229.

[10] A. Narayanan and S. Kapoor. “GPT-4 and professional benchmarks: The wrong answer to the wrong question,” AI Snake Oil. (2023), [Online]. Available: https://aisnakeoil.substack.com/p/gpt-4-and-professional-benchmarks.

[11] T. Yarkoni, “The generalizability crisis,” *Behavioral and Brain Sciences*, vol. 45, e1, 2022.

[12] M. C. Frank *et al.*, *Experimentology: An Open Science Approach to Experimental Psychology Methods*. MIT Press, 2023. [Online]. Available: https://experimentology.io.

[13] K. Valmeekam, A. Olmo, S. Sreedharan, and S. Kambhampati, “Large language models still can’t plan (a benchmark for LLMs on planning and reasoning about change),” 2022. [Online]. Available: https://arxiv.org/abs/2206.10498.

[14] S. Bubeck *et al.*, “Sparks of artificial general intelligence: Early experiments with GPT-4,” 2023. [Online]. Available: https://arxiv.org/abs/2303.12712.

[15] Y. Shen, K. Song, X. Tan, D. Li, W. Lu, and Y. Zhuang, “HuggingGPT: Solving AI tasks with ChatGPT and its friends in HuggingFace,” 2023. [Online]. Available: https://arxiv.org/abs/2303.17580.

[16] S. Kambhampati. “Language imitation games and the arrival of broad and shallow AI.” (2021), [Online]. Available: https://cacm.acm.org/blogs/blog-cacm/256068-language-imitation-games-and-the-arrival-of-broad-and-shallow-ai.

[17] N. A. Smuha, M. De Ketelaere, M. Coeckelbergh, P. Dewitte, and Y. Poullet. “Open letter: We are not ready for manipulative AI – urgent need for action.” (2023), [Online]. Available: https://www.law.kuleuven.be/ai-summer-school/open-brief/open-letter-manipulative-ai.

[18] Europol. “ChatGPT - the impact of large language models on law enforcement.” (2023), [Online]. Available: https://www.europol.europa.eu/publications-events/publications/chatgpt-impact-of-large-language-models-law-enforcement.

[19] E. Drage and K. Mackereth, “Does AI debias recruitment? Race, gender, and AI’s “eradication of difference”,” *Philosophy & Technology*, vol. 35, no. 4, p. 89, 2022.

[20] E. Drage and F. Frabetti, “The performativity of AI-powered event detection: How AI creates a racialized protest and why looking for bias is not a solution,” *Science, Technology, & Human Values*, 2023.

[21] European Parliament and Council of the European Union, *Regulation (EU) 2016/679 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation)*, 2016. [Online]. Available: https://eur-lex.europa.eu/eli/reg/2016/679/oj.

[22] High-Level Expert Group on Artificial Intelligence (AI HLEG). “Ethics guidelines for trustworthy AI.” (2019), [Online]. Available: https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai.

[23] European Commission, *Proposal for a Regulation of the European Parliament and of the Council Laying Down Harmonised Rules On Artificial Intelligence (Artificial Intelligence Act) and Amending Certain Union Legislative Acts*, 2021. [Online]. Available: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A52021PC0206.

[24] C. Chen, O. Li, D. Tao, A. Barnett, C. Rudin, and J. K. Su, “This looks like that: Deep learning for interpretable image recognition,” in *Advances in Neural Information Processing Systems*, vol. 32, 2019.

[25] N. Tilmes, “Disability, fairness, and algorithmic bias in AI recruitment,” *Ethics and Information Technology*, vol. 24, no. 2, p. 21, 2022.

[26] M. Favaretto, E. De Clercq, and B. S. Elger, “Big data and discrimination: Perils, promises and solutions. A systematic review,” *Journal of Big Data*, vol. 6, no. 1, pp. 1–27, 2019.

[27] B. F. Malle, *How the mind explains behaviour: Folk explanations, meaning, and social interaction*. MIT Press, 2006.

[28] T. Lombrozo, “Simplicity and probability in causal explanation,” *Cognitive Psychology*, vol. 55, no. 3, pp. 232–257, 2007.

[29] H. Chockler and J. Y. Halpern, “Responsibility and blame: A structural-model approach,” *Journal of Artificial Intelligence Research*, vol. 22, pp. 93–115, 2004.

[30] T. Miller, “Contrastive explanation: A structural-model approach,” *The Knowledge Engineering Review*, vol. 36, e14, 2021.

[31] T. Lombrozo, “Causal–explanatory pluralism: How intentions, functions, and mechanisms influence causal ascriptions,” *Cognitive Psychology*, vol. 61, no. 4, pp. 303–332, 2010.

[32] S. Prasada and E. M. Dillingham, “Principled and statistical connections in common sense conception,” *Cognition*, vol. 99, no. 1, pp. 73–112, 2006.

[33] A. Cimpian and E. Salomon, “The inherence heuristic: An intuitive means of making sense of the world, and a potential precursor to psychological essentialism,” *Behavioral and Brain Sciences*, vol. 37, no. 5, pp. 461–480, 2014.

[34] B. Rehder, “A causal-model theory of conceptual representation and categorization.,” *Journal of Experimental Psychology: Learning, Memory, and Cognition*, vol. 29, no. 6, p. 1141, 2003.

[35] T. Lombrozo, “Explanation and abductive inference,” *The Oxford Handbook of Thinking and Reasoning*, pp. 260–276, 2012.

[36] ——, “The structure and function of explanations,” *Trends in Cognitive Sciences*, vol. 10, no. 10, pp. 464–470, 2006.

[37] M. T. Ribeiro, S. Singh, and C. Guestrin, ““Why should I trust you?”: Explaining the predictions of any classifier,” in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ser. KDD ’16, Association for Computing Machinery, 2016, pp. 1135–1144.

[38] D. J. Hilton and B. R. Slugoski, “Knowledge-based causal attribution: The abnormal conditions focus model.,” *Psychological Review*, vol. 93, no. 1, p. 75, 1986.

[39] P. Lipton, “Contrastive explanation,” *Royal Institute of Philosophy Supplements*, vol. 27, pp. 247–266, 1990.

[40] D. J. Hilton and L. M. John, “The course of events: Counterfactuals, causal sequences, and explanation,” in *The psychology of counterfactual thinking*, Routledge, 2007, pp. 56–72.

[41] K. Uttich and T. Lombrozo, “Norms inform mental state ascriptions: A rational explanation for the side-effect effect,” *Cognition*, vol. 116, no. 1, pp. 87–100, 2010.

[42] J. Samland and M. R. Waldmann, “Do social norms influence causal inferences?” In *Proceedings of the Annual Meeting of the Cognitive Science Society*, vol. 36, 2014.

[43] V. Arya *et al.*, “One explanation does not fit all: A toolkit and taxonomy of AI explainability techniques,” 2019. [Online]. Available: https://arxiv.org/abs/1909.03012.

[44] D. J. Hilton, “Conversational processes and causal explanation,” *Psychological Bulletin*, vol. 107, no. 1, p. 65, 1990.

[45] R. Dazeley, P. Vamplew, C. Foale, C. Young, S. Aryal, and F. Cruz, “Levels of explainable artificial intelligence for human-aligned conversational explanations,” *Artificial Intelligence*, vol. 299, p. 103 525, 2021.

[46] N. Vasilyeva, D. A. Wilkenfeld, and T. Lombrozo, “Goals affect the perceived quality of explanations.,” in *CogSci*, 2015.

[47] R. M. Byrne, “The construction of explanations,” in *AI and Cognitive Science*, Springer, 1991, pp. 337–351.

[48] A. Arioua and M. Croitoru, “Formalizing explanatory dialogues,” in *Proceedings of the 9th International Conference for Scalable Uncertainty Management*, Springer, 2015, pp. 282–297.

[49] A. Vassiliades, N. Bassiliades, and T. Patkos, “Argumentation and explainable artificial intelligence: A survey,” *The Knowledge Engineering Review*, vol. 36, e5, 2021.

[50] A. Blandin, “Identifying critical design elements for increasing trust in computers through co-designing XAI for a mobile banking app,” M.S. thesis, Swansea University, 2021.

[51] A. Blandin, M. Roach, D. Doneddu, J. Pearson, M. Jones, and D. Sullivan, “Co-designing explainable AI for a mobile banking app,” 2022. [Online]. Available: https://hcxai.jimdosite.com/hcxai-22-papers-and-videos/.

[52] A. Stent and S. Bangalore, *Natural Language Generation in Interactive Systems*. Cambridge University Press, 2014.

[53] E. Reiter, “Natural language generation challenges for explainable AI,” 2019. [Online]. Available: https://arxiv.org/abs/1911.08794.

[54] T. Miller, P. Howe, and L. Sonenberg, “Explainable AI: Beware of inmates running the asylum, or: How I learnt to stop worrying and love the social and behavioural sciences,” 2017. [Online]. Available: https://arxiv.org/abs/1712.00547.

[55] E. M. Bender, T. Gebru, A. McMillan-Major, and S. Shmitchell, “On the dangers of stochastic parrots: Can language models be too big?” In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 2021, pp. 610–623.

[56] Meta Fundamental AI Research Diplomacy Team (FAIR) *et al.*, “Human-level play in the game of Diplomacy by combining language models with strategic reasoning,” *Science*, vol. 378, no. 6624, pp. 1067–1074, 2022.

[57] L. H. Gilpin, D. Bau, B. Z. Yuan, A. Bajwa, M. Specter, and L. Kagal, “Explaining explanations: An overview of interpretability of machine learning,” in *Proceedings of the IEEE 5th International Conference on Data Science and Advanced Analytics*, IEEE, 2018, pp. 80–89.

[58] M. Clinciu and H. Hastie, “A survey of explainable AI terminology,” in *Proceedings of the 1st Workshop on Interactive Natural Language Technology for Explainable Artificial Intelligence (NL4XAI 2019)*, 2019, pp. 8–13.

[59] R. Guidotti, A. Monreale, S. Ruggieri, F. Turini, F. Giannotti, and D. Pedreschi, “A survey of methods for explaining black box models,” *ACM Computing Surveys*, vol. 51, no. 5, pp. 1–42, 2018.

[60] C. Molnar, *Interpretable Machine Learning: A Guide for Making Black Box Models Explainable*. Lulu Publishing, 2020. [Online]. Available: https://christophm.github.io/interpretable-ml-book/index.html.

[61] R. Dazeley, P. Vamplew, and F. Cruz, “Explainable reinforcement learning for Broad-XAI: A conceptual framework and survey,” *Neural Computing and Applications*, pp. 1–24, 2023.

Towards European Standards supporting the AI Act: alignment challenges on the path to Trustworthy AI.

Alessio Tartaro
Department of Humanities and Social Sciences
University of Sassari
Sassari, Italy
https://orcid.org/0000-0002-0382-3083

Abstract— **The European Commission has issued a draft standardisation request to CEN and CENELEC for the development of standards for safe and trustworthy AI. The standards will serve as the basis for the implementation of the AI Act when the Regulation is eventually adopted. International standardisation activities in the field of AI have already been underway for several years. European standard organisations can leverage the proliferation of international standards by adopting them to support the AI Act. However, the standardisation of artificial intelligence in Europe shall consider the specific requirements and objectives of the AI Act, including respect for fundamental rights and Union values. This paper discusses the main alignment challenges for European standardisation, and puts forward possible solutions to address these issues.**

Keywords—AI Act, AI standards, AI regulation, AI ethics

I. INTRODUCTION

The European Commission has issued a draft standardisation request to CEN and CENELEC in support of safe and trustworthy artificial intelligence [1]. The draft identifies ten areas in which CEN and CENELEC are called upon to draft and adopt standards. These standards will serve as the basis for the development of European harmonised standards to support the implementation of the AI Act [2] when the Regulation is finally adopted, paving the way for European standards for Trustworthy AI.

At the international level, standardisation in the field of artificial intelligence has been progressing for several years. ISO/IEC JTC 1 established Subcommittee 42 on Artificial Intelligence in 2017 [3]. In 2016, IEEE launched the Global Initiative on Ethics of Autonomous and Intelligent Systems, which includes the P7000 series of standards [4]. ITU has focus groups dedicate to AI since 2017 [5]. According to AI Standard Hub [6], a UK initiative dedicated to monitoring the field of standardisation for AI technologies, these three standard development organisations (SDOs) alone have published or are developing nearly 150 AI-related standards.

European standardisation organisations (ESOs) can leverage the proliferation of international standards in the field of artificial intelligence. Rather than starting from scratch, ESOs can adopt international standards to support the AI Act. Several international standards by ISO/IEC and IEEE that could support the AI Act have already been published, others are under development, and still others will certainly emerge in the near future.

Specific agreements between international SDOs and ESOs exist to facilitate the adoption of international standards as European standards. The Vienna Agreement and the respective implementing guidelines set out the terms of technical cooperation between ISO and CEN [7], [8]. The Frankfurt Agreement describes the cooperation arrangements between IEC and CENELEC [9]. These institutional arrangements are crucial to avoid duplication of effort, reduce time, and bring together the expertise needed to realise AI standards.

While international standards can provide a valuable foundation for the development of European standards supporting the AI Act, the adoption of these international standards should consider the potential misalignments between the content of the international standards and the specific requirements and objectives of the AI Act. This issue is particularly pressing, as the AI Act aims to ensure that AI systems marketed and used in Europe comply with existing laws on fundamental rights and Union values. However, international standards are typically developed to be neutral with respect to national and regional interests, as required by the WTO Principles for the Development of International Standards [10]. Therefore, the adoption of international standards as European standards should take into account any possible misalignments and work to address them to ensure that the resulting standards support the objectives and requirements of the AI Act. In addition, the use of European harmonised standard is an instance of co-regulation, where private market actors are entrusted with the implementation EU law through the codification of technical knowledge. Given this delegation of powers, it is crucial to ensure that the resulting standards are legitimate, accessible, and representative of the interests of all the affected stakeholders.

This raises two alignment challenges faced by ESOs and European institutions. The first narrow challenge involves aligning the content of standards with the requirements of the AI Act. When European and international standards are adopted as harmonised standards, they can be used to demonstrate compliance with the Regulation. Therefore, it is crucial that the content of these standards effectively operationalise the requirements of the AI Act.

The second broad challenge concerns the legitimacy of delegating regulatory power from public institutions to ESOs and private international SDOs. The AI Act primarily provides high-level requirements for high-risk AI systems, while the specific technical criteria necessary to satisfy these requirements will be established by harmonised standards. As a result, it has been argued that "standardization is arguably where the real rule-making in the Draft AI Act will occur." [11] This raises an important question of legitimacy, particularly given the AI Act's explicit goal of safeguarding fundamental rights.

This paper aims to examine the alignment challenges faced by European standardisation. It provides a clarification of the issues at stake and advances some possible solutions. It is organised as follows. The second section of the paper analyses the role of standards in the AI Act and the Commission's draft standardisation request. In the third section, the paper focuses on the main alignment challenges, distinguished between narrow and broad challenges. In section four, the paper makes some proposals to address these challenges. Finally, the conclusion summarises the key findings of this paper and describes future lines of research.

II. THE ROLE OF STANDARDS IN THE AI ACT

The EU AI Act establishes a risk-based regulatory framework to develop and use AI systems in Europe, providing different regulatory regimes for AI applications according to their category of risk. The proposed Regulation identifies four categories of risk: unacceptable risk, high risk, limited risk, and minimal risk. These risk categories are determined according to the impact of AI systems on health, safety, or fundamental rights of individuals.

Unacceptable risk AI systems are forbidden under the proposed Regulation. They include manipulative systems, social scoring systems, and (with some exception) biometric systems. The rationale behind the ban to these applications of AI systems is the level of harm that the manipulative, exploitative and social control practices enabled by these kinds of AI systems may cause to the health, safety, and fundamental rights of people. Although the scope and effectiveness of these prohibitions are debated [11], these provisions set a red line for Trustworthy AI.

High-risk AI systems, as defined in Article 6 and Annex III of the AI Act, include, among others, those used in education and vocational training, employment, workers management and access to self-employment, law enforcement, administration of justice and democratic processes [12]. For high-risk AI systems, the Regulation sets out a number of requirements that these systems must comply with in order to obtain the CE marking and be legally placed on the EU market. These requirements, outlined in Chapter II of the Regulation, cover the following areas: risk management, data quality and governance, technical documentation, record keeping, transparency and provision of information to users, human oversight, accuracy, robustness, and cybersecurity.

The AI Act is based on the regulatory approach defined in the New Legislative Framework (NLF) [13]. The NLF consists of a set of Regulations and Decisions establishing rules to ensure that products commercialised in the EU market meet certain safety and quality standards. The New Legislative Framework is implemented through a mix of conformity assessment, accreditation, and market surveillance. Under the NLF, the content of the legislation is restricted to the definition of essential requirements for products, while the technical specification of these requirements is left to the European harmonised standards. A European harmonised standard is a European standard adopted at the request of the Commission for the purpose of implementing Union legislation. ESOs, i.e., CEN, CENELEC, and ETSI, are responsible for creating harmonised standards following a request from the European Commission. When a harmonised standard meets the requirements outlined in the corresponding Union legislation that it aims to cover, the Commission publishes a reference in the Official Journal of the European Union. Only harmonised standards can be used to show compliance with relevant legislations and benefit from a presumption of conformity with legislative requirements [14].

The AI Act includes provisions for the use of harmonised standards to comply with the requirements of the Regulation. Article 40 provides that high-risk AI systems that conform to harmonised standards or parts thereof shall be presumed to be in conformity with the AI Act's requirements. This means that providers of high-risk AI systems can demonstrate compliance with the requirements of the Regulation by following harmonised standards that cover those requirements [15]. That is why, as Vaele and Borgesius state, "standardisation is arguably where the real rule-making in the Draft AI Act will occur" [11].

The Commission's recent draft standardisation request paves the way for the realisation of these standards [1]. The request contains the rationale and terms of the request, a set of requirements for the prospected standards, and a ten-item list of standards and standards deliverables to be drafted and adopted. CEN and CENELEC are the primary recipients of the request, with ETSI mentioned as a possible contributor to fulfil the request. The required standards will serve as a basis for the development and adoption of European harmonised standards to support the implementation of the AI Act when it becomes applicable.

The Commission encourages collaboration between ESOs and International SDOs, and the possible adoption of standards by ISO/IEC on the basis of the Vienna and the Frankfurt agreements. Furthermore, the Commission requires ESOs to involve SMEs and civil society organisations in the standardisation process. Moreover, the standards produced in response to the request should be aligned with the Commission's policy objectives in the field of AI. They include, in addition to the specific objectives of the AI Act, the safety of AI products and services, the respect of fundamental rights and European Values, the digital sovereignty of the Union, the growth of the AI market, the public interest, and the rights of persons with disabilities. Finally, the 10 areas on which CEN and CENELEC are to produce or adopt standards corresponds to some of the requirements and obligations put forward in Chapters II and III of the AI Act. They include risk management, governance and quality of datasets, logging, transparency, human oversight, accuracy, robustness, cybersecurity, quality management, and conformity assessment.

To meet the standardisation request, CEN/CENELEC may develop its own homegrown standards or adopt international standards as European standards. The latter is a common practice. Almost half of the 3,500 CEN / CENELEC harmonised standards cited in the Official Journal are based on international standards [16].

While the AI Act's reliance on harmonised standards may provide clarity and consistency for providers of high-risk AI systems, there are challenges ahead in aligning these standards with the AI Act's requirements. Additionally, there are concerns about the compatibility between the standardisation processes and democratic principles. These issues are discussed in the next section to provide a deeper understanding of the complexities of AI standardisation in Europe.

III. ALIGNMENT CHALLENGES FOR EUROPEAN AI STANDARDISATION

Whether CEN/CENELEC develops its own home-grown standards or adopts international standards, two types of "alignment challenges" arise. On the one hand, standards need to be aligned with the requirements and objectives of the Regulation ("Alignment with the AI Act" issue). This is particularly pressing in the case of the adoption of international standards as European harmonised standards. On the other hand, it must be ensured that the standardisation process is aligned as much as possible with democratic and participative processes ("Alignment with democratic processes" issue). This is particularly significant since these standards will be implementing a Regulation aimed at safeguarding the fundamental rights of individuals, a topic hitherto unknown to technical standardisation bodies.

A. Alignment with the AI Act

Several of the ten areas identified in the standardisation request are covered by international standards that are either published or currently under development. For example, there are standards addressing risk management (ISO/IEC 23894), data quality and governance (ISO/IEC CD 5259 series), transparency (IEEE 7001, ISO/IEC AWI 12792), human oversight (ISO/IEC 8200, ISO/IEC PWI 18966), accuracy (ISO/IEC TS 4213:2022), robustness (ISO/IEC TR 24029-1, ISO/IEC DIS 24029-2), and cybersecurity (ETSI SAI series). Other standards, although not directly related to the ten areas mentioned in the standardisation request, are relevant to some of the requirements of the AI Act, e.g., standards on bias (ISO/IEC AWI TS 12791, ISO/IEC TR 24027, P7003), testing (ISO/IEC TR 29119-11), verification and validation (ISO/IEC TS 17847). However, there is currently a lack of standards in the areas of logging, AI quality management system, and conformity assessment. Nevertheless, it is expected that standards will emerge for these areas, particularly at the level of ESOs, under regulatory pressure.

Although numerous standards are related to the domains outlined in the standardisation request, their suitability for operationalising the requirements of the AI Act is not guaranteed. In order to determine the appropriateness of a particular standard for a specific requirement of the AI Act, a thorough comparison between the requirement and the pertinent standard's content is needed. This methodology has already been adopted by a recent report on the subject by the JRC, the European Commission's knowledge and science service [17]. However, such "suitability analysis" is currently restricted to published standards. Consequently, the present analysis will focus on ISO/IEC 23894:2023 [18] in relation to the AI Act's Article 9 on risk management, and IEEE 7001 [19] in relation to Article 13 on transparency.

1) Risk Management: Article 9 and ISO/IEC 23894:2023

Article 9 of the AI Act requires the establishment of a risk management system (RMS) for high-risk AI systems to be implemented, documented and maintained throughout their entire lifecycle. The RMS shall be a methodical process, comprising several steps. It starts with identifying and analysing known and foreseeable risks associated with high-risk AI systems, followed by estimating and evaluating risks emerging from the intended use and any reasonably foreseeable misuse. The RMS shall also evaluate other possible risks based on data gathered from post-market monitoring. Upon completion of this analysis, suitable risk management measures shall be adopted.

The goal of the RMS is to ensure that the overall residual risk is acceptable. These residual risks shall be communicated to the user. When implementing risk management measures, the focus should be on eliminating or reducing risks through appropriate design and development choices, implementation of mitigation and control measures for risks that cannot be eliminated, and provision of adequate information and training to users.

Article 9 also provides that high-risk AI systems shall be tested to identify suitable risk management measures and ensure they perform consistently. The testing procedures shall be appropriate for the intended purpose of the high-risk AI system and performed at suitable points during the development process, before placing on the market.

Presently, the sole published standard concerning AI risk management is ISO/IEC 23894. An examination of this standard reveals that although it integrates certain aspects of Article 9 and the corresponding area of the draft standardisation request, it fails to incorporate others. On the one hand, the standard aligns with Article 9 by establishing that risk management is executed iteratively throughout the entire life cycle, documented comprehensively, and aimed at reducing risk to acceptable levels. On the other hand, it does not mandate communication of residual risks to users or consideration of risks associated with foreseeable misuses. Additionally, it lacks specific testing procedures for identifying risk management measures and does not provide metrics or probabilistic thresholds for conducting testing. Nevertheless, guidance on this matter may be found in standards for testing of AI system like ISO/IEC TR 29119-11 [20].

Another element that misaligns ISO/IEC 23894 from Article 9 is the different conception of risk that underlies the standard compared to the AI Act. The standard focuses on organisational risks, which are uncertainties that could impact an organisation's objectives, while the Regulation considers risk as potential harm to individuals' health, safety, and fundamental rights. This factor limits the adequacy of ISO/IEC 23894 in operationalising the AI Act's RMS requirements.

Finally, a key limitation of the ISO/IEC 23894 is that it provides recommendations rather than requirements. In the standardisation field, the distinction between requirements, typically introduced by "shall", and recommendations, introduced by "should", is crucial because only standards containing requirements can be used for conformity assessment. Standards supporting the AI Act must be suitable for conformity assessment in order to verify that an AI system comply with the Regulation's requirements. However, since it is not possible to demonstrate compliance with ISO/IEC 23894, this standard is not an appropriate solution for operationalising the RMS requirements of the AI Act.

2) Transparency: Article 13 and IEEE 7001

Requirements on transparency and provisions of information to users are provided in Article 13 of the AI Act. It requires that high-risk AI systems be designed and developed to ensure a sufficient level of transparency that allows users to interpret the system's output and use it properly. High-risk AI systems shall also be accompanied by instructions providing concise, complete, correct, and clear information that is relevant, accessible, and easy for users to comprehend.

The information provided shall include several elements concerning the characteristics, capabilities, and performance limitations of the high-risk AI system, including its intended purpose, the level of accuracy, robustness, and cybersecurity tested and validated for the system, and any known or foreseeable circumstances that may impact these levels. The information shall also specify any known or foreseeable circumstances that could pose risks to health, safety, or fundamental rights. Moreover, the system's performance concerning the persons or groups it is intended to be used on shall be described.

Additionally, Article 13 establishes that human oversight measures, including the technical measures implemented to facilitate users' interpretation of AI system outputs, be detailed. Lastly, the high-risk AI system's expected lifetime should be disclosed, along with necessary maintenance and care measures to ensure proper functioning, including software updates.

IEEE 7001 is pertinent to Article 13 of the AI Act and the corresponding area of the draft standardisation request. The standard defines six levels of transparency, ranging from level 0 (no transparency) to level 5 (highest attainable level), and specifies these levels for three stakeholder groups: users of autonomous systems, the general public and bystanders, and expert stakeholders. For each group, criteria are given to assess the transparency level of an AI system from 0 to 5. The standard can serve two purposes: evaluating the transparency of an existing AI system through a System Transparency Assessment (TSA) or guiding the design of a new system by addressing the transparency needs of each stakeholder group. However, IEEE 7001 does not provide specific guidance on incorporating transparency into AI systems.

IEEE 7001 aligns with Article 13 in multiple ways. Firstly, it considers transparency as the disclosure of information concerning the system's purpose, context of use, capabilities, and limitations. Secondly, the standard assumes that the goal of transparency is to enable users understand the AI system. Thirdly, it acknowledges the necessity to differentiate the type of information provided according to users. This is in line with the requirements of Article 13. In addition, IEEE 7001 explicitly mandates information disclosure on AI system performance, particularly for the stakeholder category of "validation and certification agencies and auditors" at transparency level 3. As required by Article 13, IEEE 7001 also requires the release of information about data and maintenance instructions at transparency level 1 for the stakeholder category of "users". As discussed in the JRC report [17], IEEE 7001 also establishes requirements relevant to Articles 12 and 14 of the AI Act on logging and human oversight, respectively.

IEEE 7001 may be suitable for assessing whether an AI system is compliant with the requirements of Article 13. To achieve this, it is essential to determine the appropriate level of transparency required for each stakeholder category in accordance with the AI Act's requirements. The highest transparency levels in IEEE 7001, such as levels 4 and 5, establish requirements that go far beyond those of the AI Act, for example on explainability. In some cases, level 1 is sufficient for compliance with some of AI Act's requirements, while in other instances, it is necessary to reach level 3, depending on the stakeholder category.

3) Narrow challenges for European AI standardisation

These suitability analyses enable us to pinpoint key challenges in ensuring that future harmonised standards align with the requirements of the AI Act.

First, the current alignment, or lack thereof, between international standards and the AI Act's requirements is coincidental. Both ISO 23894 and IEEE 7001 were being developed prior to the European Commission's release of the AI Act's initial draft. As a result, the alignment of IEEE 7001, for example, is unintentional. Nevertheless, IEEE 7001 demonstrates that international standards can be adopted as harmonised standards when their requirements are more stringent than those of the AI Act. In such instances, it is possible to identify which clauses of the standard are to be followed to meet the requirements [21].

Better alignment can be achieved through the development of new standards by ESOs, which can explicitly consider and operationalise the AI Act's requirements within the standards. Moreover, this approach promotes greater alignment with European specificities, such as EU policy objectives and European values and fundamental rights. However, this seemingly straightforward solution encounters significant practical obstacles, such as time constraints, lack of expertise, and market resistance to regional standards in favour of global ones.

Another challenge lies in linking the AI Act's requirements to appropriate standards or their respective sections. The complexity of this task arises from the low probability that a single standard will match exactly with one of the AI Act's requirements. Often, an article in the AI Act contains provisions that are operationalised across multiple standards. For example, Article 9, which focuses on risk management, also covers testing requirements that are separately dealt with in the standardisation process. On the other hand, a single standard may be relevant to several AI Act requirements, as illustrated by IEEE 7001's connection to parts of Articles 12, 13, and 14.

It is therefore crucial to evaluate a standard's suitability, preferably during the development phase rather than *ex post*. However, the assessment of compliance with legislative requirements for documents drafted by ESOs remains unclear. According to the Regulation on European standardisation [22], the Commission and the ESOs "shall assess the compliance of the documents drafted by the ESOs with its initial request". Nonetheless, the scope and depth of the Commission's assessment are debated [23]. In the past, the publication of harmonised standards references in the Official Journal did not involve a systematic *ex ante* control of the standard by the Commission. More recently, however, the Commission has established a more thorough assessment of harmonised standards, involving external private entities such as HAS consultants [24]. Specifically, the assessment of a document's compliance with a request and any related legal requirements should focus on two main issues: identifying and assessing the extent to which the document covers and addresses the requirements specified in the request, and evaluating whether the document thoroughly addresses the covered requirements of a request. [25].

Beyond the question of the legal legitimacy of this assessment [23], such an evaluation seems necessary to ensure alignment between the standards and the AI Act's requirements. In particular, this process should involve examining both substantive and formal aspects to guarantee that the standard's content effectively operationalises the Regulation's requirements.

When determining the compliance of ESOs' documents with the standardisation request, fundamental rights and European values will be of utmost importance. In areas where technical and value concerns intersect, such as risk management, data quality and governance, and human oversight, ensuring the adequacy of standards can be challenging. For example, Article 10 requires datasets to be complete and error-free. This requirement, however, seems difficult if not impossible to implement [26]. In addition, Article 10 mandates an examination of training datasets for high-risk AI systems for potential bias. However, it is uncertain whether the technical specifications of the standards can provide a clear determination of what bias are and how to identify them, and especially whether they are entitled to do so [27].

This issue highlights a broader issue: the capacity of standards to sufficiently protect fundamental rights and European values and their legitimacy in doing so. This topic is further explored in the subsequent section.

B. Alignment with democratic processes

The delegation of powers to ESOs entailed by the NLF has been criticised in the past for many reasons. Constitutional concerns have been raised about the use of harmonised standards to support European legislation [28]. ESOs operate as private rule makers that are not bound by the same transparency requirements as public authorities. Moreover, these organizations are not held to the same level of oversight and accountability as governmental bodies. Since these standards are formulated by non-elected private entities, they lack the democratic legitimacy typically associated with public rule-making.

Another issue revolves around the inclusiveness and accessibility of the standardisation process for all interested stakeholders. While the Regulation on European standardisation encourages the involvement of organisations representing civil society interests in standardisation, there are doubts about the effectiveness of these provisions in ensuring the effective participation of social stakeholders [29].

Because of these concerns, the new European standardisation strategy [30] aims to improve the governance of ESOs and make their decision-making process more transparent, accessible and inclusive in order to increase their legitimacy. This is particularly relevant to the standardisation of artificial intelligence and become even more significant due to the fact that harmonised standards must contribute to the objectives of the AI Act, which aims to safeguard fundamental rights.

1) Legitimacy of the European standardisation process: Transparency, inclusiviness, and accessibility

The legitimacy of using harmonised standards to support European legislation is a subject of intense debate among legal scholars [31]. Legal scholarship distinguishes between three dimensions of legitimacy. First, "input legitimacy" concerns the involvement of stakeholders in the development of standards that directly impact them. This element ensures that their perspectives and interests are taken into consideration. Second, "throughput legitimacy" pertains to the presence of mechanisms that ensure transparency and accountability during the standard-setting process. These mechanisms aim to maintain the integrity and fairness of the procedures involved. Lastly, "output legitimacy" focuses on the positive outcomes resulting from the standardisation process, such as increased efficiency, predictability, and overall benefits for those affected by the standards [31].

As ESOs are private entities, determining the legitimacy of their role as *de facto* rule-makers can be challenging. To ensure legitimacy, ESOs should at least consider and protect all citizens' interests in standard setting, guaranteeing equal representation and balancing stakeholders' interests. However, this is not always the case.

The Regulation on European standardisation attempts to provide legitimacy by establishing participation and transparency obligations for ESOs. However, many stakeholders face significant challenges in effectively participating in and influencing the standardisation process. SMEs and social stakeholders, such as Annex III organisations representing consumers (ANEC), the environment (ECOS), trade unions (ETUC), and smaller companies (SBS), only possess observer status without voting rights [32]. As a matter of fact, standardisation is predominantly driven by industry. The European Commission acknowledges this by stating that the industry is "the key stakeholder in standardisation, the 'engine' of all standardisation and the main influencer of European standardisation" [25]. Even among industry actors, large players generally wield much more influence than other market participants. Small and medium-sized enterprises are often practically excluded from participating in technical committees due to financial and time constraints [31]. Finally, the standardisation process is completely intransparent to the general public. No information is publicly available on the technical committees, the participants in the standardisation process or the decisions taken [27].

The problem of insufficient legitimacy is further exacerbated for international standards not developed by the ESOs, as they are not required to comply with the procedural prerequisites outlined in the Regulation on European standardisation. When an international standard is adopted as a European standard, there is no assurance that all stakeholders' interests have been considered in the process. However, when these standards are adopted as harmonised standards to implement European legislative acts, they have legal effect such as presumption of conformity.

Finally, as discussed in more detail in the previous section, legitimacy issues also arise in relation to the effective *ex ante* control by the Commission over ESOs' standards before their adoption as harmonised standards. Strict control by the Commission would be in opposition to the core philosophy of the New Legislative Framework, which encourages public-private partnerships and co-regulation as a means to develop better policies [23].

Ensuring a sufficient degree of transparency, inclusiveness, and accessibility is essential for guaranteeing an acceptable level of legitimacy in the standardisation process. This issue has grown increasingly urgent following the European Court of Justice's judgment according to which European standards "form part of EU Law" as "measures implementing or applying an act of EU Law" [34]. Despite this judgment, however, several legal questions remain unresolved and continue to be debated [35]. Unlike other legislative acts, standards are not subject to judicial review. In addition, they are protected by copyright, exempting them from the publicity requirement of the law. Moreover, ESOs maintain their status as private entities while producing standards that carry legal effects. The notion that standards are part of EU law has consequently raised additional questions [36], but these matters fall beyond the scope of this paper and will not be discussed here.

Following the ECJ judgment, the European Commission implemented a series of changes to the standardisation procedure in order to comply with the court's decision. One significant change was the shift from publishing decisions on the publication of harmonised standards references in the C series (Communication) of the Official Journal to the L series (Legislation). The Commission also replaced the previous system of NA Consultants (New Approach Consultants) with HAS Consultants (Harmonised Standards Consultants) since mid-2018. Finally, the Commission modified its standardisation requests, providing more detailed information about the harmonised standards to be adopted and introducing an expiry date for the validity of the decision [23].

2) *The new standardisation strategy*

The changes triggered by the ECJ judgment culminated in the new standardisation strategy [37] and a proposal to amend the Regulation on European standardisation [38]. Both initiatives aim to improve the internal governance and decision-making procedures of ESOs. On the one hand, they want to improve the transparency, accessibility, and inclusiveness of the standardisation process in order to strengthen its legitimacy. On the other hand, they want to improve the ability of harmonised standards to foster EU values, policy objectives and regulatory implementation.

The standardisation strategy underscores the significance of harmonised standards within the EU single market, which have facilitated companies' compliance with EU regulations and advanced interoperability, safety, and environmental protection. Nevertheless, the strategy also recognises the mounting global competition in standardisation, as countries outside the EU adopt aggressive approaches to provide their industries with a competitive advantage. In many cases, private, non-European industry-led consortia are more efficient in developing standards, leading to Europe losing its "first mover" advantage in new and emerging technologies. Consequently, the strategy advocates for a shift, proposing measures to reinforce the centrality of standards in fostering a resilient, environmentally friendly, and digitally driven EU single market, while bolstering the European standardisation system's global influence. This is even more important since "more than ever, standards do not only have to deal with technical components, but also incorporate core EU democratic values and interests, as well as green and social principles".

The strategy aims to enhance the standardisation system by addressing urgent standardisation needs in crucial areas such as medicine production, recycling of critical raw materials, semiconductor chips, data interoperability, and digital technologies. To address these urgencies and anticipate future needs, the Commission intends to implement several measures. First, they will establish work programmes to set standardisation priorities. Second, the Commission has created a new High-Level Forum to advise on future standardisation needs, coordinate European interests in international standardisation, and ensure that standardisation activities align with EU goals of a green, digital, fair, and resilient economy. Additionally, the Commission will undertake a revision of existing standards to meet the EU policy objectives. They will also establish an EU excellence hub on standards to better coordinate existing standardisation expertise. The Commission will also implement measures to accelerate the development of standards that underpin EU legislation, ensuring a more efficient and effective standardisation system.

Finally, the strategy also focuses on improving access to standard development processes and standards themselves for SMEs and other societal stakeholders. ESOs are expected to make proposals to modernise their governance, address uneven and intransparent representation of industrial interests, and increase the involvement of SMEs, civil society, and users. The ESOs should also consider providing free access to standards and other deliverables.

In summary, the new standardisation strategy highlights the growing importance of European standardisation. As the significance of harmonised standards increases, it necessitates improvements in transparency, inclusiveness, and accessibility within the standardisation process. The standardisation of artificial intelligence serves as a test-bed for the new course of European standardisation. Nevertheless, it continues to confront substantial challenges, which will be explored in the following section.

3) *Broad challenges for European AI standardisation*

The concerns outlined in Section III.B.1 have also been raised in relation to the AI Act. Furthermore, unique issues stem from the distinct nature of the AI Act itself. While the Regulation is based on the New Legislative Framework, typically employed for product safety, one of its primary objectives is to protect fundamental rights. This discrepancy creates tensions between the chosen instrument and its intended goals.

Regarding the legitimacy of harmonised standards supporting the AI Act, concerns surrounding input, throughput, and output legitimacy have been emphasised [39]. In terms of input legitimacy, the concerns mentioned earlier are applicable. The nature of ESOs, which are essentially private entities dominated by large market players, raises questions about their suitability for specifying technical standards intended to safeguard fundamental rights. Moreover, the minimal control by public authorities over the standardisation process gives rise to two further challenges. Firstly, it becomes difficult to shape the content of these standards in a way that aligns with the objectives of the AI Act, as public authorities and social stakeholders have limited means to ensure that fundamental rights and ethical considerations are adequately addressed. Secondly, the risk of "regulatory capture" increases, where specific interest groups wield disproportionate influence over the standardisation process to serve their own interests [40]. This situation could potentially undermine the goals of the Regulation that the harmonised standards are meant to support, ultimately compromising the effectiveness of the AI Act.

Additionally, considering the lack of accountability, transparency, inclusivity, and accessibility of standardisation process, issues of throughput legitimacy arise. In the context of the AI Act, Small Business Standards, one of the Annex III organizations, observes that “SMEs are not included in standards development as they are underrepresented in standards organisations” and that “this leads to standards that are written in a way that is impractical and not applicable for SMEs” [41]. Similar concerns extend to other social stakeholders as well.

In a recent report by the Ada Lovelace Institute on “Inclusive AI Governance,” [42] representatives from ANEC express concerns that consumer perspectives are “often absent and disregarded” during standardisation processes. The report highlights those societal stakeholders face significant barriers to effectively participating in CEN/CENELEC JTC 21, the technical committee responsible for AI standardisation. These challenges encompass a range of issues such as the time commitment required for participation, the opacity and complexity of the standardization process, industry dominance in decision-making, and a general lack of awareness about the relevance of European standards for AI.

Lastly, it remains uncertain whether harmonised standards, which generally focus on technical matters, can ensure sufficient protection of fundamental rights. This casts doubt on the output legitimacy of these standards, as the benefits might be undermined by an exclusively technical approach that overlooks the broader ethical and societal implications, as well as the impact on fundamental rights, generated by AI systems.

A position paper by ANEC highlights the challenges of transposing fundamental rights and EU values and principles into technical standards from both substantive and process perspectives [43]. They contend that standardisation is not the suitable approach for guaranteeing sufficient protection of fundamental rights and should not be employed for such purposes. In a similar vein, the European Digital Rights Association (EDRI) asserts that technical standards should be restricted to addressing exclusively technical elements, without encroaching upon public policy and legal domains [44]. This viewpoint is further supported by a joint statement from various civil society organisations, which emphasises that standardization should focus solely on technical aspects, leaving other issues to be resolved through legitimate legislative processes [45].

IV. ADDRESSING THE CHALLENGES

These concerns about harmonised standards, i.e., alignment with AI Act’s requirements, issues of legitimacy, lack of transparency, inclusiveness and accessibility in the standardisation process, as well as inadequate protection of fundamental rights, challenge the compliance system underpinning the AI Act [46]. This section puts forward some suggestions for addressing these challenges.

First, it is crucial to ensure that the content of the standards effectively operationalises the AI Act’s requirements. Several paths are available to achieve this goal. The most direct approach, but also the most difficult in practice, is to develop internal CEN/CENELEC JTC 21 standards that explicitly operationalise each requirement of the AI Act. Alternatively, adopting international standards under specific conditions can be considered. Since the adoption of international standards as harmonised standards raises legitimacy concerns, European standardisation could employ “common modifications” [47] to address this issue. “Common modifications” allow parts of international standards to be modified to align better with European specificities and stakeholders’ needs. In any case, adequate control by European institutions during the standardisation process is key to overcoming the lack of alignment with policy and legal requirements that has characterised European standardisation in recent years, with only 27% of standards aligned with legislative requirements in the period 2016-2022 [48].

To address throughput legitimacy issues, it is essential to increase transparency in ESOs and make the standardisation process more accessible and inclusive. The report by Ada Lovelace Institute emphasises the importance of expanding civil society participation in JTC 21 for several reasons. Civil society representatives with expertise in human rights law and public policy can offer valuable input, providing missing perspectives and information on various ethical and cultural aspects. Additionally, a more balanced representation between civil society and large companies can prevent decisions made solely in companies’ interest, potentially conflicting with the public interest. The report suggests several ways to increase civil society representation in JTC 21, such as broadening the categories of Annex III organisations eligible for funding and mandated participation, funding more individuals from civil society and academia, and exploring the creation of a central hub to support civil society participation. These measures can enhance viewpoint diversity, leading to more balanced decision-making and higher-quality standards that consider the broader public interest.

V. CONCLUSION

This paper investigates alignment challenges for harmonised standards that will support the forthcoming AI Act. Firstly, it highlights the complexities in aligning a standard’s content with regulatory requirements. This challenge arises from factors such as the nature and procedures of ESOs, the influence of major market players in the standardisation process, the adoption of international standards, unattainable regulatory requirements, and the inherent difficulty of translating critical issues, like bias, into technical specifications. The analysis of two examples reveals the challenges in achieving alignment, which, when it does occur, seems coincidental. Recent data shows that only 27% of harmonised standards are genuinely aligned with their corresponding legislative acts, indicating that increased oversight by European institutions could be a potential solution.

Secondly, the paper delves into the broader issue of alignment regarding the legitimacy of the standardisation process. Drawing from existing literature, a distinction is made between input, throughput, and output legitimacy, and it is observed that standardisation faces challenges across all three dimensions. Specifically, in AI standardisation, harmonised standards may not sufficiently address fundamental rights issues and often neglect the needs and interests of diverse social stakeholders. By enhancing transparency, inclusiveness, and accessibility within ESOs, as recommended by the EU’s recently launched standardization strategy, these concerns could be mitigated.

However, several questions remain unresolved and are open to further investigation. The implications of harmonised standards’ status as part of European law are still unclear. ESOs are private organizations, and this casts doubt on whether their documents should produce legal effects such as a presumption of conformity. Given the implications of the AI

Act for fundamental rights, this issue becomes increasingly pressing for legal scholars.

Moreover, the uncertainty remains whether standards are adequate instruments to address fundamental rights issues, even with enhanced transparency, inclusiveness, and accessibility in the standardisation process. The protection of fundamental rights cannot be a one-time, set-and-forget process, relying solely on adherence to technical norms within a standard. If this were the case, the AI Act's structure might be undermined by an overreliance on the effectiveness of standards-based conformity assessments.

REFERENCES

[1] European Commission, Draft standardisation request to the European Standardisation Organisations in support of safe and trustworthy artificial intelligence. 2022. [Online]. Available: https://ec.europa.eu/docsroom/documents/52376

[2] European Commission, 'Proposal for a regulation of the European Parliament and of the Council laying down harmonised rules on artificial intelligence (Artificial Intelligence Act) and amending certain union legislative acts.' 2021. [Online]. Available: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex%3A52021PC0206

[3] 'ISO/IEC JTC 1/SC 42 Artificial Intelligence', ISO/IEC JTC1. https://jtc1info.org/sd-2-history/jtc1-subcommittees/sc-42/

[4] R. Chatila and J. C. Havens, 'The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems', in Robotics and Well-Being, M. I. Aldinhas Ferreira, J. Silva Sequeira, G. Singh Virk, M. O. Tokhi, and E. E. Kadar, Eds., in Intelligent Systems, Control and Automation: Science and Engineering, vol. 95. Cham: Springer International Publishing, 2019, pp. 11–16. doi: 10.1007/978-3-030-12524-0_2.

[5] 'AI for Good Global Summit 2017', ITU. https://www.itu.int:443/en/ITU-T/AI/Pages/201706-default.aspx (accessed Dec. 15, 2022).

[6] 'Standards Database', AI Standards Hub. https://aistandardshub.org/ai-standards-search/

[7] 'Agreement on technical co-operation between ISO and CEN (Vienna Agreement)'. 2001. [Online]. Available: https://boss.cen.eu/media/CEN/ref/vienna_agreement.pdf

[8] 'Guidelines for the implementation of the Agreement on Technical Cooperation between ISO and CEN (the Vienna Agreement)'. 2016. [Online]. Available: https://boss.cen.eu/media/CEN/ref/va_guidelines_implementation.pdf

[9] 'CENELEC Guide 13. IEC-CENELEC Agreement on Common planning of new work and parallel voting (the Frankfurt Agreement)'. 2016. [Online]. Available: https://boss.cenelec.eu/media/Guides/CLC/13_cenelecguide13.pdf

[10] WTO, 'Principles for the Development of International Standards, Guides and Recommendations'. 2000. [Online]. Available: https://www.wto.org/english/tratop_e/tbt_e/principles_standards_tbt_e.htm

[11] M. Veale and F. Z. Borgesius, 'Demystifying the Draft EU Artificial Intelligence Act', Computer Law Review International, vol. 22, no. 4, pp. 97–112, 2021, doi: 10.9785/cri-2021-220402.

[12] Annexes to the Proposal for a Regulation Laying Down Harmonised Rules on Artificial Intelligence and Amending Certain Union Legislative Acts. 2021.

[13] 'New legislative framework'. https://single-market-economy.ec.europa.eu/single-market/goods/new-legislative-framework_en (accessed Dec. 14, 2022).

[14] 'Harmonised Standards'. https://single-market-economy.ec.europa.eu/single-market/european-standards/harmonised-standards_en (accessed Dec. 16, 2022).

[15] M. McFadden, K. Jones, E. Taylor, and G. Osborn, 'Harmonising Artificial Intelligence: The role of standards in the EU AI Regulation'. Oxford Information Labs, 2021. [Online]. Available: https://oxil.uk/publications/2021-12-02-oxford-internet-institute-oxil-harmonising-ai/

[16] CEN/CENELEC, 'Feedback from CEN-CENELEC on the Roadmap of the "European Strategy for Standardization"'. [Online]. Available: https://ec.europa.eu/info/law/better-regulation/have-your-say/initiatives/13099-Standardisation-strategy/F2665566_en

[17] G. J. Soler et al., 'AI Watch: Artificial Intelligence Standardisation Landscape Update', JRC Publications Repository, Jan. 09, 2023. https://publicationstest.jrc.cec.eu.int/repository/handle/JRC131155

[18] ISO/IEC, 'ISO/IEC 23894- Information technology- Artificial intelligence - Guidance on risk management'. 2023. [Online]. Available: https://www.iso.org/standard/77304.html

[19] IEEE, 'IEEE Std 7001-2021. IEEE Standard for Transparency of Autonomous Systems'. 2021. [Online]. Available: https://standards.ieee.org/ieee/7001/6929/

[20] ISO/IEC, 'ISO/IEC AWI TS 29119-11 - Software and systems engineering - Software testing - Part 11: Testing of AI systems', ISO. https://www.iso.org/standard/84127.html (accessed Jan. 20, 2023).

[21] European Commission, 'Vademecum on European standardisation in support of Union legislation and policies. PART III - Guidelines for the execution of standardisation request'. 2015. [Online]. Available: https://ec.europa.eu/docsroom/documents/13509/attachments/1/translations

[22] Regulation (EU) No 1025/2012 of the European Parliament and of the Council of 25 October 2012 on European standardisation, amending Council Directives 89/686/EEC and 93/15/EEC and Directives 94/9/EC, 94/25/EC, 95/16/EC, 97/23/EC, 98/34/EC, 2004/22/EC, 2007/23/EC, 2009/23/EC and 2009/105/EC of the European Parliament and of the Council and repealing Council Decision 87/95/EEC and Decision No 1673/2006/EC of the European Parliament and of the Council Text with EEA relevance, vol. 316. 2012. [Online]. Available: http://data.europa.eu/eli/reg/2012/1025/oj/eng

[23] R. K. Dingemann and M. Kottmann, 'Legal Opinion On the European System of Harmonised Standards'. 2020. [Online]. Available: https://www.bmwk.de/Redaktion/EN/Downloads/L/legal-opinion-on-the-european-system-of-harmonised-standards.pdf?__blob=publicationFile&v=3

[24] EY, 'Call for expression of interest for Harmonised Standards (HAS) Consultants'. [Online]. Available: https://assets.ey.com/content/dam/ey-sites/ey-com/en_be/topics/advisory/ey-has-call-for-expression-of-interest.pdf

[25] European Commission, 'Vademecum on European standardisation in support of Union legislation and policies. PART I - Role of the Commission's Standardisation requests to the European standardisation organisations'. 2015. [Online]. Available: https://ec.europa.eu/docsroom/documents/13507/attachments/1/translations

[26] F. F. Liza, 'Challenges of Enforcing Regulations in Artificial Intelligence Act - Analyzing Quantity Requirement in Data and Data Governance', presented at the CEUR Workshop Proceedings, 2022. [Online]. Available: https://ceur-ws.org/Vol-3221/IAIL_paper9.pdf

[27] N. Maccabiani, 'The European path towards Data Quality and its standardisation in AI: a legal perspective', BioLaw Journal - Rivista di BioDiritto, no. 4, Art. no. 4, Dec. 2022, doi: 10.15168/2284-4503-2487.

[28] P. Verbruggen, 'Does Co-regulation Strengthen EU Legitimacy?', 2009, [Online]. Available: https://cadmus.eui.eu/handle/1814/13457

[29] M. Kallestrup, 'Stakeholder Participation in European Standardization: A Mapping and an Assessment of Three Categories of Regulation', Legal Issues of Economic Integration, vol. 44, no. 4, pp. 381–393, 2017.

[30] 'Standardisation strategy'. https://single-market-economy.ec.europa.eu/single-market/european-standards/standardisation-policy/standardisation-strategy_en

[31] M. Eliantonio and C. Cauffman, The Legitimacy of Standardisation as a Regulatory Technique: A Cross-disciplinary and Multi-level Analysis. Edward Elgar Publishing, 2020. [Online]. Available: https://www.elgaronline.com/display/edcoll/9781789902945/9781789902945.xml

[32] M. Eliantonio and C. Cauffman, 'The Legitimacy of Standardisation as a Regulatory Technique in the EU – A Cross-disciplinary and Multi-level Analysis: An Introduction', The Legitimacy of Standardisation as a Regulatory Technique, pp. 1–18, Jun. 2020.

[33] M. Eliantonio, 'Judicial Control of the EU Harmonized Standards: Entering a Black Hole?', Legal Issues of Economic Integration, vol. 44, no. 4, Nov. 2017, [Online]. Available: https://kluwerlawonline.com/api/Product/CitationPDFURL?file=Journals\LEIE\LEIE2017022.pdf

[34] Judgment of the Court (Third Chamber) of 27 October 2016. James Elliott Construction Limited v Irish Asphalt Limited. [Online]. Available: https://curia.europa.eu/juris/liste.jsf?language=en&num=C-613/14

[35] B. Lundqvist, 'European Harmonized Standards as "Part of EU Law": The Implications of the James Elliott Case for Copyright Protection and, Possibly, for EU Competition Law', Legal Issues of Economic Integration, vol. 44, no. 4, Nov. 2017, [Online]. Available:

https://kluwerlawonline.com/api/Product/CitationPDFURL?file=Journals\LEIE\LEIE2017024.pdf

[36] C. Tovo, ‘Judicial review of harmonized standards: Changing the paradigms of legality and legitimacy of private rulemaking under EU law’, Common Market Law Review, vol. 55, no. 4, Aug. 2018, [Online]. Available: https://kluwerlawonline.com/api/Product/CitationPDFURL?file=Journals\COLA\COLA2018096.pdf

[37] European Commission, ‘An EU Strategy on Standardisation - Setting global standards in support of a resilient, green and digital EU single market’. [Online]. Available: https://ec.europa.eu/docsroom/documents/48598

[38] European Commission, ‘Proposal for a Regulation amending Regulation (EU) No 1025/2012 as regards the decisions of European standardisation organisations concerning European standards and European standardisation deliverables’. Accessed: Apr. 01, 2023. [Online]. Available: https://ec.europa.eu/docsroom/documents/48599

[39] M. Almada and N. Petit, ‘The EU AI Act: Between Product Safety and Fundamental Rights’. 2022. doi: 10.2139/ssrn.4308072.

[40] M. Ebers, ‘Standardizing AI - The Case of the European Commission’s Proposal for an Artificial Intelligence Act’. Rochester, NY, Aug. 06, 2021. doi: 10.2139/ssrn.3900378.

[41] SBS, ‘Position Paper - SBS reply to the Artificial Intelligence Act consultation’. 2021. [Online]. Available: https://www.sbs-sme.eu/sites/default/files/publications/SBS%20position%20paper_AI%20Act%20consultation_08.21_2.pdf

[42] C. Galvagna, ‘Discussion paper: Inclusive AI governance’. 2023. [Online]. Available: https://www.adalovelaceinstitute.org/report/inclusive-ai-governance/

[43] ANEC, ‘Position paper: The role of standards in meeting consumer needs and expectations of AI in the European Commission proposal for an Artificial Intelligence Act - ANEC: The European consumer voice in standardisation’. 2021. [Online]. Available: https://www.anec.eu/publications/position-papers/924-anec-position-paper-the-role-of-standards-in-meeting-consumer-needs-and-expectations-of-ai-in-the-european-commission-proposal-for-an-artificial-intelligence-act

[44] EDRI, ‘The role of standards and standardisation processes in the EU’s Artificial Intelligence (AI) Act’. 2022. [Online]. Available: https://edri.org/wp-content/uploads/2022/05/The-role-of-standards-and-standardisation-processes-in-the-EUs-Artificial-Intelligence-AI-Act.pdf

[45] ‘An EU Artificial Intelligence Act for Fundamental Rights. A Civil Society Statement’. 2021. [Online]. Available: https://edri.org/wp-content/uploads/2021/12/Political-statement-on-AI-Act.pdf

[46] C. Castets-Renard and P. Besse, ‘Ex ante Accountability of the AI Act: Between Certification and Standardization, in Pursuit of Fundamental Rights in the Country of Compliance’. 2022. [Online]. Available: https://papers.ssrn.com/abstract=4203925

[47] ‘Common modifications to IEC (draft) standards’. https://boss.cenelec.eu/fadel/pages/commonmods/pages/

[48] European Commission. Directorate General for Internal Market, Industry, Entrepreneurship and SMEs. and Ernst & Young., Study on the implementation of the Regulation (EU) No. 1025/2012 (Article 24): executive summary. LU: Publications Office, 2021. Accessed: Apr. 03, 2023. [Online]. Available: https://data.europa.eu/doi/10.2873/593923

Artefactual ethics as opportunity to rethink “natural” ethics

Joel Parthemore[1] & **Blay Whitby**[2]

Abstract. This paper argues that, within the ethics community, the wider philosophical establishment and society in general, people have been far too quick to let themselves and, all too often, each other off the hook, at the same time as setting impossibly high standards for artefactual moral agents to meet, such that the artefactual agents should be guaranteed to make no mistakes. If artefacts are ever to be considered candidates for moral agency, then they should be held to no higher (and, at the same time, not significantly lower) a standard than what human beings can achieve. Meanwhile, the prospects of artefactual moral agency invite the opportunity for human moral agents to reconsider the standards they set for themselves and hold themselves to a higher standard.

1 Introduction

If one starts from the presupposition that ***any*** genuine moral agent should be held to the same standards regardless of the agent’s nature or origins – a position we argued for in [1] – the goal becomes finding an appropriate “middle” position between setting-up-for-failure and too-easy success. A follow-up paper [2] argued that, for an action to be morally right for a moral agent, one must have a convergence of the right motivations, the right means, and the right consequences. The underlying insight is that deontological, virtue-ethics-based, and consequentialist accounts all have their necessary role to play, but each tends to get too focused on itself and its merits to the loss of the bigger picture; while utilitarian accounts, as perhaps the most prominent division within consequentialism, face the further problem of failing to allow for those occasions where the needs of the few, or the one, outweigh the needs of the many, as Ursula K. LeGuin [3] so devastatingly addressed. Although the requirement to align motivations, means, and consequences may seem impossibly onerous, it need not be, provided one is prepared to allow that moral behaviour is far more difficult to achieve than it might seem at first glance and that mistakes will be made throughout the moral agent’s lifetime. Perhaps it matters more to take responsibility for those mistakes than to assure oneself, despite reasonable argument to the contrary, that one has avoided them. We argue that it is time to hold natural agents to a higher standard and purported artefactual agents to an achievable one.

2 The right consequences: Human beings, artefactual agents, and the responsibility game

For purposes of this paper, we take moral agency as the capacity to take responsibility and be held responsible for one’s actions with respect to ethical questions. This is in contrast to ***moral patienthood,*** where an entity has certain moral responsibilities attached to it on the part of moral agents; see, e.g., [4]. Moral agents are always moral patients; but many moral patients – say, any nonhuman animal that does not qualify for moral agency, or a contemporary robot that sufficiently resembles and behaves like a person or other animal to invite unconscious or pre-reflectively conscious comparison – are not moral agents.

2.1 The responsibility question

Intimately wrapped up in all matters moral is the ***responsibility question***: who individually has, and who collectively have, responsibility for any given action (or inaction) or set of events? People have been attributing agency of various kinds to virtual and physical artefacts at least since Eliza, when people expressed feelings for Eliza or complained bitterly about conversations with Eliza being interrupted. Given the advent of “self-driving” cars promising heightened safely (see, e.g., [5]) and “autonomous” battlefield robots (see, e.g., [6, 7]) promising a transformation of warfare – even more so with the rise of ChatGPT3/4 and kin, with all the hype attached to them – the responsibility question has only grown.

When a Tesla in full self-driving mode crashes and kills someone, who is responsible? Is it the driver who failed to supervise adequately or simply lost focus? Is it Tesla and, if so, who within the company? . . . The engineers who designed the systems and signed off on them? . . . The marketing department who pushed the hype? . . . The management who egged the marketing department on? . . . The car or the car’s AI itself?

To the last possibility, we offer a firm “no”, based on our earlier arguments. Attempts to assign responsibility to existing artefacts pose, we believe, a clear example of responsibility avoidance. Does this mean that artefacts cannot take responsibility, cannot be moral agents? To that, too, we would answer “no”. It is only that such agents would not appear to exist as yet.

---

[1] Guest researcher, University of Skövde, Sweden; joel.parthemore@his.se
[2] Visiting lecturer, University of Sussex, UK; B.R.Whitby@sussex.ac.uk

Otherwise, we would suggest that responsibility needs to be shared among the remaining players, with perhaps the greatest share reserved for management or the "person at the top", on the principle that "the buck stops here". If management is going to take a greater (even disproportionate) share of the profits then perhaps, too, management can take a greater share of the responsibility. Meanwhile, the driver arguably has diminished responsibility, given that the design of the car and the very name "fully self driving" invite the driver to relax and leave the car to take care of things, maybe even better than she can. She is still responsible for her (in-)actions and their consequences, but surely not nearly so much as others.

This raises a problem already seen on a more modest level with automatic transmissions, introduced in the late 1930s, and cruise control, invented in its modern form in 1948: by not requiring the driver to be as intimately aware of and involved in controlling the car, these innovations inadvertently facilitate a level of distraction. A frequently heard complaint with automatic transmissions is "I feel less a part of the car". Autonomous vehicles only heighten the problem.

Consider a parallel from aviation. The failure of an autopilot without warning produces what is known as "startle effect" in human pilots (Martin and Murray 2103). Despite being trained for this, it can take some time – typically 30 seconds – for a flight crew to appraise the situation and take back control. Perhaps the most famous failure of a flight crew to do so is provided by the crash of Air France flight 447, which collided with the south Atlantic ocean *four minutes* after the automatic system handed over control (exactly as it was designed to do in the circumstances). If highly trained, alert, and professional pilots cannot take control faster than 30 seconds – and sometimes never manage to pull it off – it seems absurd to suggest that an average driver could take back control from an autonomous vehicle in the fraction of a second that is typically required. The current practice of insisting (unrealistically) on constant monitoring by the human driver is clearly inadequate and unethical. Even though no one is claiming to be handing over control to the "full self-driving mode" system (in effect, a pseudo moral agent), that is – in effect – what is happening already today.

For an artefactual agent to *replace* the human driver or the pilot, it would need to function in all relevant ways like the driver or pilot, able to draw on a wealth of experience from outside highway driving or aviation. It would be both the driver *and* the vehicle, both the pilot *and* the plane. It might not always make the right choice – indeed, its very sophistication of design would preclude that – but it could then take responsibility, and it could be held accountable.

2.2 Who's to blame?

Just because existing artefacts do not qualify as moral agents does not mean that they cannot be a key part of the problem. It is striking how we're ready to blame existing artefacts and hand them responsibility in certain circumstances, just because they incorporate artificial intelligence, as a way to let ourselves off the hook; while in other circumstances we use artefacts' *lack* of intelligence to let ourselves off the hook. Consider a common response to attempts at gun-control legislation in the US: "guns don't kill people; people do". The point, presumably, is that guns can assume no responsibility for their actions, which is true but trivial. Guns – particularly certain kinds of guns including but not limited to semi-automatic assault weapons – make killing people (a lot of people quickly) a great deal easier.

More to the point for our purposes, there is a responsibility avoidance problem here, too – only here, instead of placing responsibility onto an "intelligent" artefact, there is the attempt, by claiming to deflect responsibility from obviously non-intelligent artefacts, to place blame solely on the individual who carried out the killings. One finds the frequent insistence, without evidence, that the way to treat gun violence is to have *more* guns in circulation, not fewer, accompanied by a refusal to acknowledge that society itself may be unwell. Opponents of gun-control legislation often say of someone involved in a mass killing, "he's just a sick person" or "he chose to be evil" or, with seemingly more thought, "obviously he's mentally unwell, but we can't cure mental illness". Although we agree that such an individual has to take ultimate responsibility for their actions and the consequences, that does not mean there is no responsibility to be shared – quite the contrary.

We let ourselves off the hook in other ways at other times, of course. One that is surely defensible – to a point – is when we acknowledge diminished responsibility, as when the Tesla driver is lulled by the seeming cleverness of full self-driving mode. Perhaps – as with a pilot deciding to override the automatic systems during flight – the penalty for being wrong will be particularly high. In the best case, she must be prepared to face a disciplinary hearing that could end her career; in the worst case, people die.

Perhaps one has acted with diminished cognitive capacity through no fault of one's own (e.g., as a result of hallucinations from high fever) or one's well-intended actions produce unintended and unforeseeable consequences (e.g., driving carefully at a safe rate of speed but nevertheless running over a small child who darts out suddenly from between parked cars). Diminished responsibility though is not absence of responsibility; and, even when the consequences could not even conceivably have been foreseen, that does not of itself make the actions morally right (only perhaps unavoidable). A less defensible excuse for letting ourselves off the hook is when we claim no alternative to acting in a way we would otherwise acknowledge as morally wrong – more on that later.

2.3 The ethical-standard question

Closely related to the responsibility question is the ***ethical-standard question*****:** if one accepts for sake of argument that a purported agent is a genuine moral agent, what is the appropriate standard to hold that agent to?

In attempting to answer this question, researchers interested in artefactual moral agency (e.g., [8, 9]) have tended to focus more-or-less equally on what artefacts do and what they fail to do – what morally relevant “choices” they make or fail to make – and here the standard objection is that existing artefacts either do the “wrong” things (e.g., battlefield robots producing “friendly fire”) or fail to do the “right” ones (say, making no response on seeing someone in danger, as with the Uber car that failed to brake for the pedestrian in Arizona).

Setting aside whether it provides an adequate litmus test for moral agency – as it is surely attempting to do – Colin Allen and colleagues' [9] proposed Moral Turing Test[3] sets a standard that, it would seem, no existing artefact could pass. That is to say that the standard is simultaneously too low (requiring no intentionality on the part of the moral agent) and ***too high*** (assuming implicitly that the artefactual moral agent should be incapable of making a moral mistake: because, after all, everyone “knows” that computers can't make mistakes[4]). Over-attribution of artefactual moral agency is met by bald under-performance. The Moral Turing Test invites the possibility of artefactual agents passing the test whom one, reasonably, would not want to acknowledge as genuine moral agents (perhaps because one is interested in more than actions and outcomes). It might be possible to rescue the test if one is a strict consequentialist, but otherwise there might seem an unavoidable sense of something missing.

The test faces other difficulties. It assumes that artefactual moral agents come off the assembly line having everything they need to be moral agents. The general consensus is that human newborns are not moral agents and that the process of their becoming moral agents is a gradual one. Why should one expect artefactual newborns to be different?

The test quite explicitly sets one standard for “natural” moral agency, another for artefactual moral agency – just as others have attempted to do with human vs. artefactual creativity, arguing that it is unfair to the artefacts to hold “man” and machine to the same standard. Such disjunctive accounts may have their place in certain discussions, but at least with respect to moral agency, it raises the distinct possibility that attitudes and prejudices to the one kind of moral agency will unconsciously colour discussion of the other kind. In any case, it is a position that we ourselves have repeatedly rejected ([1, 2]).

We believe that the Moral Turing Test is right to treat doing what is morally right, not *failing to do* what is morally right, and *not* doing what is morally wrong all on an equal footing: there is no morally relevant distinction that can be maintained solely on the basis of action vs. inaction. The mistake is in thinking we can just give the artefact a list it can follow when we cannot do that with children. The problem is that there is nothing close to universal agreement (outside *very* broad guidelines: “you shall not kill”; almost always with major caveats) for what is “right” and “wrong”, never mind one that can safely be applied regardless of context.

3 The right motivations: Action, inaction, and intention

> *A certain event may be desired under one of its descriptions, unwanted under another, but we cannot treat these as two different events, one of which is aimed at and the other not. And even if it can be argued that there are here two different events... the two are obviously much too close for an application of the doctrine of double effect* [10].

By contrast to their attitude toward artificially intelligent artefacts, psychology tells us that people are, ***ceteris paribus***, far more willing to excuse inaction in themselves or others – a failure to act – than to excuse actions they consider morally problematic.[5] To fail to save someone's life – to allow that death to happen (consider the highly skilled swimmer who, despite facing no apparent risk to herself, nevertheless stands by while someone is drowning) – is generally considered less morally wrong than to take a life, even if the two circumstances are, in all other relevant aspects, the same. At the same time, it seems difficult how one might logically justify how the passive vs. active nature of the behaviour could make the necessary difference – as, e.g., Sisela Bok [11] has pointed out in discussing the nature of lies. How is a ***lie of omission*** (what I fail to tell you) any less a lie than a ***lie of commission*** (what I tell you wrongly)? If the one is morally problematic, then so is the other.

Along similar lines, the ***Doctrine of Double Effect*** (DDE) – often invoked to uphold Roman Catholic thinking on abortion – holds that knowing that something otherwise morally unacceptable will happen as the unintended consequence of one's actions or inactions is at least sometimes acceptable whereas intending that same thing to happen would not be. The doctrine is necessary for reconciling moral absolutes (human life is sacred; therefore killing of innocent human beings is always wrong; human foetuses, having done nothing on which they could be morally judged, are innocent human beings; therefore abortion is always wrong) with real-world cases that would otherwise pose problems for those moral absolutes. (What if allowing the pregnancy to go to term – not performing an abortion – would kill the mother or both the mother and the child? Many defenders of the DDE would argue that that is morally preferable because the death of the child, though foreseeable and unfortunate, is not intended; whereas abortion is always an intentional act.[6])

---

[3] In brief, a purported agent is a moral agent if it takes what they consider the morally “right” decision a sufficiently high percentage of the time.

[4] This is already demonstrably false with existing systems. which are ***not*** Turing machines (a mathematical abstraction) but physically embodied systems (albeit ones that are embodied in very different ways from living creatures). Operate them outside their allowable boundaries of temperature and humidity, and they will “make mistakes” – likewise if certain of their physical components become sufficiently worn from use or, say, overclocking.

[5] If one objects that no one could excuse the human equivalent of the Arizona Uber case, the authors have personally encountered such excuses more than once.

[6] It is telling that almost if not without exception, recently passed abortion bans in the US make no exception for the termination of a non-viable foetus, forcing women to carry the foetus to term even though, say, it is effectively lacking a brain and no life is being saved. Most do make some allowance for the health of the mother, but in practice doctors have been reluctant to carry out an abortion if there is *any* shadow of a doubt that maternal death is imminent; mental health is strictly not considered – a reaction

The Trolley Problem, as originally (and succinctly) formulated[7] raises difficulties here, as the DDE can equally be used to argue for saving the life of the one person on the one track (with the unintended consequence of killing five on the other) or for saving the lives of the five at the cost of the one: it all depends on one's intentions, which Foot (rightly, we believe) declares an unacceptable state of affairs.

For Foot, intention is important but insufficient; means matter; and, clearly, she takes a utilitarian-inspired interest in numbers in favouring the lives of the five over that of the one: that is to say, a purely deontological approach is inadequate. For Foot, the outcome *must* be weighed along with the means and intention, where ***ceteris paribus*** one looks for the greatest good for the greatest number.

The ***ceteris paribus*** clause is necessary because of the equal inadequacy of a purely utilitarian approach, as LeGuin's short story – adapted for a first-season episode of ***Star Trek: Strange New Worlds*** – points out. It describes a society that is utopian in every conceivable way save one: all the prosperity and happiness is dependent on the interminable suffering of a single, neglected child. Even if, as in the Star Trek version, the child "willingly" accepts the burden, intuition suggests that some critical ethical line has been crossed.

For all her sympathy with those who oppose abortion and support the DDE, Foot sees merit not only in saving the mother's life at the deliberate loss of the child's – i.e., via abortion – when both would otherwise be certain to die; but also in pursuing abortion in cases where only one or the other might be saved. Foot rescues a version of the DDE at the loss of the possibility of absolute moral principles; but one might see this as a good thing. Claims to absolute moral principles may serve to excuse behaviour – as that by persons inclined to take a dogmatic as opposed to pragmatic position on abortion – that perhaps should ***not*** be excused. If artefacts are not allowed resort to sophistry, whether we think them capable of genuine sophistry or not, then neither should people be.

4 The right means: Hard-and-fast rules, rules of thumb, and ground rules

> *. . . Morality is not properly the doctrine of how we should make ourselves happy, but how we should become worthy of happiness.* [12]
>
> *The claim to have been 'just following orders' functions to mitigate one's culpability insofar as it positions oneself as having being a mere subsidiary actor in a chain of authority.* [13, p. 585]

The virtuous agent is the one who knows, within herself, what she needs to do and how to go about it, and who then cultivates the habits needed for putting that into practice in a faithful and disciplined way. For Aristotle, virtue cannot be arrived at by memorizing what one has to learn, nor can it be arrived at by some simple calculus; it needs to be lived. Aristotle's ***eudaimonia*** is commonly mistranslated as "happiness" (see, e.g., [14, pp. 7-8]), but it has more to do with being in the right place with oneself: living a life in accordance with virtue, whether that life happens (for circumstances beyond one's control) to be happy or not. ***Eudaimonia*** is a state of being. Most simply expressed, it is, for Aristotle, the highest human good.

So, virtue ethics is another essential piece of the puzzle. At the same time, it is no more successful on its own than deontological or consequentialist approaches – not because its formulations lead, as theirs do in isolation, to morally unacceptable outcomes, but because they are too vague to determine in any meaningful way whether an action is morally right or wrong. One knows a virtuous agent when one sees one. Virtuous agents lead by example, but one cannot become a virtuous agent simply by copying those examples.

Much ink has been spilled within the machine ethics community on what rules to hardwire into artefactual moral agents, and much effort has been made to draw inspiration or even borrow directly from Isaac Asimov's ***Three Laws of Robotics*** – despite the many times, in his stories, where Asimov showed just what impossible conundrums those rules created: a rule intended to anticipate every possible circumstance rarely if ever can. Such rules set a bar so high that not even those who clearly qualify as moral agents can reach it, never mind those whose moral agency may be considered in dispute. It is the problem of moral absolutes once again, in a slightly different guise: rules that broker no exceptions, no dissent. Even if one argues with Immanuel Kant that they should, human beings do not, in practice, function that way, even when they claim to do so. They are notoriously bad at following rules strictly; worse, when they try to, the results are often morally outrageous. Not only do human beings make mistakes, but they break rules, and sometimes the rule breaking seems to be a necessary part of achieving the morally right behaviour: putting that another way, blind-rule following does not qualify as the right means to the end; it exemplifies the opposite of virtue. The virtuous person knows when to follow the rules, when to bend them, when to break them.

Artefacts, by contrast, ***can*** be good at strict rule-following behaviour – but only when they are kept fairly uninterestingly simple. It is in the nature of an automaton – what Descartes took all non-human animals to be, lacking as he believed a human soul to give them free will – only to do what one is told. Part of being a moral agent depends on being able to make mistakes, recognize them as mistakes, and learn from them. Being a moral agent depends, too, on sometimes expressly *not* doing what one is told to do: consider the My Lai massacre, where William Calley defended his actions, in part, by saying that he was just following the captain's orders: what has elsewhere been called the Nuremberg Defence.

In practice, hard-and-fast rules ("always follow orders") morph into rules of thumb, which seem to fare better any-

---

against the means by which many US abortions were carried out prior to ***Roe v. Wade***.

[7] ". . . It may be supposed that [the man] is the driver of a runaway tram which he can steer from one narrow track onto another; five men are working on one track and one man on the other; anyone on the track he enters is bound to be killed" [10].

way. First-order predicate logic may rely on universal quantification, but the *lifeworld* [15]with which people engage on a daily basis has a habit of throwing up exceptions. If Foot is right – and we think she is – then *any* strictly rule-based approach will fail.

Perhaps the lesson to be learned from present-day artefacts, and the reason so few are willing to grant them moral agency – despite the haste with which others would do so – is not that they lack the right rules with which to make the right decisions; rather it is that they lack the capacity to make decisions or take responsibility for them in the first place – in no small part because they lack intentionality (they lack intentionality because they lack consciousness; they lack consciousness because they lack a mind). Remember that, by our definition, moral agency requires the capacity to take responsibility: something that – in company with newborn infants and certain among the mentally infirm (who nevertheless qualify as moral patients!) – present artefacts appear to lack. Most infants and at least some mentally infirm persons can be expected to outgrow their present conditions; by contrast, no amount of time and patience will, of itself, change present-day artefacts or their close kin into moral agents.

This is not to say that one can or should avoid hard-and-fast rules altogether. At least at first blush, the principle that what is acknowledged as morally wrong under a given set of circumstances should never simultaneously be accepted as morally right under those circumstances seems like a suitable candidate. Indeed, if one holds that moral right and wrong are mutually exclusive, then it follows of logical necessity.

Yet "lesser of two evils" arguments, widely used, require that the "lesser" evil is, at the least, morally acceptable if not strictly speaking "right"; and "just war" accounts – to take one example – critically depend on such arguments. The evil action (or inaction) becomes the good because, it is said, there is no alternative. Jean-Paul Sartre argued strenuously that, on nearly every occasion where people claim a lack of alternatives, there *are* alternatives; the problem is either that we fail to see them (our perspective, being non-Godlike, is limited) or that we fail to acknowledge them (we find them too distasteful, perhaps because they come at too great a cost to us). "Lesser of two evils"-type arguments are simply another way of letting ourselves off the hook. If people would not accept "lesser of two evils" arguments to excuse artefactual behaviour – and we believe that few would – then they likewise should not accept them to excuse their own. The alternative is bluntly acknowledging that sometimes we will do the morally wrong because we cannot see the morally right.

5 Weaving the three strands together

> *We often hear it asserted that most people do not really want freedom, because freedom involves responsibility, and most people are frightened of responsibility.*[8] *Whether or not this applies to 'most people' there is, I am sure, a vital element of truth in it. Accepting responsibility for our lives involves continually facing difficult choices and decisions, and bearing the consequences of them when we are wrong, and this is burdensome, not to say alarming. And there is something in all of us, something infantile perhaps, which would like to escape it by having the load taken from our shoulders.* [16, pp. 87-88]

The solution we propose is to let go of moral absolutes on the guiding premise that few things are *always* morally right or wrong regardless of circumstance and instead embrace personal responsibility, owning our chosen means, our honest motivations, our actions and their consequences in the thoroughly honest and uncompromising way Sartre [17] challenges us to do: taking ultimate responsibility for our actions and inactions alike, even when external forces seemingly have guided or even forced us there. Free will in the broadly libertarian sense does not depend on absolute freedom to do absolutely anything without constraint,[9] only a measure of genuine capacity in some small way, whenever we act, to do otherwise.

Such a solution means acknowledging both when we believe that we *have* done right, despite all evidence and arguments to the contrary, with a willingness and ability to defend the reasoning that led us there; *and* when we know or have reason to believe we have done wrong, either because we could not see an alternative or lacked the courage to embrace it. The proper solution to determining moral agency in artefacts is not to have one system for us, one for them; neither is it to suggest that we can reasonably hold artefacts to standards we ourselves cannot meet. It is rather to come up with one set of standards for any purported agent to qualify for moral agency: not a litmus test like the Moral Turing Test presents, but a set of reasonable – if painfully tough – expectations one is expected to fulfill if one can reasonably expect to be treated as, and continue to be treated as, a moral agent (not least including the capacity to perceive oneself as a person potentially deserving to be treated as a moral agent). Moral agency should be a lifelong challenge! The very real conceivability of future artefacts that qualify as moral agents on the terms set out here – artefacts that, perhaps, will have more in common with the lab-grown agents of Karel Čapek's play than the often clunky, mechanical, generally rule-following artefacts we are familiar with at present – invites the opportunity to raise the moral bar for ourselves as we wait for them to arrive.

References

[1] J. Parthemore and B. Whitby, "What makes any agent a moral agent? reflections on machine consciousness and moral agency," *International Journal of Machine Consciousness*, vol. 5, no. 2, pp. 1–25, 2013.

[2] ——, "Moral agency, moral responsibility, and artifacts: What existing artifacts fail to achieve (and why), and why they, nevertheless, can (and do!) make moral claims upon us," *International Journal of Machine Consciousness*, vol. 6, no. 2, pp. 1–21, 2014.

[3] U. K. LeGuin, "The ones who walk away from omelas," in *New Dimensions*, R. Silverberg, Ed. Doubleday, 1973, vol. 3, pp. 1–8.

---

[8] Note that this line is frequently misattributed to Sigmund Freud.

[9] . . . Which is not what Sartre had in mind, despite the frequent tendency to read him that way; "absolute" or "radical" freedom for him is more about choosing to follow or break the rules, behave as we're expected to *or not*, while taking full responsibility not just for our visible actions but our thoughts as well.

[4] E. Pluhar, “Moral agents and moral patients,” *Between the Species*, vol. 4, no. 1, pp. 32–45, 1988.

[5] S. Nyholm and J. Smids, “The ethics of accident-algorithms for self-driving cars: An applied trolley problem?” *Ethical Theory and Moral Practice*, vol. 19, pp. 1275–1289, 2016, https://doi.org/10.1007/s10677-016-9745-2.

[6] N. Sharkey, “Automating warfare: Lessons learned from the drones,” *Journal of Law Information and Scienc*, vol. 21, pp. 140–154, 2011.

[7] ——, “Killing made easy: From joysticks to politics,” in *Robot Ethics: The Ethical and Social Implications of Robotics*, P. Lin, K. Abney, and G. A. Bekey, Eds. MIT Press, 2011, ch. 7, pp. 111–128.

[8] W. Wallach, C. Allen, and S. Franklin, “Consciousness and ethics: Artificially conscious moral agents,” *International Journal of Machine Consciousness*, vol. 3, no. 1, pp. 177–192, 2011.

[9] C. Allen, G. Varner, and J. Zinser, “Prolegomena to any future artificial moral agent,” *Journal of Experimental & Theoretical Artificial Intelligence*, vol. 12, no. 3, pp. 251–261, November 2000.

[10] P. Foot, “The problem of abortion and the doctrine of double effect,” *Oxford Review*, vol. 5, pp. 5–15, 1967, available online from https://philpapers.org/archive/footpo-2.pdf (accessed 26 January 2020).

[11] S. Bok, *Lying: Moral choice in public and private life.* Vintage, 1999, first published 1978.

[12] *Kant’s critiques.* Start Publishing, 2013, ebook.

[13] S. Gibson, G. Blenkisopp, E. Johstone, and A. Marshall, “Just following orders?? the rhetorical invocation of "obedience" in stanley milgram’s post-experiment interviews,” *European Journal of Social Psychology*, vol. 48, no. 5, pp. 585–599, 2017, https://doi.org/10.1002/ejsp.2351.

[14] J. D. Mikalson, *The essential Isocretes.* University of Texas Press, 2022, https://doi.org/10.7560/325520.

[15] E. Husserl, *The Crisis of European Sciences and Transcendental Phenomenology: An Introduction to Phenomenological Philosophy*, D. Carr, Ed. Northwestern University Press, 1970, tr. David Carr. First published (in German) 1954.

[16] B. Magee, *Popper.* Routledge, 2013.

[17] J.-P. Sartre, “The flies,” in *The Flies and In Camera.* Hamish Hamilton, 1946, tr. Stuart Gilbert.

Part V.

Computational Creativity

Genetic Car Designer: A Large-Scale User Study of a Mixed-Initiative Design Tool
(A Work in Progress)

Sean P. Walton, Jakub Vincalek and
Alma A. M. Rahat
Computational Foundry
Faculty of Science and Engineering
Swansea University, Wales, UK
Email: s.p.walton@swansea.ac.uk

James Stovold
School of Computing and Communications
Lancaster University Leipzig
Nikolaistraße 10
Leipzig 04109, Germany

Ben J. Evans
Department of Aerospace
Faculty of Science and Engineering
Swansea University, Wales, UK

***Abstract*—This work-in-progress paper presents the initial results from a large-scale user study of a mixed-initiative design tool. A simple mixed-initiative design tool using a mutant shopping approach, coupled with MAP-Elites was created and made available to the public. In the tool participants were tasked with designing a simple car; the computer then presented sets of suggestions to participants to use and/or edit. Three out of four of these sets were generated using MAP-Elites and the forth was simply a control set of random designs. A total of 808 sessions were recorded and we found that participants who interacted with the algorithm produced better quality designs than the algorithm working alone. We also found that participants were statistically more likely to use suggestions from views which contained designs generated using a MAP-Elites algorithm compared to a randomised control.**

I. INTRODUCTION

Mixed-initiative design tools allow humans to collaborate with computers to perform creative tasks such as design [1]. One field which has driven development in the area of mixed-initiative design is procedural content generation (PCG) for video games. Initially, PCG algorithms were designed to fully automate creative tasks, removing the human from the loop completely with the goal of saving time and money [2]–[4]. As the field matured the emphasis moved away from replacing humans to save time, and towards augmenting and supporting human creativity [5]. This shift has resulted in a wide range of mixed-initiative algorithms and techniques which all attempt to support human creativity [1], [3], [4], [6]–[8].

Despite the increasing sophistication of mixed-initiative techniques there is a gap in our understanding of how these tools affect the creative process [9]. There has been increasing interest in performing user studies related to mixed-initiative systems [10], [11], but these generally have limited size and scope. Our contribution is to adapt an approach followed by Secretan et al. [12] for their work studying an interactive evolution tool, Picbreeder, and create a system which is interesting to the general public and use this to gather data for evaluation.

In this 'work in progress' paper we present some preliminary results, made up of 808 data points, obtained from the public deployment of a mixed-initiative design tool. This represents a field study where participants have interacted with our tool without our supervision and all the data collected is quantitative. We are currently undertaking a qualitative lab study with a smaller set of participants which will complete our data set for the final paper.

A. Research Questions

The research questions we aim to answer with the final paper, combining both the field and lab studies are as follows.

1) Do users interact with suggestions from MAP-Elite algorithms differently to a purely random set of suggestions?
2) Do MAP-Elite algorithms influence the engagement of users during the design process?
3) Do MAP-Elite algorithms influence the decision making of users during the design process?
4) Do MAP-Elite algorithms have an influence of the quality of designs produced in the design process?

II. BACKGROUND

A. Mixed-Initiative Design Systems

Lai et al. [13] provide a comprehensive survey of mixed-initiative design systems and propose a taxonomy of common techniques.

1) Grading: A human effort intensive approach, *grading* requires that the user rates the quality of every design presented to them by the algorithm numerically. For example, in [14] an evolutionary algorithm is used to evolve seat designs based on several computer calculated fitness functions. These fitness values are then augmented each generation by user assigned values, whereby the top ten solutions in the current population are presented to the user who then assigns a numerical fitness to each one. These user assigned fitness' are then used in future generations to select parents during tournament selection.

2) Rank-Based Interactive Evolution (RIE): Unlike with *grading*, RIE techniques do not require the user to rate the quality of every design numerically, but instead, asks the user to rank a set of designs. Subsequently the system will use this ranking to generate a numerical quality metric. For example, Liapis et al. [15] used RIE to generate maps for strategy games. A user preference model was constructed using the order in which users ranked a number of maps in terms of preference each iteration.

3) Overriding the Computational Algorithm: With these approaches the user can decide to lock parts of designs created by the computer such that the computer does not change it in the future. Some approaches also allow the user to manually edit parts of a design created by the computer. For example, in Tanagra [16], a mixed-initiative level design tool, the user and computer take turns to build on the work of the other. The developers of Tangara carefully ensured that the computer does not override decisions made by the user, but only augments human-placed geometry.

4) Mutant Shopping: This is where the computer presents the user which a set of candidate designs to the user to select from, and use or edit this design. For example, in the evolutionary dungeon designer [17] an underlying evolutionary algorithm generates various candidate maps based on a variety of developer specified fitness functions. The user can then click on one of the candidates to replace their current design with it.

B. MAP-Elites Algorithms

In many of the mixed-initiative approaches discussed above the computer presents a set of candidate designs to the human to select from and either use or edit. Generating this set of candidate designs can be done using a variety of algorithms, for example a tool could present the population of designs in a single generation of an evolutionary algorithm. More recently the community has turned towards using quality diversity algorithms, such as MAP-Elites [18], to create a diverse set of high quality designs [19] to present to the designer. Quality diversity (QD) algorithms are a subset of evolutionary algorithms which aim to create a diverse set of high quality solutions to a problem [19]. Modern QD algorithms evolved from methods developed to solve multi-modal function optimisation problems, such as niching [20], where it is important for the algorithms to maintain the quality and location of multiple solutions throughout the domain. One of the most common QD algorithms, which has spawned numerous variations, is MAP-Elites [18]. Algorithms based on Multi-dimensional Archive of Phenotypic Elites (MAP-Elites) create a map of high-quality solutions throughout a space whose dimensions are user defined and may not directly correspond to inputs and outputs of the function being optimised. Crucially, for the context of creating algorithms to support designers, MAP-Elites allows the designer to specify a meaningful ‘design possibility’ space to map.

MAP-Elite algorithms have been increasingly applied to mixed-initiative systems which fall under the category of mutant shopping. An early example of this is presented by Alvarez et al. [21], in an approach they call interactive constrained MAP-Elites (IC MAP-Elites) applied to dungeon map design. IC MAP-Elites allows the designer to select which dimensions (from a predetermined set) to use for the MAP-Elite algorithm. The underlying evolutionary algorithm then runs for a set number of generations and presents the elites to the designer as *suggestions*. MAP-Elites have also been applied to the design and re-balancing of the card game Hearthstone [22] illustrating their potential wide application to many aspects of design and creativity. The MAP-Elite methodology can also be used to assist in the exploration and expansion of crowd sourced content. For example, in ‘Baba is Y’all’, users can submit level designs which are then allocated to a cell in the MAP-Elites matrix [1]. Users are then guided towards designing levels which would fill gaps in the MAP-Elites matrix.

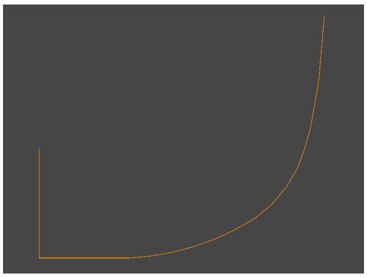


Fig. 1. An example of one of the courses participants can select from when starting the task. Cars are dropped into the course at the far left and simulated for 30 seconds. The fitness of a design is then the distance travelled from the first place the car hit the ground.

III. METHODOLOGY

A mixed-initiative design tool based on a mutant shopping approach was created and made available to the general public. Part of the tool uses MAP-Elites [18] to suggest designs to the participant, whereas a control view simply presents random designs to the participant. By collecting analytical data we are able to compare the way participants interact and use these different groups of suggestions—if at all. This approach is an extension of a triple-blind mixed-method approach to evaluating mixed-initiative design systems we introduced and validated in the context of video game level design [11]. At present we have only carried out the quantitative portion of the study and are presenting an overview of the data in this paper. Ethical approval was given to this work by the Swansea University Faculty of Science and Engineering Ethics Committee (SU-Ethics-Staff-030822/505). The tool is available for free on an independent video games store itch.io[1] and participants were recruited through a number of social media channels, primarily Reddit.

A. Design Task

Participants were given the task to design a simple car to travel as far as possible on a selection of courses. An example of one of the courses is shown in Figure 1. This is based on a popular web toy where users watch a genetic algorithm

---

[1] https://pillbuginteractive.itch.io/genetic-car-designer

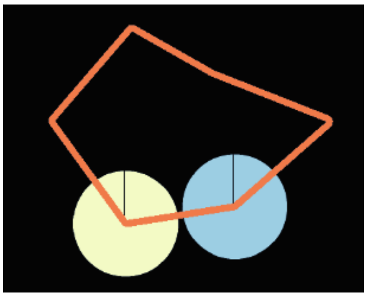


Fig. 2. An example of car design with 6 vertices and 2 wheels

design a car in real time [23]. The car, seen in Figure 2, is a closed polygon with wheels attached to some of the vertices. This task was chosen because (a) the existing popular web toy gave us confidence that we could recruit participants and (b) it is simple to understand. All physics is simulated using the two-dimensional rigid body engine in Unity3D.[2]

In terms of a formal definition of the task, this is an high-dimensional optimisation problem. Depending on the settings selected by the participant there could be between 9 and 85 degrees of freedom describing the car design. These degrees of freedom are made up of:

1) The mass of the car.
2) The radius of each vertex.
3) The vertex each wheel is attached to.
4) The radius of each wheel.
5) The mass of each wheel.
6) The parameters for the wheel joint[3] which connects each wheel to its vertex.

The fitness of a design is calculated by dropping the car at at the start of the course and recording the maximum signed distance travelled from the start location achieved during a 30 second simulation. This means designs which travel backwards at the start have a negative fitness.

Our aim was to encourage participants to explore and spend time with the tool, therefore we included options which encouraged multiple sessions. Participants were free to spend as long as they wished on the task and could perform the task multiple times with different courses and different numbers of degrees of freedom. We also implemented a simple replay recording system as well as the ability to export and share designs with other participants.

B. *Mixed-Initiative Approach*

Our tool uses MAP-Elites to implement a mutant shopping approach to mixed-initiative design. The design process is primarily driven by a simple evolutionary algorithm initialised with a random set of designs at the first generation. Each generation the designer has the opportunity to influence the optimiser by specifying which designs are to be used during tournament selection, and which designs to allow to survive to the next generation. The designer can either select existing designs from a series of views or edit their own design and select for breeding or survival. The views include a live view of the current generation, a control randomised sample of previous designs and several MAP-Elites driven views.

In MAP-Elites dimensions of interest related to the problem are defined. The MAP-Elites algorithm then generates a number of bins for this dimension and stores an archive of the best design found in each bin as the optimisation continues. In our tool the designer can select from three dimensions to explore using MAP-Elites: *speed insights*, *wheel insights* and *geometry insights*. Speed insights uses the mean speed of the wheels on the car and wheel insights uses the mean wheel radius. Geometry insights calculates the centre of mass of the car (including the contribution from the wheels) and measures the mean shift from the centre with respect to each vertex. An example of a view the user is presented with when browsing the MAP-Elites is shown in Figure 3. In addition to views which use MAP-Elites to present suggestions to the designer, our tool also includes a view labelled *history insights* which does not use MAP-Elites, but instead designs are assigned to bins at random every generation, this acts as a control in order to evaluate the effectiveness of using MAP-Elites. For clarification, designs only replace existing designs in the view if they are of higher quality. The order in which these views are presented to participants in the navigation bar, which can be seen at the top of Figure 3, was randomised each time the software is launched.

C. *Preliminary Results From Field Study*

The preliminary data presented in this paper was collected between 4th August and 4th September 2022. In total the tool was launched 2,274 times during this period, of these sessions 808 provided valid data with consent to be used for analysis. Participants spent between 1 minute and 4.3 hours on the task, with a mean time of 12 minutes. During each session our artefact recorded a range of analytical data for each session. This included:

1) The fitness of the best design.
2) The number of interactions with each view in the system. For example, the number of times a design from the editor was used in the next generation, or the number of times a design from a particular MAP-Elites view was exported to the editor.
3) The amount of time a participant spent in each view.

Most of these data points were recorded per generation enabling us to reconstruct a time line for each session.

An overview of the full data set is shown in Figure 4, where each session is plotted based on the improvement from the best design in the first generation to the best design found throughout the session and the total number of user interactions recorded in that session. Distributions of each of these dimensions is shown as a violin plot on the relevant axis. The Spearman rank-order correlation coefficient between these data is 0.401 with $p < 0.001$, indicating that sessions with

---

[2] https://unity.com/

[3] https://docs.unity3d.com/Manual/class-WheelJoint2D.html

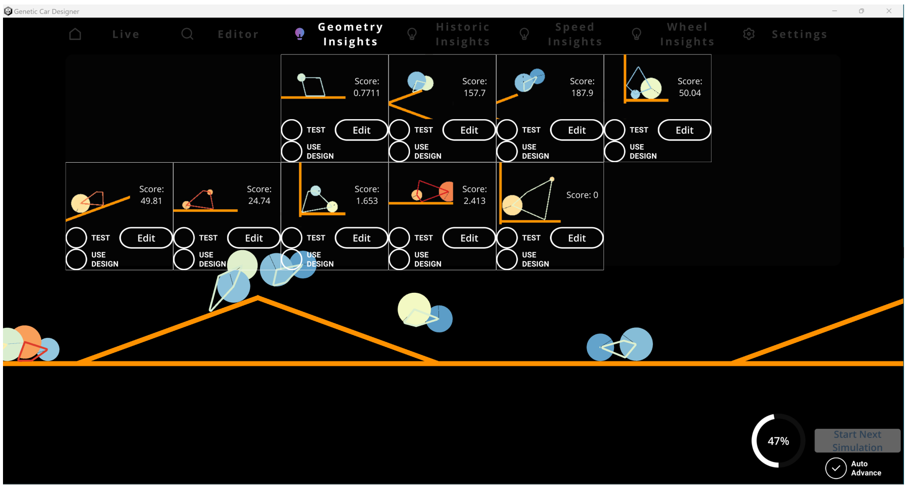


Fig. 3. An example of a view which shows a set of designs selected using MAP-Elites. From here the user can select which designs to use to create the next generation and which to test in the next generation. They can also click edit to edit the design in the edit view.

more interactions are likely to have achieved a better overall design. The size of each data point indicates the session length, larger is longer as indicated in the legend. Each session was placed into a participant group based on how the participant interacted with the system.

1) In 50% of the sessions, participants did not interact with the system at all, but instead simply watched the evolutionary algorithm design the car automatically.
2) In 36% of the sessions, participants only interacted with the editor, manually creating designs along side the evolutionary algorithm.
3) In 14% of the sessions, participants interacted with the editor and other views.

The group each session belongs to is indicated in Figure 4 by the colour and symbol of each data point.

Figure 5 shows a comparison between the distributions of improvement between sessions where participants did not interact, only interacted with the editor and interacted with both the editor and views. Table I shows the mean percentage improvements for each group. The no interactions group can be considered equivalent to the algorithm running without any input from the human designer and therefore be used as a baseline. Using a Mann–Whitney U test we find that both the sessions where participants interacted with the views and editor, and those where participants interacted with just the editor have statistically significant higher improvements than the base line ($p < 0.001$). There is however no statistical difference between those who interacted with just the editor and those who interacted with the editor and views.

Focusing on the sessions where participants interacted with the mutant shopping views we found that participants were more likely to use designs from the MAP-Elite views compared to the randomised control view. The distributions of the percentage of interactions on each view are shown in Figure 6, where an interaction is when a participant has either used or edited a design from a particular view.

D. Discussion

Preliminary results from a large scale user study of a mixed-initiative design tool have been presented. These results currently help us answer the first research question "Do users interact with suggestions from MAP-Elite algorithms differently to a purely random set of suggestions?". We found that participants were statistically more likely to use designs recommended using a MAP-Elites algorithm compared to a randomised control group. Furthermore, we found that participants who interacted with the algorithm were more likely to produce better designs than the algorithm working on its own. Our current data can not confirm whether or not a human working alone is likely to produce better designs than the algorithm or a participant working with the algorithm. Although we have data that allows us to identify participants who did not use designs from the various views, we can not verify if they did not take inspiration from those views and use that to inform their own design. This highlights the key weakness of the presented field study, a lack of context of the quantitative data. To address this we are currently carrying out a small scale lab study with a smaller number of participants. In this lab study we are using questions designed to assess recommender algorithms [24], [25] to compare the randomised control to a MAP-Elite algorithm. These questions are designed to evaluate the accuracy, diversity, user satisfaction and novelty presented by each recommender algorithm was evaluated using a 5-point Likert scale. An altered version of the original Genetic Car Designer was made for the purposes of the lab study. The changes included limiting the number of generations to 40, starting each session from the same seed, and simplified header tabs. Rather than the *Speed, Geometry, Wheel*, and *Historic Insights* tabs, the illumination

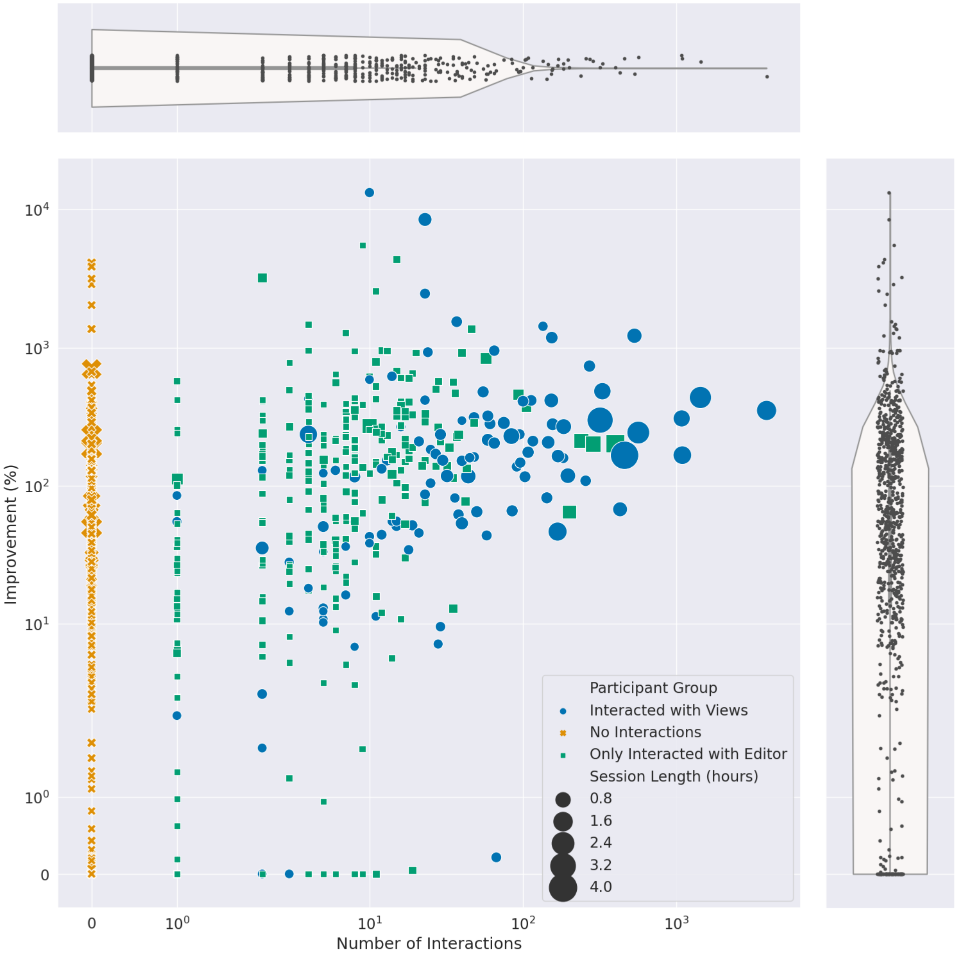


Fig. 4. An overview of the data collected in the field study, plotted using the number of interactions on the x-axis and percentage improvement in the fitness of the best design comparing the best in the first generation to the best found throughout the entire session. Each marker on the main central graph represents a single design session. The colour and shape of each marker groups the sessions into three groups based on how the participant approached the design task, and the size of the marker indicates the session length.

TABLE I

MEAN PERCENTAGE IMPROVEMENT IN DESIGN ACHIEVED IN EACH SESSION SEPERATED BY PARTICIPANT GROUP. THE ERROR PRESENTED IS THE STANDARD ERROR IN THE MEAN.

Editor Interaction	Views Interaction	Mean Percentage Improvement (%)
—	—	124 ± 19
✓	—	243 ± 30
✓	✓	420 ± 140

algorithm and the control were presented to participants in tabs labelled *Insights 1* and *Insights 2*. To prevent the order of the tabs influencing the participants, the illumination algorithm was randomly assigned to either *Insights 1* and *Insights 2*. Double-blind conditions were maintained during the study to prevent the authors from influencing participants. In addition to the Likert scale questions we have also asked open ended questions which can be analysed qualitatively.

In conclusion, our preliminary results suggest that the mixed-initiative tool implemented does have an effect on the creative process, and furthermore participants are more likely to use suggestions generated using MAP-Elites compared to a randomised control. A key weakness of our current approach is that all of the data is quantitative, we can see that there is

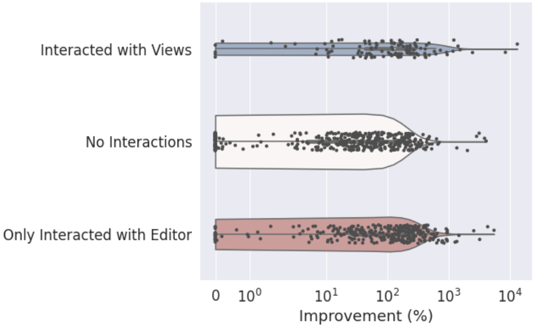


Fig. 5. Comparing the percentage improvement from the initial design for participants in different groups. Using a Mann–Whitney U test (rejecting at $p < 0.001$) the *interacted with views* is significantly different to the other distributions.

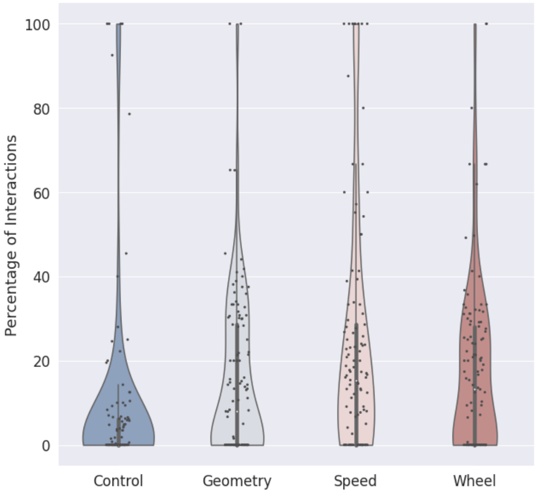


Fig. 6. Comparing the percentage of interactions for each of the mutant shopping views. Using a Mann–Whitney U test (rejecting at $p < 0.05$) the *control* group is significantly different to the other distributions.

an effect but we cannot determine the context or reason for that effect. To address this we will be carrying out a smaller scale lab study as discussed above.

REFERENCES

[1] M. Charity, I. Dave, A. Khalifa, and J. Togelius, "Baba is y'all 2.0: Design and investigation of a collaborative Mixed-Initiative system," *IEEE Trans. Comput. Intell. AI Games*, pp. 1–15, 2022.

[2] A. S. Melotti and C. H. V. de Moraes, "Evolving roguelike dungeons with deluged novelty search local competition," *IEEE Trans. Comput. Intell. AI Games*, vol. 11, no. 2, pp. 173–182, Jun. 2019.

[3] A. S. Ruela and K. Valdivia Delgado, "Scale-Free evolutionary level generation," in *2018 IEEE Conference on Computational Intelligence and Games (CIG)*, Aug. 2018, pp. 1–8.

[4] A. Baldwin, S. Dahlskog, J. M. Font, and J. Holmberg, "Mixed-initiative procedural generation of dungeons using game design patterns," in *2017 IEEE Conference on Computational Intelligence and Games (CIG)*, Aug. 2017, pp. 25–32.

[5] M. Cook, "A vision for continuous automated game design," in *Proceedings of the 13th Experimental AI and Games Workshop*. AIIDE, Jul. 2017.

[6] A. Liapis, G. N. Yannakakis, and J. Togelius, "Sentient sketchbook: Computer-aided game level authoring," in *FDG*, 2013, pp. 213–220.

[7] B. von Rymon Lipinski, S. Seibt, J. Roth, and D. Abé, "Level graph – incremental procedural generation of indoor levels using minimum spanning trees," in *2019 IEEE Conference on Games (CoG)*, Aug. 2019, pp. 1–7.

[8] S. Earle, J. Snider, M. C. Fontaine, S. Nikolaidis, and J. Togelius, "Illuminating diverse neural cellular automata for level generation," in *Proceedings of the Genetic and Evolutionary Computation Conference*, ser. GECCO '22. New York, NY, USA: Association for Computing Machinery, Jul. 2022, pp. 68–76.

[9] R. Craveirinha and L. Roque, "Studying an Author-Oriented approach to procedural content generation through participatory design," in *Entertainment Computing - ICEC 2015*. Springer International Publishing, 2015, pp. 383–390.

[10] A. Alvarez, J. Font, S. Dahlskog, and J. Togelius, "Assessing the effects of interacting with MAP-Elites," *AIIDE*, vol. 17, no. 1, pp. 124–131, Oct. 2021.

[11] S. Walton, A. Rahat, and J. Stovold, "Evaluating Mixed-Initiative procedural level design tools using a Triple-Blind Mixed-Method user study," *IEEE Trans. Comput. Intell. AI Games*, 2021.

[12] J. Secretan, N. Beato, D. B. D'Ambrosio, A. Rodriguez, A. Campbell, J. T. Folsom-Kovarik, and K. O. Stanley, "Picbreeder: a case study in collaborative evolutionary exploration of design space," *Evol. Comput.*, vol. 19, no. 3, pp. 373–403, May 2011.

[13] G. Lai, F. F. Leymarie, and W. Latham, "On Mixed-Initiative content creation for video games," *IEEE Trans. Comput. Intell. AI Games*, vol. 14, no. 4, pp. 543–557, Dec. 2022.

[14] A. T. Machwe and I. C. Parmee, "Enabling generative behavior within an interactive evolutionary design system using a component-based representation," in *2007 IEEE Congress on Evolutionary Computation*, Sep. 2007, pp. 2964–2971.

[15] A. Liapis, H. P. Martínez, J. Togelius, and G. N. Yannakakis, "Adaptive game level creation through rank-based interactive evolution," in *2013 IEEE Conference on Computational Inteligence in Games (CIG)*. ieeexplore.ieee.org, Aug. 2013, pp. 1–8.

[16] G. Smith, J. Whitehead, and M. Mateas, "Tanagra: Reactive planning and constraint solving for Mixed-Initiative level design," *IEEE Trans. Comput. Intell. AI Games*, vol. 3, no. 3, pp. 201–215, Sep. 2011.

[17] A. Alvarez, S. Dahlskog, J. Font, J. Holmberg, C. Nolasco, and A. Österman, "Fostering creativity in the mixed-initiative evolutionary dungeon designer," in *Proceedings of the 13th International Conference on the Foundations of Digital Games*, ser. FDG '18, no. Article 50. New York, NY, USA: Association for Computing Machinery, Aug. 2018, pp. 1–8.

[18] J.-B. Mouret and J. Clune, "Illuminating search spaces by mapping elites," *arXiv preprint*, Apr. 2015.

[19] J. K. Pugh, L. B. Soros, P. A. Szerlip, and K. O. Stanley, "Confronting the challenge of quality diversity," in *Proceedings of the 2015 Annual Conference on Genetic and Evolutionary Computation*. ACM, Jul. 2015, pp. 967–974.

[20] S. W. Mahfoud, "Niching methods for genetic algorithms," Ph.D. dissertation, USA, 1996.

[21] A. Alvarez, S. Dahlskog, J. Font, and J. Togelius, "Empowering quality diversity in dungeon design with interactive constrained MAP-Elites," in *2019 IEEE Conference on Games (CoG)*. ieeexplore.ieee.org, Aug. 2019, pp. 1–8.

[22] M. C. Fontaine, S. Lee, L. B. Soros, F. De Mesentier Silva, J. Togelius, and A. K. Hoover, "Mapping hearthstone deck spaces through MAP-elites with sliding boundaries," in *Proceedings of the Genetic and Evolutionary Computation Conference*, ser. GECCO '19. New York, NY, USA: Association for Computing Machinery, Jul. 2019, pp. 161–169.

[23] R. Matsunaga, "Html5 genetic cars: A genetic algorithm car evolver in HTML5 canvas," *GitHub repository*, 2020. [Online]. Available: https://github.com/red42/HTML5_Genetic_Cars

[24] M. D. Ekstrand, F. M. Harper, M. C. Willemsen, and J. A. Konstan, "User perception of differences in recommender algorithms," in *Proceedings of the 8th ACM Conference on Recommender systems*. ACM, oct 2014. [Online]. Available: https://doi.org/10.1145%2F2645710.2645737

[25] B. P. Knijnenburg, M. C. Willemsen, Z. Gantner, H. Soncu, and C. Newell, "Explaining the user experience of recommender systems," *User Modeling and User-Adapted Interaction*, vol. 22, no. 4-5, pp. 441–504, mar 2012. [Online]. Available: https://doi.org/10.1007%2Fs11257-011-9118-4

Exploring a Collaborative and Intuitive Framework for Combined Application of AI Art Generation Tools in Architectural Design Process

Lok Hang Cheung
Department of Architecture, Xi'an Jiaotong-Liverpool University

Suzhou, China
lokhang.cheung19@student.xjtlu.edu.cn

Juan Carlos Dall'Asta
Department of Architecture, Xi'an Jiaotong-Liverpool University

Suzhou, China
juancarlos.dallasta@xjtlu.edu.cn

***Abstract*— Artificial Intelligence (AI) art generation tools are widely explored in the conceptual design process in architecture. However, common applications view these tools primarily as rapid and high-quality image generators rather than design collaborators that converse with the designers. This limited application is further intensified by focusing on one specific tool for each use and a requirement for expertise in machine learning to design and access advanced functionalities beyond mere image generation. Therefore, the potential for adopting AI art generation tools in the architectural design process remains limited.**

This research compares several AI art generation tools commonly explored in architectural design. Followed by providing case studies of the two authors' exploration in university, an online workshop and personal explorations from 2020 to 2023. Based on the findings, a collaborative and intuitive framework for the combined use of AI art generation tools in the architectural design process is proposed, facilitating immediate application by architectural designers.

Keywords*— *AI art generation, architectural design process

I. THE FIRST WAVE OF AI ART IN ARCHITECTURE

Artificial intelligence (AI) has been widely explored in the architecture discipline, especially computer vision (CV) and natural language processing (NLP) techniques.

After generative adversarial network (GAN) [1] was proposed, it was applied in the architecture discipline of three main categories, including data synthesis, design automation and data augmentation, while 'design assistance' and 'design inspiration' were considered as part of data automation [2]. Although only 11% of GAN research in architecture is related to design explorations, it has increased by 85% from 2015 to 2021 [3].

More convenient AI art generation tools like DeepDream [4] and StyleTransfer [5] were explored afterwards. However, the intrinsic relationship between AI and the human creative mind remains unclear [6]. It urges a critical examination of applying AI art generation tools for the architectural design process.

II. THE SECOND WAVE OF AI ART IN ARCHITECTURE

Since 2021, AI art generation tools combining CV and NLP to generate high-quality images have been developed. The most common tools are StableDiffusion [7], Midjourney [8] and DALL-E 2 by OpenAI [9]. Designers can enter "prompt" (series of text commands) to generate high-quality images.

A. Misuses of AI art generation tools were criticized

Criticized by John Frazer while being asked about his comments on the AI art generation experiment at Zaha Hadid Architects, he stated it was "very dangerous" and "risking trivializing their work or Zaha's work" because focusing on visual aspect only while neglecting space, movement, environment, etc. was not appropriate architectural design process [10]. Likewise, Daniel Bolojan stated that misusing AI art generation tools to generate images has little to no difference from using a blender to mix different images and create new images [11].

B. Three observations from designers and researchers

First, despite the capability of producing "realistic visual representation" [12], there appears to be confusion regarding applying AI art generation tools to abstract concepts. Therefore, the application of AI art generation tools should evolve from an automatic process [13] to a collaborative [12], [14], and iterative process [15].

Second, while comparisons of various AI art generation tools have been initiated, the evaluations primarily focused on the AI techniques, speed and image quality [12], [16].

Third, challenges exist for designers in providing suitable prompts for the AI art generation tool to output images consistent with the design intent [17]. Therefore, designers should be able to adjust certain parameters (instead of using a black-box approach) [15]. Design researcher Erik Ulberg proposed and illustrated the potential of 'hand-crafting' neural networks [18], [19]. However, it could not be applied widely because the designer must possess expertise in machine learning knowledge to execute it effectively.

III. PROPOSAL OF A COLLABORATIVE, COMBINATIONAL AND INTUITIVE FRAMEWORK FOR AI ART GENERATION TOOLS IN ARCHITECTURAL DESIGN

A novel framework with three guiding principles is proposed to address the challenges that have emerged from the three identified research gaps.

Firstly, the architectural design process is a collaborative, iterative, and conversational endeavour instead of an automatic, linear, and scientific approach. It is necessary to emphasize that every design decision made will create a new "wicked" design problem; there is no "immediate" design solution [20]. Design activities such as sketching or modelling in design are considered conversational processes, serving as explorative thinking rather than explicit representations of thought [21], [22], [23].

The application of AI art generation tools in the architectural design process aligns with this notion. Consequently, this paper builds upon the same theoretical framework within the architectural design context, emphasizing our objective to facilitate the conversational process between designers and AI art generation tools.

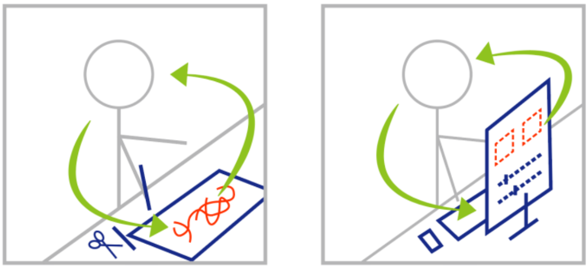


Fig. 1. Architectural design process as a conversational process (left: with physical sketching and modelling tools, right: with AI art generation tools)

Secondly, the criteria for selecting AI art generation tools should be based on the requirements of each design stage and the designers' understanding of the tools.

Instead of identifying a specific AI art generation tool that provides the best image quality or the fastest generation speed, the emphasis could be placed on the potential for combining various tools in a customized manner to address specific design needs.

Thirdly, AI art generation tools should be intuitive for designers, facilitating immediate use. While AI art generation tools nowadays rely on inputting prompts (a series of phrases or sentences) for operation, 'prompt-engineering' should not be the principal focus for designers. Emphasis should be placed on using images as inputs, given that architectural designers predominantly work with visual mediums such as sketches or 3D models.

The adjustable parameters provided by AI art generation tools should be aligned with designers' knowledge, ensuring ease of use. Additionally, it is important to avoid lengthy or expensive local computer setups and steep learning curves, thus making the tools more accessible for designers.

TABLE I. THREE OBSERVATIONS AND RESPECTIVE PROPOSALS

	Current observation	***Proposal***
Architectural design process	Automatic	Collaborative, iterative, conversational
AI art generation tool selection criteria	Best image quality, fastest generation speed	Combinational uses according to design need
Designer-friendly	Need to learn 'prompt-engineering' and AI to adjust the tool	Intuitive for immediate widely-use

IV. RESEARCH STRUCTURE

This research is composed of three parts. First, the common AI art generation tools are compared, focusing on identifying their potential applications. Second, a framework schematic is presented. Finally, case studies and illustrations of the two authors' attempts in an architecture university module in 2020, an online workshop on the DigitalFutures platform in 2021 and self-exploratory research in 2022 and 2023.

The self-research aimed to demonstrate the framework's flexibility by exploring various tools that cater to designers with differing levels of experience and knowledge. This includes designers with no computational knowledge, those who have previously used AI tools, and individuals with AI knowledge who intend to customize the tools for more advanced applications.

V. COMPARISON OF CURRENT AI ART GENERATION TOOLS AND IDENTIFY THEIR POTENTIAL USES

Common AI art generation tools were compared and evaluated from the perspective of architectural designers rather than focusing on the underlying techniques. In addition to the first wave of AI art generation tools, such as DeepDream and StyleTransfer, the analysis also included the second wave of tools, such as MidJourney, DALL-E2, and Stable Diffusion.

With rapid advancements, an increasing number of derived applications have emerged. While it is impossible to enumerate all the most advanced tools, some notable examples, such as Playground.AI and InvokeAI, are incorporated. The primary objective was to showcase various tools with distinct characteristics and operation methods, which can be employed in various architectural design contexts.

TABLE II. COMPARISON AND POTENTIAL DESIGN APPLICATIONS

	Advantages	***Limitations***	***Potential uses***
DeepDream	Intuitive, Free	Limited uses	Understand Colab environment
StyleTransfer	Intuitive, Free	Difficult to iterate	Understand Colab environment

	Advantages	*Limitations*	*Potential uses*
Midjourney	High quality	Less intuitive, 25 free quota	High quality
DALL-E 2	High-quality, intuitive UI	50 free quotas (Monthly reset to 15 credits)	High quality, Edit, expand, combine images
StableDiffusion	Intuitive, Free, with WebUI	Need basis of Google Colab	Developable
Playground .AI	Intuitive, 1000 free images/day	Couldn't expand/ combine	Image-to-image as the starting point
InvokeAI	All-rounded	Using local machine	When you have a powerful machine

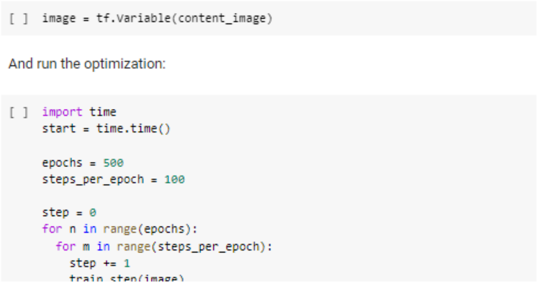
VI. CASE STUDIES

The analysis encompasses three case studies from the authors' prior investigations, comprising a university architecture module, an online workshop, and a collection of recent research efforts.

A. XJTLU Architecture Module in 2020

The first exploration was the ARC 411 Module (ARC411 Practice based enquiry and architectural representation) of Xi'an Jiaotong Liverpool University Master of Architectural Design course in 2020. There was a two-week exercise focusing on exploring AI art generation tools.

One author was the tutor, while another was among the twelve students with minimal or no programming knowledge. The case study explored the use of DeepDream and StyleTransfer through Google Colab, with files provided by a tutor from the computer science discipline. Students gained inspiration from the resulting images and translated them into architectural approaches. However, the interface proved challenging for non-programmers to comprehend.



```
[ ]  image = tf.Variable(content_image)
```

And run the optimization:

```
[ ]  import time
     start = time.time()

     epochs = 500
     steps_per_epoch = 100

     step = 0
     for n in range(epochs):
       for m in range(steps_per_epoch):
         step += 1
         train_step(image)
```

Fig. 2. An extracted screenshot of StyleTransfer on Google Colab

During the first week, students explored the use of DeepDream. They conducted numerous trials by adjusting input images and switching between various "hallucination" modes preset within the DeepDream program. However, students struggled to interpret the unpredictable results.

Eventually, students discovered that DeepDream provided the most value when applied to a top view or conceptual images, as illustrated below. Overall, this process relied heavily on students' imagination, with the 'black-box' trials offering limited assistance in the architectural design process.

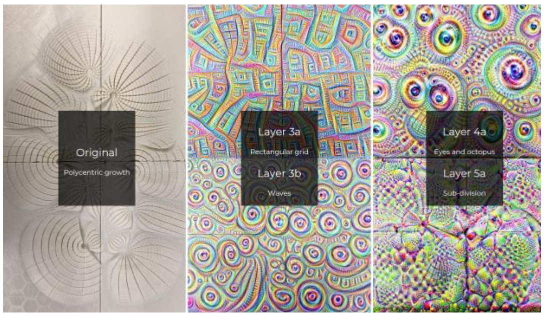


Fig. 3. Output of applying DeepDream to a physical concept model

The application of StyleTransfer in the architectural design process initially appeared intuitive for the students. However, instead of merely transferring a design style from a 'texture' image to a 'content' image, the output often appeared blurry and confusing.

Despite the underwhelming outcomes of StyleTransfer in the architectural design process, students provided both input images, unlike the default hallucination modes in DeepDream. Consequently, students attempted to modify input images to comprehend the mechanism behind StyleTransfer better. Some even tried "hacking" StyleTransfer by switching context and texture images, yielding surprising results.

To facilitate a meaningful design process using AI art generation tools, incorporating adjustable elements controlled by designers is crucial. These elements include the designer's direct input (input images) and the AI art generation tool parameters, which serve as the medium for the collaborative design process between designers and the AI program.

In the second week, students were encouraged to manually edit the output images and propose ways of application in their own design studio project.

Like the DeepDream exploration, it required a lot of students' imagination. However, since it is more controllable than DeepDream, students could provide a more convincing narrative of applying AI art generation tools to the architectural design process.

The following example demonstrates how the author combined three inspirations from StyleTransfer outputs into three distinct architectural design strategies at different scales (urban and architectural) for the studio project.

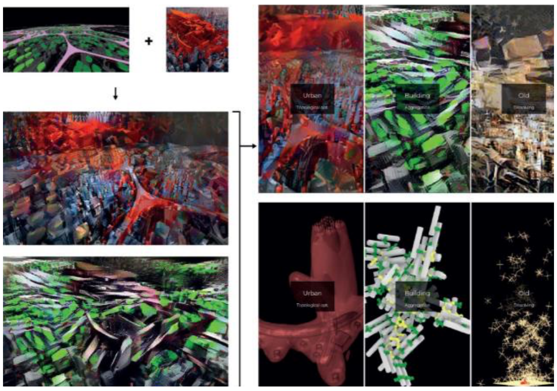


Fig. 4. Inputs, outputs of StyleTransfer and architectural strategies translations by the author

In conclusion, a user-friendly interface, the flexibility for designers' manual iterations of both input and output images, and the degree of control over AI art generation tools must be balanced to facilitate a meaningful architectural design process.

The Google Colab notebooks used above can be accessed at https://colab.research.google.com/drive/1T4RbQsmeVABxQv_Kdl4G6wS1NYPOvxFO?usp=sharing and https://colab.research.google.com/drive/1FUscVRUcJZ5_Oy8LjSWNL9N1bPO_zj0i?usp=sharing respectively.

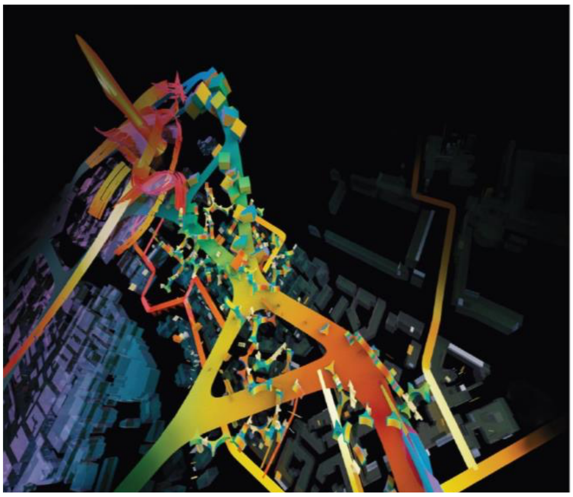


Fig. 5. The final coursework output after translating the StyleTransfer outputs above into a conceptual urban planning scheme by 3D modelling

B. *DigitalFUTURES Workshop in 2021*

The second exploration was a four-day online workshop in 2021 titled "Hacking Machine Learning Style Transfer," organized on the DigitalFUTURES platform. The two authors participated as one of the three instructors and one of the two teaching assistants. The workshop was attended by approximately twenty global architecture students with little or no programming knowledge.

Students were guided to explore StyleTransfer using a web application with a simple interface. They were then instructed to edit the outputs as collages and create 3D models. It was to enhance the intuitive controls to designers and provide a clear expected outcome as a familiar design target to the designers. On the first day, an introduction to machine learning was provided, including sharing experiences from the previous architecture module explorations. A StyleTransfer demonstration was given, and students were prompted to use a web-based StyleTransfer application featuring a clean interface with minimal adjustments (content image, texture image, and strength).

On the second day, the discussion session focused on the results of StyleTransfer. Similar to the previous exploration, many students were uncertain about how to incorporate the outputs into an architectural design process. Therefore, the conversation between students and tutors centred on strategies for combining the outputs into a target collage and the subsequent 3D translation. Additionally, one of the tutors provided an extra live session to demonstrate how a generated image could be translated into a 3D model as part of the architectural design process.

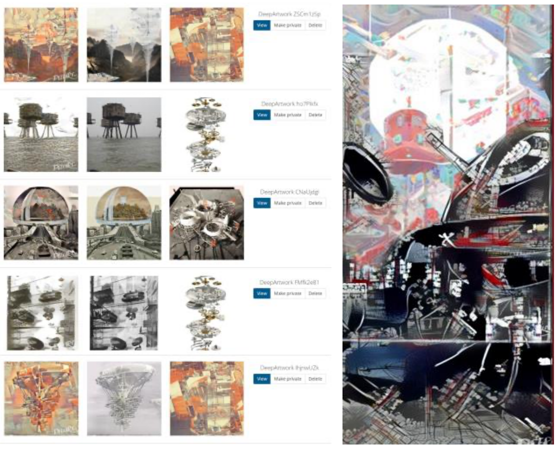


Fig. 6. StyleTransfer trials and edited collage from one of the students

On the third day, after discussing the design development of the edited collages into 3D models, students finalized their 3D models and prepared slides for the final presentations.

Fig. 7. A student 3D modelling while using StyleTransfer collage as a reference

The final presentation and discussions took place on day four. The outcomes were more promising than those of the initial exploration. They employed a more user-friendly AI art generation tool and provided a well-defined expected output within the architectural design process. However, due to the significant portion of time spent manually editing output images for collage creation and 3D modelling, the potential impact of AI art generation tools appeared to be diluted.

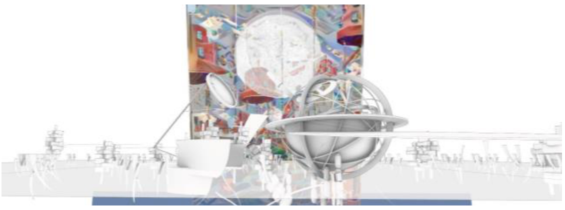


Fig. 8. A workshop outcome example of translating StyleTransfer outputs into a reference for 3D modelling

C. *Self-research in 2022, 2023*

This series of self-exploratory research sought to investigate various tools within simulated architectural design processes, building upon previous explorations and the proposed framework.

Within this framework, three different sets of combined tool usage were examined for designers with varying levels of computational knowledge and interest in developing customized AI tools.

1) *Exploration 1: Immediate use for designers with little to no machine learning knowledge.*

The first exploration targeted designers with little to no computational knowledge. The selected tools were characterized by a user-friendly interface and increased free quotas for the exploration stage. Playground.AI, which offered both text-to-image (txt2img) and image-to-image (img2img) generation functions, was essential for an intuitive design process. The platform featured adjustable parameters that designers could quickly learn through self-testing, including photo strength, prompt strength, and image quality. Consequently, designers could manipulate understandable parameters, fostering collaboration between humans and AI.

The first exploration example aimed to allow designers to iterate design and engage in dialogue with the AI art generation tool through intuitive parameter adjustments and feedback from the output images.

One screenshot was captured from each stage of the described process below. Img2img was first applied to the designer's initial sketch input to search for variations as inspirations. Then preferred option would be used for seeking further variations. The design process would then be concluded with editing functions, including overall quality control for presentation and partial generation of the iteratively-generated image.

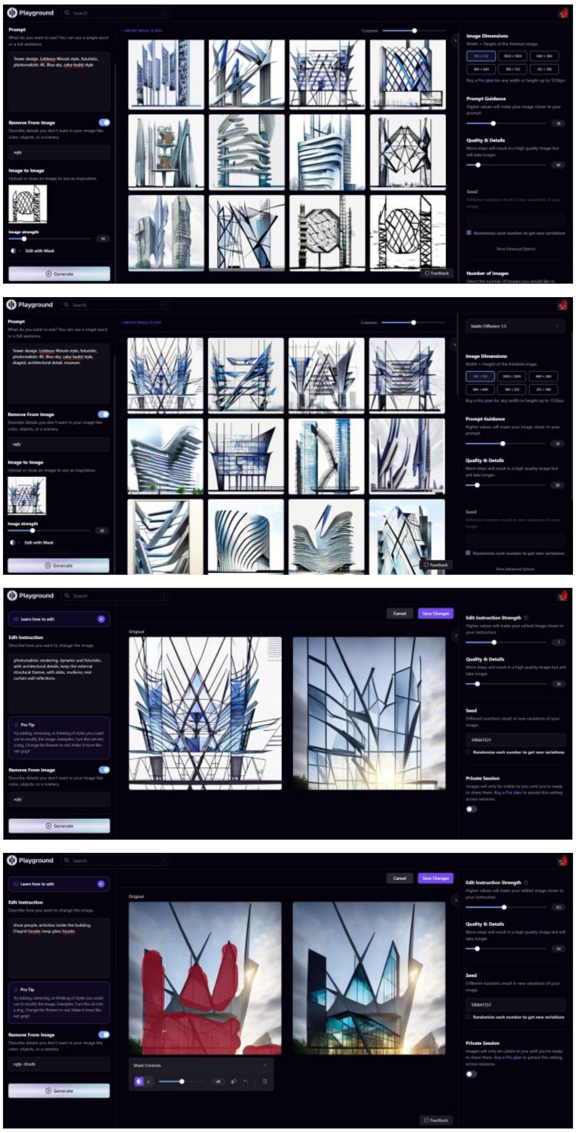


Fig. 9. Four steps from design sketch to developing an iterative design option (on Playground.AI)

2) *Exploration 2: For designers with more computational knowledge or experience in using AI art generation tools*

Exploration 2 focused on more experienced designers, utilizing Google Colab and DALL-E 2 on the OpenAI platform. In the first step, designers explored design variations on Google Colab based on their initial design sketch. Unlike the initial complex interfaces from the examples of DeepDream and StyleTransfer, a simplified code from Fast.AI's course by Jeremy Howard [24] facilitated designers' familiarity with Google Colab, allowing for further development.

In the case study, the author simulated a design scenario by inputting a self-sketched image to create design variations on the Google Colab platform. The designer experimented with different strengths (which control the degree of variations from the original image) and compared results simultaneously after understanding the parameters, adjusting and comparing all outcomes through self-revising the codes.

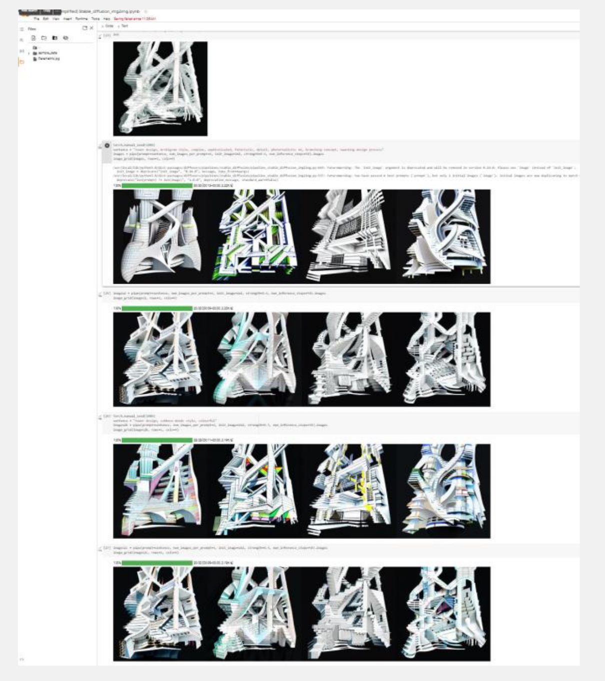


Fig. 10. From a design sketch to iterative design options on the Google Colab platform

After a few rounds of iterative generation, the output was uploaded onto DALL-E 2 for further design development. Although the OpenAI platform offered less free quota than Playground.AI, it provided more editing functions useful for more advanced design development. The example illustrated how the image produced in the previous stage could be further expanded, partially regenerated, and combined with additional images, as shown in the screenshots.

Findings indicate that some initial learning was required using developable platforms such as Google Colab, but the potential was significant due to the programmable environment. Fine-tuning could yield more promising results based on specific needs. The platform could be adaptive to emerging technology, offering more potential for customized development and adaptation in the future.

The Google Colab link used in this example can be accessed at https://colab.research.google.com/drive/1ruMnNuKizQ7Umznz_5T59ZcVlUvEyr-e?usp=sharing.

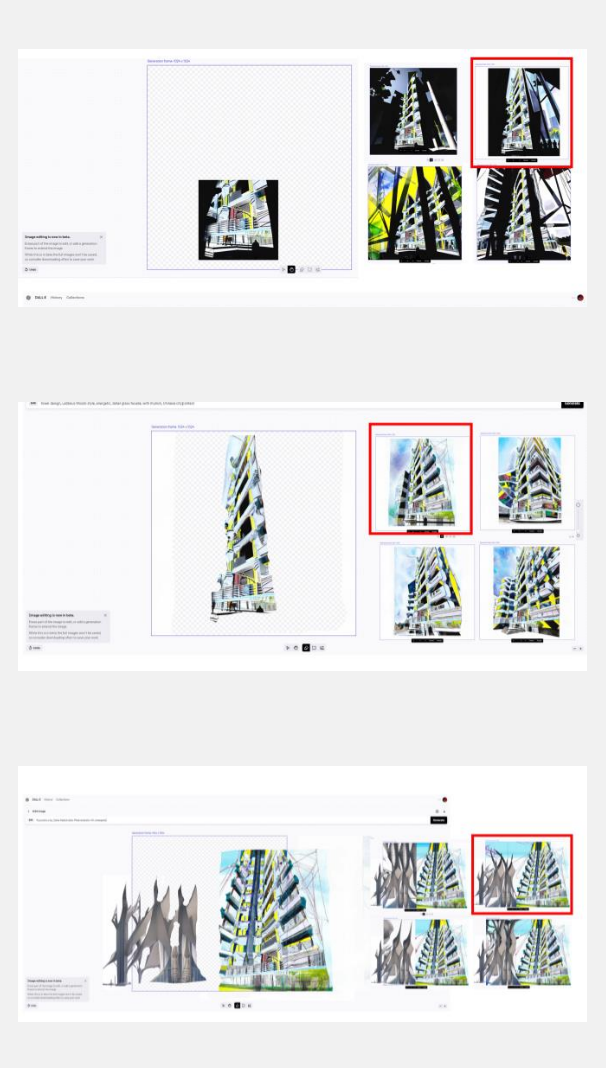


Fig. 11. Editing (expansion, partial regeneration, combination) for design development on DALL-E 2 (OpenAI platform)

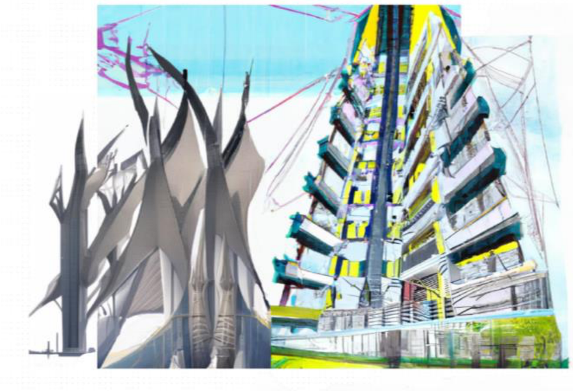


Fig. 12. The final output from the second exploration

3) Exploration 3: *For experienced designers with the intention to further customize the tool*

We further explored the more advanced use of AI art generation tools for the design development stage, specifically for experienced designers. In a simulated design scenario, a designer sought to test different colour combinations in a specific style for a conceptual diagram of an urban planning project. This required the AI art generation tool to have a deeper understanding of the intended style and recognize the input image more effectively for collaboration. Despite the complex collaboration between designers and the AI art generation tool, the attempt still aimed for immediate use for most users without a time-consuming setup or steep learning curve.

Applied the technique of Low-Rank Adaptation (LoRA) [25], designers could fine-tune stable diffusion models with only a few images. A customized model was trained within an hour through cloud computing services. Subsequently, the lightweight model was uploaded onto the Google Colab platform for immediate use. By employing the WebUI library, an intuitive interface for designers was provided. Initially, the output results did not perform as expected, where the output images were not following the 3D model structure.

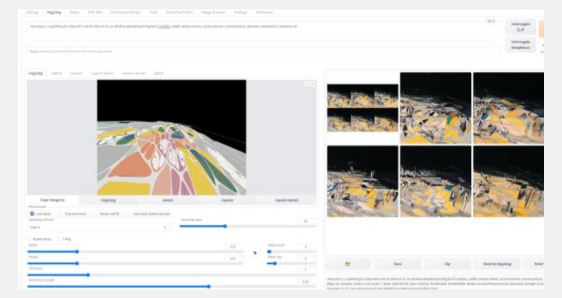


Fig. 13. Initial test of inputting a 3D model screenshot through a more user-friendly interface

Because of the flexibility of the Google Colab platform, the ControlNet plugin was installed to provide edge detection function, enabling the AI art generation tool to understand the input images based on the concept of edges, thus distinguishing the intended areas of generations. As a result, the outputs demonstrated significant improvement. This finding illustrates that a more developable AI art generation tool could be chosen for designers with more computational knowledge to perform a wider variety of design actions rather than merely increasing image quality or production speed.

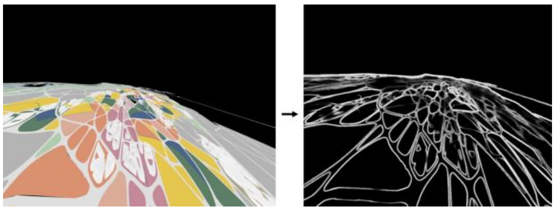


Fig. 14. Apply edge detection by ControlNet

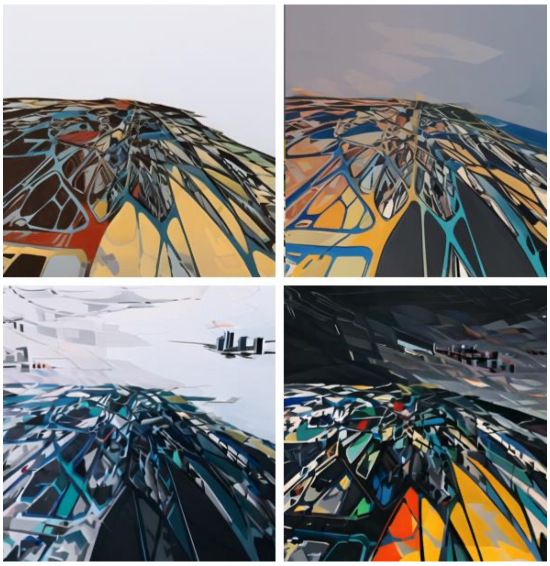


Fig. 15. Outputs of colour variations after applying edge detection and fine-tuned models

The Google Colab link used in this example can be accessed at https://drive.google.com/drive/folders/1xhlIIL3-wj4WbG94RV2KHTpkoiJDYtcf?usp=sharing.

VII. DISCUSSIONS

A. *Conclusion*

We aimed to demonstrate that, under the three principles of the proposed framework, different AI art generation tools could be utilized more effectively in various stages of the architectural design process, serving as a design partner.

TABLE III. CONCLUSIONS DESIGN APPLICATIONS

	Initial design	***Design development***	***Design development (later stage)***
Main focus	Variations	Fine adjustment of images	For presentation
Suitable tools	Tools with adjustable parameters to control the degree of variations	Tools with editing functions, such as expanding, combining, erasing and partial generation function	Tools capable of generating higher image quality Tools with a better understanding of images (such as edge detection)

B. *Limitations*

The non-intuitive switching between tools is challenge designers face, as most tools offer different advantages in producing various types of images and are rapidly evolving.

Consequently, designers must invest extra time deciding which tools are more suitable for their design process.

Another challenge is the limited understanding of architectural terms by most AI art generation tools, which are built for general purposes rather than architectural design. Although fine-tuning enables these tools to understand a specific style based on training data, introducing fundamental architectural terms could facilitate a more intuitive and in-depth conversation between architectural designers and AI art generation tools.

C. Future Research

First, in the second workshop, we attempted to translate 2D images into 3D models. Another intuitive architectural design workflow starts with physical models in addition to images. Utilizing ControlNet, the depth of an image can be detected, allowing photographs of physical models to be analyzed and AI art generation tools to create images based on these models.

Second, a streamlined application specifically for architectural designers would be beneficial. Currently, designers must switch between different applications for different design stages, and more advanced websites often provide more functions than needed. Compiling the essential functions for architectural designers into a single user interface would enhance immediate intuitive application.

Third, future design sessions should be organized for architecture students, allowing them to gain meaningful experience using AI art generation tools as part of their design process. Collecting feedback from these experiences could contribute to further development in this field.

ACKNOWLEDGEMENT

This research is supported by XJTLU Postgraduate Research Scholarship (PGRSB211206) offered by Xi'an Jiaotong-Liverpool University.

REFERENCES

[1] I.J. Goodfellow, *et al*., 'Generative Adversarial Nets', presented at the NIPS, Dec. 2014. Accessed: Mar. 04, 2023. [Online]. Available: https://www.semanticscholar.org/paper/Generative-Adversarial-Nets-Goodfellow-Pouget-Abadie/54e325aee6b2d476bbbb88615ac15e251c6e8214.

[2] A. N. Wu, R. Stouffs, and F. Biljecki, 'Generative Adversarial Networks in the built environment: A comprehensive review of the application of GANs across data types and scales', *Building and Environment*, vol. 223, Sep. 2022, doi: 10.1016/j.buildenv.2022.109477.

[3] M. L. Castro Pena, A. Carballal, N. Rodríguez-Fernández, I. Santos, and J. Romero, 'Artificial intelligence applied to conceptual design. A review of its use in architecture', *Automation in Construction*, vol. 124, p. 103550, Apr. 2021, doi: 10.1016/j.autcon.2021.103550.

[4] A. Mordvintsev, C. Olah, and M. Tyka, "Inceptionism: Going Deeper into Neural Networks", *Google Research*. 18-June-2015. [Online]. Available: https://ai.googleblog.com/2015/06/inceptionism-going-deeper-into-neural.html. [Accessed: 4-Mar.-2023]..

[5] L. A. Gatys, A. S. Ecker, and M. Bethge, 'A Neural Algorithm of Artistic Style'. arXiv, Sep. 02, 2015. doi: 10.48550/arXiv.1508.06576.

[6] M. Radhakrishnan, 'Is Midjourney-Ai the New Anti-Hero of Architectural Imagery & Creativity?', vol. 11, pp. 94–114, Jan. 2023, doi: 10.11216/gsj.2023.01.102270.

[7] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, 'High-Resolution Image Synthesis with Latent Diffusion Models'. arXiv, Apr. 13, 2022. doi: 10.48550/arXiv.2112.10752..

[8] MidJourney [@MidJourney], _"We're officially moving to open-beta! Join now at https://discord.gg/midjourney. **Please read our directions carefully** or check out our detailed how-to guides here: https://midjourney.gitbook.io/docs. Most importantly, have fun!"_, Twitter. 13-July-2022. [Tweet]. Available: https://twitter.com/midjourney/status/1547108864788553729. [Accessed: 4-Mar.-2023].

[9] A. Ramesh, P. Dhariwal, A. Nichol, C. Chu, and M. Chen, 'Hierarchical Text-Conditional Image Generation with CLIP Latents'. arXiv, Apr. 12, 2022. doi: 10.48550/arXiv.2204.06125.

[10] DigitalFUTURES World. *DigitalFUTURES / Legends : John Frazer*. (June 29, 2022). Accessed: Mar. 4, 2023. [Online Video]. Available: https://www.youtube.com/watch?v=_GE2PsI2Tdw.

[11] DigitalFUTURES World. *Doctoral Consortium: Creative AI in Architecture--Daniel Bolojan*. (June 13, 2022). Accessed: Mar. 4, 2023. [Online Video]. Available: https://www.youtube.com/watch?v=A8r0Q_tbmTU.

[12] E. Yildirim, 'Text-to-Image Generation AI in Architecture', in *Art and Architecture: Theory, Practice and Experience,* 1st Ed. Livre de Lyon, 2022, pp. 97–120.

[13] J. Cudzik and K. Radziszewski, 'Artificial Intelligence Aided Architectural Design', *MOST Wiedzy - portal z wiedzą dla Ciebie*, 2018. https://mostwiedzy.pl/pl/publication/artificial-intelligence-aided-architectural-design,146441-1 (accessed Mar. 02, 2023).

[14] T. Stojanovski et al., 'Rethinking Computer-Aided Architectural Design (CAAD) – From Generative Algorithms and Architectural Intelligence to Environmental Design and Ambient Intelligence', in *Computer-Aided Architectural Design. Design Imperatives: The Future is Now*, Singapore, 2022, pp. 62–83. doi: 10.1007/978-981-19-1280-1_5.

[15] J. Vermillion, 'Iterating the Design Process Using AI Diffusion Models', *Creative Collaborations*, Oct. 2022, [Online]. Available: https://digitalscholarship.unlv.edu/cfa_collaborate/9.

[16] J. Ploennigs and M. Berger, 'AI Art in Architecture'. arXiv, Dec. 19, 2022. doi: 10.48550/arXiv.2212.09399.

[17] Enjellina, E. V. P. Beyan, and A. G. C. Rossy, 'A Review of AI Image Generator: Influences, Challenges, and Future Prospects for Architectural Field', *Journal of Artificial Intelligence in Architecture*, vol. 2, no. 1, Art. no. 1, Feb. 2023, doi: 10.24002/jarina.v2i1.6662.

[18] E. Ulberg, D. C. Llach, and D. Byrne, 'Hand-Crafting Neural Networks for Art-Making', In *Proceedings of the 11th International Conference on Computational Creativity*, 508-511. 2020.

[19] E. Ulberg, 'Crafting the Weights of a Convolutional Neural Network to Make a Drawing', thesis, Carnegie Mellon University, 2021. doi: 10.1184/R1/14135663.v1.

[20] H. W. J. Rittel and M. M. Webber, 'Dilemmas in a general theory of planning', *Policy Sci*, vol. 4, no. 2, pp. 155–169, Jun. 1973, doi: 10.1007/BF01405730.

[21] R. Glanville, 'Try again. Fail again. Fail better: the cybernetics in design and the design in cybernetics', *Kybernetes*, vol. 36, no. 9/10, pp. 1173–1206, Jan. 2007, doi: 10.1108/03684920710827238.

[22] R. Glanville, 'A (Cybernetic) Musing: Certain Propositions Concerning Prepositions', *Cybernetics And Human Knowing*, vol. 12, no. 3, pp.87-95, Jan. 2005, Accessed: Apr. 02, 2023. [Online]. Available: https://www.academia.edu/995476/A_Cybernetic_Musing_Certain_Propositions_Concerning_Prepositions

[23] B. Sweeting, 'Design Research as a Variety of Second-Order Cybernetic Practice', in *Series on Knots and Everything*, 2017, pp. 227–238. doi: 10.1142/9789813226265_0035.

[24] Jeremy Howard. *Lesson 9: Deep Learning Foundations to Stable Diffusion*. (Oct. 20, 2022). Accessed: Apr. 2, 2023. [Online Video]. Available: https://www.youtube.com/watch?v=_7rMfsA24Ls

[25] E. J. Hu et al., 'LoRA: Low-Rank Adaptation of Large Language Models'. arXiv, Oct. 16, 2021. doi: 10.48550/arXiv.2106.09685.