DIGITAL 2022

Advances on Societal Digital Transformation

ISBN: 978-1-68558-014-8

November 13 - 17, 2022

Valencia, Spain

DIGITAL 2022 Editors

Atriya Sen, University of New Orleans, USA

DIGITAL 2022

Forward

Advances on Societal Digital Transformation (DIGITAL 2022), held between November 13 and November 17, 2022, continues a series of international events covering a large spectrum of topics related to the digital transformation of our society.

The society is continuously changing with a rapid pace under digital transformation. Taking advantage of a solid transformation of digital communications and infrastructures and with great progress in AI (Artificial Intelligence), IoT (Internet of Thinks), ML (Machine Learning), Deep Learning, Big Data, Knowledge acquisition and Cognitive technologies, almost all societal areas were redefined. Transportation, Buildings, Factories, and Agriculture are now a combination of traditional and advanced technological features. Digital citizen-centric services, including health, well-being, community participation, learning and culture are now well-established and set to advance further on. As counter-effects of digital transformation, notably fake news, digital identity risks and the digital divide are also progressing in a dangerous rhythm, there is a major need for digital education, fake news awareness, and legal aspects mitigating sensitive cases.

We take here the opportunity to warmly thank all the members of the DIGITAL 2022 technical program committee, as well as all the reviewers. The creation of such a high-quality conference program would not have been possible without their involvement. We also kindly thank all the authors who dedicated much of their time and effort to contribute to DIGITAL 2022. We truly believe that, thanks to all these efforts, the final conference program consisted of top-quality contributions. We also thank the members of the DIGITAL 2022 organizing committee for their help in handling the logistics of this event.

We are convinced that the participants found the event useful and communications very open. We hope that Valencia provided a pleasant environment during the conference and everyone saved some time to enjoy the charm of the city.

DIGITAL 2022 Chairs

DIGITAL 2022 Steering Committee
Adel Aneiba, Birmingham City University, UK
Fernando Joaquim Lopes Moreira, Universidade Portucalense, Portugal
Yunpeng (Jack) Zhang, University of Houston, USA
Wanwan Li, University of South Florida, USA

DIGITL 2022 Publicity Chairs
Mar Parra, Universitat Politecnica de Valencia, Spain
Sandra Viciano Tudela, Universitat Politecnica de Valencia, Spain

DIGITAL 2022

COMMITTEE

DIGITAL 2022 Steering Committee

Adel Aneiba, Department of Networks and Cybersecurity (NCS), School of Computing and Digital Technology Birmingham City University, UK
Fernando Joaquim Lopes Moreira, Universidade Portucalense, Portugal
Yunpeng (Jack) Zhang, University of Houston, USA
Wanwan Li, University of South Florida, USA

DIGITAL 2022 Publicity Chair

Mar Parra, Universitat Politecnica de Valencia, Spain
Sandra Viciano Tudela, Universitat Politecnica de Valencia, Spain

DIGITAL 2022 Technical Program Committee

Qammer Abbasi, University of Glasgow, UK
Aaroud Abdessadek, Chouaib Doukkali University, Morocco
Kawiwat Amnatchotiphan, Thai-Nichi Institute of Technology, Thailand
Daniel Amo Filvà, La Salle, Ramon Llull University, Spain
Mariia Andriyiivna Nazarkevych, Lviv Polytechnic National University, Ukraine
Adel Aneiba, Birmingham City University, UK
Sakthi Balan Muthiah, The LNM Institute of Information Technology, Jaipur, India
Mohamed Basel Almourad, Zayed University, United Arab Emirates
Louiza Bouallouche-Medjkoune, University of Bejaia, Algeria
An Braeken, Vrije Universiteit Brussel, Belgium
Roberta Calegari, University of Bologna, Italy
Jundong Chen, Dickinson State University, USA
Siming Chen, Fudan University, China
Marta Chinnici, ENEA, Rome, Italy
Mohamed Dahmane, CRIM – Computer Research Institute of Montreal, Canada
Babu R. Dawadi, Tribhuvan University, Nepal
Burcu Demirdöven, Pamukkale University, Turkey
Zakaria Abou El Houda, University of Montreal, Canada
Nikos Fakotakis, University of Patras, Greece
Fadi Farha, University of Science and Technology Beijing, China / Aleppo University, Syria
Nasim Ferdosian, Cergy Paris University, France
Allan Fowler, University of Auckland, New Zealand
Ali Mohsen Frihida, University of Tunis El Manar, Tunisia
Andrea Gentili, Telematic University eCampus, Italy
Jing Gong, Uppsala University, Sweden
Teresa Guarda, CIST Research and Innovation Center | UPSE, Ecuador / ALGORITMI Research Centre of Minho University, Portugal
A S M Touhidul Hasan, University of Asia Pacific, Dhaka, Bangladesh

Copyright Information

For your reference, this is the text governing the copyright release for material published by IARIA.

The copyright release is a transfer of publication rights, which allows IARIA and its partners to drive the dissemination of the published material. This allows IARIA to give articles increased visibility via distribution, inclusion in libraries, and arrangements for submission to indexes.

I, the undersigned, declare that the article is original, and that I represent the authors of this article in the copyright release matters. If this work has been done as work-for-hire, I have obtained all necessary clearances to execute a copyright release. I hereby irrevocably transfer exclusive copyright for this material to IARIA. I give IARIA permission or reproduce the work in any media format such as, but not limited to, print, digital, or electronic. I give IARIA permission to distribute the materials without restriction to any institutions or individuals. I give IARIA permission to submit the work for inclusion in article repositories as IARIA sees fit.

I, the undersigned, declare that to the best of my knowledge, the article is does not contain libelous or otherwise unlawful contents or invading the right of privacy or infringing on a proprietary right.

Following the copyright release, any circulated version of the article must bear the copyright notice and any header and footer information that IARIA applies to the published article.

IARIA grants royalty-free permission to the authors to disseminate the work, under the above provisions, for any academic, commercial, or industrial use. IARIA grants royalty-free permission to any individuals or institutions to make the article available electronically, online, or in print.

IARIA acknowledges that rights to any algorithm, process, procedure, apparatus, or articles of manufacture remain with the authors and their employers.

I, the undersigned, understand that IARIA will not be liable, in contract, tort (including, without limitation, negligence), pre-contract or other representations (other than fraudulent misrepresentations) or otherwise in connection with the publication of my work.

Exception to the above is made for work-for-hire performed while employed by the government. In that case, copyright to the material remains with the said government. The rightful owners (authors and government entity) grant unlimited and unrestricted permission to IARIA, IARIA's contractors, and IARIA's partners to further distribute the work.

Table of Contents

War-Gaming Needs Argument-Justified AI More Than Explainable AI

John Licato
Department of Computer Science and Engineering
Advancing Machine and Human Reasoning (AMHR) Lab
University of South Florida
Tampa, FL, USA
licato@usf.edu

***Abstract*—I argue that a planning agent in a societal- or war-gaming environment, whether that agent is a sole AI or part of a human-AI team, should behave in a way that is more than just *explainable*. Rather, its actions should be *argument-justified*; i.e., it must produce as justification of its actions the equivalent of an argument graph demonstrating how its choice is superior to, *and fairly considers*, the strongest possible arguments for a sufficient number of alternative choices. Although argument-justified AI might be considered a subset of interpretable AI, the requirement that a qualified argument graph be part of the model's output imparts multiple desirable properties over alternatives, namely: trustworthiness, understandability, persuasiveness, thoroughness, and others.**

***Index Terms*—AI, justification, reasoning, argumentation, war gaming, decision-making, explainable AI**

I. Introduction

Complex environments necessitate complex rules; this is true particularly when the range of choices an agent has available to them at any given moment is large (or infinite), and the range of possible consequences of those actions is also large (or infinite). As anyone who has spent time designing or playing a sophisticated war-game knows, making the game increasingly realistic (and thus more useful as a simulation environment for training both human and AI actors) requires game rules and mechanisms of a complexity that can quickly rival that of a full-fledged legal system. And real-world legal systems unavoidably contain *open-textured terms* [1, 2], terms denoting concepts whose boundaries are virtually impossible to fully formalize, whose applicability must be determined dynamically through the use of interpretive reasoning [3, 4, 5, 6].

We have previously argued for the importance of interpretation-capable reasoning in AI, particularly when that AI must act in accordance with human-created rules such as laws, ethical codes, rules of engagement, and so on [3, 4]. According to what we have called the *MDIA position*, Rule-following AI should act in accordance with the interpretation best supported by Minimally Defeasible Interpretive Arguments (MDIA) [4]. In this paper, I discuss the need for interpretation-capable reasoning in war-gaming. In short, I argue that a planing agent—whether AI or human-AI hybrid—in a societal- or war-game must be argument-justified; i.e., it must produce a justification of its conclusions which is the equivalent of an argument graph demonstrating how its final course of action is superior to, and properly considers, the strongest possible arguments for all alternative plausible actions. I first discuss why war-gaming is a domain in which interpretive reasoning is particularly important (I-A), and introduce minimal defeasibility (I-B). I then introduce argument-justified AI and argue for its benefits over merely explainable AI (II), and close by anticipating objections (III).

A. Interpretive Reasoning in War-Gaming

"War-gaming" encompasses a range of games that is so broad, it can be futile to make sweeping claims that apply equally to all of them. In this paper, I focus instead on the fuzzy subset of war-games that is typically played on a board between teams of human or human/AI players, which serves as "a dynamic representation of conflict or competition in which people [or artificial agents] make decisions and respond to the consequences of those decisions". Here we borrow a definition from the NPS (https://nps.edu/web/wargaming-activity-hub/what-is-wargaming-). This class of games includes popular games with multiple paths to victory and agreements between players (such as the Civilization© series of computer games), as well as what might be considered more "serious" board games with instruction manuals complex enough to fill entire books (such as the GMT Next War© game series).

In such war-games, interpretive reasoning can be so prevalent as to occur unnoticed by players. But getting it wrong can be disastrous. Consider, for example, a game in which two players, a and b, make an agreement that because player c is so far ahead of both of them, a and b will observe a non-aggression pact with each other until c is eliminated, and the first to violate this peace must pay a large financial penalty to the other player. As such, they refuse to attack each other for a few turns, but then a decides to block the trade routes surrounding b's territory and refuses to trade anything with b whatsoever. b considers these actions by a to constitute aggressions in violation of their agreement. But should such economic actions really be interpreted as violations of peace, particularly in the sense of the open-textured term "peace" in the agreement between a and b?

Clearly, this disagreement hinges a question of interpretation. They enlist a neutral fourth player, d, to settle their dispute. In doing so, they will both need to argue to convince d

 ISBN: 978-1-68558-014-8

that the terms of their agreement, prior precedents, reasonable assumptions, and so on support their claims. And it is exactly interpretive argumentation that will allow them to do so, and it is interpretive reasoning that will allow d to compare those arguments against each other and decide which case should prevail.

Likewise, it is easy to imagine scenarios in which disagreements about how official rules of the game are to be interpreted must be resolved by the players. Such disagreements are common with complex war-games which introduce terminology that may draw on real-world phrases whose applicability to game actions is not immediately clear. For example, the collectible card game Battlespace Next™: Multi-Domain Operations (https://www.printplaygames.com/product/battlespace-next/) has rules disallowing "kinetic attacks" under certain conditions, but it may not be immediately clear to non-military players what exactly constitutes a kinetic attack. Disagreements about whether an action consisting of a single person physically breaking and entering a secured facility constitutes a kinetic attack, again, will need to be settled using interpretive reasoning. And if an artificial agent is asked to adjudicate such disagreements, a simple output declaring who the winner is and with what confidence is not going to be very satisfying to the disputants. On the other hand, were the adjudicating AI to output a full argument graph demonstrating exactly how all of the arguments presented factor into the final consideration (as in Figure 1b), the final conclusion may be more palatable to all—at the very least, it allows for the arguers to see whether there are points in which their arguments were misunderstood or misrepresented.

B. Minimal defeasibility

Often, the boundaries between argument text and argument are blurred. For instance, in a dialogical debate, one participant might say "cats are funny because they make me laugh," and a second participant might attack this by saying "That conclusion is not warranted; I know of at least one cat that isn't funny." The first might reply, "I did not mean that *all* cats are funny. I meant that there are at least some cats which make me laugh, therefore some cats are funny." In this admittedly silly example, the two participants are mistaken about the proper interpretation of the text "cats are funny"—does the text denote a claim that is quantified universally (all cats are funny) or existentially (some cats are funny)?

Dialogical debates will often proceed in this way. A claim is made by one participant, which is then met by rebuttals, counterarguments, or clarification requests. The first participant may adjust the argument text in order to better match their intended argument, or they may adjust the intended argument itself, or some blurred combination of the two. That adjustment may open them up to further attacks, in response to which the participant will either defuse the attacks or further adjust their argument and argument text. This iterative process might continue until the participants are satisfied with the strength of their respective arguments (or, in practice, such discussions are more often terminated because of a subject shift, time constraint, or an exhaustion of patience). And in an ideal dialogical debate, each iteration of this process results in arguments that are less *defeasible*—less subject to attacks, less need for clarification, fewer weak points, and a more robust ability to both pre-empt and defend against possible counter-arguments and other argumentative attacks. ('Defeasibility' as a term is often credited either to Chisholm [7] or Hart [8], but was perhaps made most famous by Pollock [9, 10].) The goal of the iterative process we describe here, then, is to achieve a state of *minimal defeasibility* for arguments: a state in which a minimal amount and quality of possible attacks can be levied against it.

In real-world argumentation, the vast majority of arguments are defeasible—they are always subject to possible counter-attacks. That is why minimal defeasibility must be a goal direction, but should not be considered something that can ever practically be reached. At some point, limitations of time, computing power, or available information will restrict iterative improvement of an argument's defeasibility. It should also be noticed that the way in which I have defined minimal defeasibility here means that it will not do as a general definition of argument strength, merely because the definition itself relies on the concept of argument strength. My intent here is for minimal defeasibility to serve as a way of conceptualizing the high-level search strategy that I believe can lead to the generation and evaluation of high-quality *interpretive* arguments.

In order to become minimally defeasible, an argument must be able to anticipate what sort of attacks might be levied against it. But in order to be sure that we have successfully considered the best arguments from all possible sides, we need to understand what kinds of processes generate the best arguments from each side; after all, considering only strawman counterarguments is not a productive strategy that will lead to minimal defeasibility. Fortunately, we can draw from the examples in human domains that deal with the presentation and evaluation of argumentative exchanges. For example, many processes in legal settings employ some variant of an adversarial approach, in which representatives from each side of an issue put forth the strongest arguments they can come up with.

The paradigm example of the adversarial approach is a court trial, where opposing counsel argue their respective cases before an (ideally) impartial judge and/or jury. In the ideal case, the judge or jury thoroughly considers the strongest arguments presented on each side and produces a decision that takes all of them into account. This leads to a division of labor, in which the representatives of each side only need to focus on producing the most impactful arguments for their respective side, and the strongest counterarguments for those of the opposition. Indeed, it seems to be a feature of human reasoning that we excel at producing arguments for one side at a time (typically the side we already agree with), but struggle when forced to generate or evaluate arguments from multiple perspectives. Manifestations of this phenomenon go by many names: confirmation bias, myside bias, and so on. And in

 ISBN: 978-1-68558-014-8

both individual reasoning and large-scale debates, this one-sidedness can be highly problematic, even for medical doctors [11, 12, 13, 14] or judges [15, 16, 17, 18]. Mercier and Sperber argue that this one-sidedness is a *feature, not a bug*; human reasoning evolved to work best in small groups where opposing arguers attack, and are forced to defend against, each other. According to their *argumentative theory of reasoning*, limitations such as the myside bias are due to the human reasoning capability being taken out of its natural social context (for which it evolved), and used individually where it is less suited to flourish [19, 20, 21]. Because of the myside bias, people are motivated to defend views they have, even if the best arguments they can come up with to defend such views are weak and fallacious (i.e., have high defeasibility). Indeed, growing evidence shows that the iterative dialogue approach, in which reasoning and argument development are carried out in a dialogical, argumentative form between small groups, tends to work better than individual reasoning particularly because it encourages the development of arguments to be increasingly resistant against possible attacks [20, 22, 23, 24, 25, 26, 27, 28, 29]. In other words, it works *because it strives for minimal defeasibility.*

To be sure, the adversarial approach itself has its limitations. E.g., when one side has access to more expensive legal representation, the quality of argumentation put forth by both sides may be uneven. But these are problems of implementation, not necessarily problems with the idea that if multiple sides are given the resources to properly put forward the strongest possible arguments for their side, then the resulting synthesis of arguments is better overall. And so for our current question of interest—how interpretation-capable AI might best generate and evaluate interpretive arguments—something resembling an adversarial approach is the way to go.

II. Argument-Justified AI (AJAI)

Much current work in representing computational argumentation can be traced to [31], in which an abstract argumentation framework is introduced as a tuple consisting of a set of arguments A, and a set of attacks which is a subset of $A \times A$. Simple as this definition may be, Dung was able to then define a series of semantics for individual arguments and argumentation frameworks: they could be stable, conflict-free, acceptable, admissible, etc. Let us say that an attacking argument a_1 successfully attacks (when successful, we say it *defeats*) another argument a_2. According to the ASPIC framework [32], defeating attacks fall into one of three types: a rebutting attack directly contradicts the conclusion of a_2; an undermining attack contradicts one of the premises of a_2, and an undercutting attack attacks the inference step that directly connects the premises of a_2 to its conclusion. One way of visualizing an argument being attacked in all three of these ways is contained in Figure 1b.

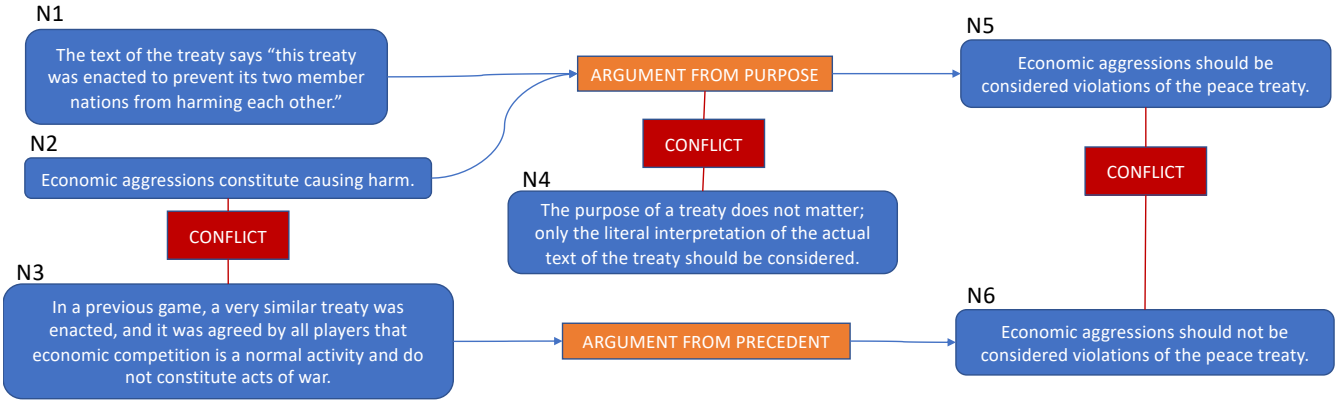
Dung's framework spawned a variety of approaches that extended it, most based on the argument interchange format [33]. Today, the amount of available tools for representing argumentation graphs is continuing to expand (see the review in [34]), particularly with the rate at which progress in natural language processing is accelerating. Because there is a wealth of options for visualizing networks of interpretive arguments and counterarguments each with its own pros and cons (for overviews, see [35, 36, 37]), I will not commit to any particular implementation here. But observe the differences between the argument graphs presented in Figures 1a and 1b; the first presents a single argument which may seem strong at first glance, whereas the second not only shows possible attacking arguments, but *how* those attackers relate to the original argument. The weights used to compare these arguments and determine whether they are defeating or merely just attacking, which might be obtained for example from a public vote or decision by experts, can be visualized as well. And thus, the precise way in which all arguments factor into the final conclusion can be made fully transparent.

A. AJAI vs. XAI

It is difficult to understate the value of the type of transparency afforded by argument graphs which contain the strongest possible arguments and counterarguments for competing positions. People who sympathize more with the counterarguments to the winning position will be more likely to be persuaded if they see how their arguments are fairly considered and factor into the final calculation (as compared to simply being told that their arguments were considered without explaining how, a favorite trick of high-level decision-makers in large organizations). On the other hand, if the argument graph consisting only of the arguments in support of the final decision is provided by the decision-maker, it may be subject to manipulation and rhetorical tricks: imagine, for example, a deceptive politician presenting their position in a way that unfairly dismisses potentially strong counterarguments. Furthermore, the properties of an argument graph may change over time. The weights that are used to determine whether one argument should be considered stronger than another, or the full set of facts and available evidence, might change over time. It will not be clear how those changes affect established conclusions if we do not preserve and present the entire graph.

Let us consider the merits and demerits of argument-justified AI as opposed to its alternatives. *Explainable AI* (XAI) comes first to mind, as it is an active area of research in machine learning. XAI work takes the outputs of black-box systems and produces explanations for them. Although there are some overlaps between explanations and arguments, and the two can productively be used in combination with each other [38], there is a fundamental difference: *explanations help people understand how an output was generated, while arguments persuade people that an output should be accepted.* It is not always clear what is meant by the word 'explainable' in XAI [39], and depending on how one defines it, what we have called AJAI may be considered a proper subset of XAI. I will use the broad sense of the word 'explainable' so that an AI is explainable if it is able to provide a human-understandable explanation of its decisions. An AJAI which provides a full argument graph along with the strongest counterarguments

(a) A Single Argument



(b) Multiple Arguments

Fig. 1: Example argument graphs containing a single argument for one position (a) and a network of conflicting claims and arguments for two positions (b). Visualizations here are loosely based on OVA+ [30].

to each of its decisions is therefore a type of XAI (as in Figure 1b), but so also is an AI which merely presents the reasons to accept its preferred conclusion without stating the counterarguments (as in Figure 1a).

Let us assume that in the future, someone comes up with a purely statistical deep neural network for war-games where all we have to do is feed as inputs: the rules to be followed, a description of a game scenario to be interpreted, and some minimal set of contextual details so that the system can infer things like intents of the rule-makers, historical interpretations of the rules, etc. Assume further that this system is an almost impermeable black box, and its outputs are explainable, but not argument-justified. Instead, this system (let's call it $\mathcal{O}$ for 'oracle') simply outputs the optimal interpretation; i.e., the interpretation that would have come about if the best possible interpretive arguments of all types were generated and combined in an optimal way. On the other hand, another system $\mathcal{A}$ is an argument-justified AI which outputs the optimal interpretation along with an argument graph that relates the strongest arguments for the optimal interpretation to its strongest counterarguments. $\mathcal{O}$ clearly provides a more concise output, and it may even output a percentage that might be understood as a measure of its confidence in its conclusion. Let us assume, for the sake of simplicity, that if $\mathcal{O}$ outputs an interpretation and a confidence of 50% or higher, then the interpretation is "recommended." Now ask yourself: If $\mathcal{O}$ were to exist today, and it produced the same conclusions as the argument-justified, interpretation-capable system $\mathcal{A}$, would $\mathcal{O}$ be preferable to $\mathcal{A}$?

I argue that $\mathcal{O}$ would *not* be preferable to an equivalent argument-justified, interpretation-capable system. To be clear, I would *not* argue that the creation of $\mathcal{O}$ is impossible. It is conceivable that in the future a massive, well-designed artificial neural network could exactly simulate the brains of the 15 greatest Supreme Court justices who ever lived, simulate a lengthy and productive debate between them, and then run iteratively until a conclusion is reached. Presumably, such a system (or another similar brute force approach) would come as close as any other decision-making algorithm to coming up with the "correct" interpretation in the largest number of cases. But what I do doubt is that any approach to designing $\mathcal{O}$ can do so without, at some stage of its deliberations, internally generating and evaluating interpretive arguments. If $\mathcal{O}$ were able to generate an interpretation without carrying out any of these steps, then in all likelihood, it has failed to consider some crucial argument or counter-argument, and will therefore be suboptimal as compared to $\mathcal{A}$ (in the sense that will not come up with the most correct interpretations). On the other hand, if $\mathcal{O}$ internally generates and evaluates interpretive arguments just like $\mathcal{A}$ would as part of its reasoning process, then it is difficult to see why it should not simply provide the optimal interpretive arguments, along with the reasoning behind their combination it evaluated internally as part of its output—but if it did so, it would make it an AJAI anyway!

Even if I am wrong about my claims in the previous paragraph, $\mathcal{O}$ would still not be preferable to $\mathcal{A}$, for several reasons. First, interpretations of open-textured rules must be subject to stakeholder analysis and approval. The interpretive argument paradigm provides a rich tapestry of justification types, and it is easy to see why interpretations that are justified

 ISBN: 978-1-68558-014-8

with clearly laid-out interpretive arguments is preferable to a simple black-box output. Even if $\mathcal{O}$ were the most powerful pattern recognizer in existence, trained on the largest data set possible, if $\mathcal{O}$ is unable to argue *why* we should accept its outputs, it will fail to persuade stakeholders. Further, there is a sense in which the correct answer to certain interpretive scenarios does not even exist until the stakeholders consider arguments for an interpretation. For example, the United States Supreme Court is not a legislative body, but when they decide on an interpretation of some open-textured term in a law, that interpretation is binding upon lower courts and also the Supreme Court itself, according to the principle of *stare decisis*. Therefore, when interpreting law, is the Supreme Court merely *discovering* correct interpretations that were always true, or are they *creating* the correct interpretation through an interaction of values, viewpoints, and arguments? For our purposes, it will suffice to say that it is likely some combination of these two (I elaborate more on this idea in [4]). And likewise, when $\mathcal{A}$ encounters new scenarios that were not anticipated by its programmers, it does not apply a discovery algorithm to find the optimal interpretation that exists independently of the values of the stakeholders that will evaluate it. Rather, a complex medley of inference, value-laden judgments, and argument evaluation interact, together *shaping* the interpretation that is ultimately accepted as the correct one. An interpretive framework which therefore fails to take any of these elements into account will be sub-optimal.

Yet another reason to prefer AJAI relates to the need for even application of rules. That rules should be applied equally to all is a principle pervasive in virtually all modern legal and ethical systems. If the same rule is assigned two different interpretations for two different target cases, the reasons for this difference must be made clear in such a way that they create a guide for future cases. Imagine that $\mathcal{O}$ was tasked with controlling who can enter a private park area, and told to follow the rule "no vehicles allowed in the park." One day, it decides to grant an exception to a group of senior citizens on motorized scooters without providing substantial interpretive argumentation to support this decision (internally, the reason it decided to do so was because its internal statistical algorithm estimated a 50.01% confidence that an exception was warranted). But then the next day, a different motorized scooter group consisting of teenagers arrives and decides to host an impromptu picnic, this time for a charitable cause. Is $\mathcal{O}$ required to grant their request? If not, why not? And how can such questions be answered in the first place, in the absence of interpretive arguments? If the second group's request were to be denied (for example, perhaps $\mathcal{O}$'s internal algorithm only had a 49.99% confidence that an exception was warranted), can we really say that $\mathcal{O}$'s judgements constitute a fair application of the rules across the two scooter groups?

$\mathcal{A}$, on the other hand, may be able and required to explain precisely how the second group should be awarded an exception *on the basis of the network of interpretive arguments used to support its decision on the first group*. For example, it may be that the first group was granted an exception primarily because providing recreational spaces to senior citizens is a good, ethical thing to support. Thus, since the second group is supporting a charity (also presumably a good, ethical thing to support), the exception should also apply. On the other hand, if the most influential reason for the first group's exception was that their proposed event was rare and the citizens of the park would not mind a single day's worth of noise, then rejecting the second group might be warranted. Either way, if $\mathcal{O}$ or $\mathcal{A}$ are to apply laws equally and fairly across multiple circumstances, they must be able to demonstrate why interpretations across multiple borderline cases are consistent—and this can best be done with explicit rationales of the sort made possible with full argument graphs.

III. Anticipated Objections

The MDIA position advocated for in this paper rests on the assumption that the overall quality of argumentative conclusions is improved when potential counterarguments are addressed. In this spirit, I conclude by addressing possible objections.

a) Why not advocate fully formalizing the law instead? Won't this remove the need for open-textured predicates?: Simply stated, *it will not*. Research into better ways of expressing rules is absolutely a worthwhile pursuit, one which can greatly reduce the scope of possible interpretations which an interpretation-capable agent must consider. Such research is complementary to the research I advocate here. But as explained in [3, 4], open-texturedness in rules is not a bug, but rather it is a feature. So long as human beings must follow, create, communicate, or reason about rules applied in non-trivial domains, open-texturedness will be a feature of those rules.

b) Why is contemporary work in explainable AI not sufficient? A powerful statistical algorithm with a robust explanation engine should be fine.: Addressing this question was largely the focus of Section II-A. To summarize: argument-justified AI overlaps with, but is ultimately different from, explainable AI. The former focuses on providing arguments for why stakeholders should accept outputs of systems, rather than simply explaining why the systems came up with those outputs. I do not claim that some black-box algorithm in the future might exist that will be capable of producing a perfectly correct interpretation of a rule every single time. But I did claim: (1) without accompanying supporting arguments, that interpretation will not be accepted by the stakeholders whose opinions matter; and (2) it seems unlikely that the black-box system could properly reach the correct interpretation without having internally done something resembling the consideration of arguments and potential counterarguments, so why not just make those considerations explicit?

c) Human beings carry out actions all the time without justifications for their actions. Why should we expect more out of artificial agents?: Let us be clear on a goal of the MDIA approach: we are asking what an operationalizable definition of "correct interpretation" should look like for AI interpreting textual rules in war-gaming. Now, in practice, carrying out

the computational effort required for MDIA may be too cumbersome. But it can still serve as a north star against which to compare other interpretation-finding algorithms—which is already more than can be said of interpretive argumentation without MDIA.

As for the fact that humans are not expected to provide justifications for their interpretations, this may be true. In fact, I do believe that the requirement to provide full argument graphs can and should be placed on humans in positions of authority, at least when transparency in decision-making is valued. But human judgment is such that justificatory argumentation in support or against a decision or action can be provided after the fact, even if that justification is post-hoc. For example, consider a law enforcement officer who performs an action that they believe is in accordance with a correct interpretation of the law. But afterwards, the officer's action is called into question, and they are compelled to testify before an oversight committee. Assuming the officer believes their actions were justified, then what sort of testimony might they provide? In most cases, it will either be a defense that their actions were justified due to some factor which overrides the law (e.g., perhaps they were attacked and were acting in self-defense), or that their actions were indeed performed within a proper interpretation of the law. And *the latter of these will come in the form of interpretive argumentation*.

Assume the officer chooses a defense on the basis of interpretive correctness, and that the oversight committee is convinced that the officer's interpretation of the law is in accordance with theirs. What if, through some futuristic technology that allows us to read past brain states, it is discovered that at the time of the action, the officer did not actually believe or reason using any interpretive argumentation whatsoever? In other words, what if it is somehow proven that the officer actually acted out of selfishness, but their action just coincidentally happened to be something that is defensible as being in accordance with a proper interpretation of the law? My inclination is to believe that the committee would let the officer off the hook for the action; after all, the officer did not *technically* break the law, rather they did the right thing for the wrong reasons. But it's not unreasonable to say that because the officer acted for the wrong reasons, some correction may be warranted; perhaps a mandatory re-training course, for example.

Now assume instead the officer was a robot. Would any of the considerations in the previous paragraphs change substantially? I do not believe that they would, save for the last: the robot officer would not take a mandatory re-training course, but would instead have its programming adjusted to ensure that in the future, it considers whether its actions are in accordance with the law. *But for the reasons described in this article, carrying out that task requires MDIA*. Thus, we are back where we started: non-MDIA rule-following AI will find itself needing to be MDIA anyway. Why delay the inevitable?

d) What if there is no such thing as a "correct" interpretation?: There is a pessimistic skepticism I have often encountered in discussing the ideas in this paper, according to which trying to understand interpretive reasoning is useless: they who have the political power will establish the correct interpretation regardless of the arguments provided. Indeed, I do not dispute that in many scenarios of importance, what determines which interpretation is ultimately adopted and enforced goes beyond considerations of rational deliberation, interpretive argumentation, and defeasibility. Furthermore, it may be that in many borderline cases, no amount of interpretive reasoning can clearly establish the dominance of one interpretation over another, and that in such scenarios, less-than-rational tiebreakers must be used. But this does not, by any stretch of the imagination, mean that there exists no possible process which can establish correct interpretations in the everyday rules which we follow a vast majority o the time—*even if what makes something a 'correct interpretation' is nothing more than whether an interpretation will be accepted by the current authoritative judicial system*.

The fact that the correct information is not necessarily the one that wins out in public discourse is not a reason to believe that correctness doesn't ever exist. Additionally, in many mundane cases (which are the types that our interpretation-capable agents in war-gaming environments will be faced with), there is general agreement on when certain interpretations are completely wrong. An example we have previously cited [40] comes from the *Amelia Bedelia* children's books [41]. The titular maid is presented with a written list of instructions on what to do around the house of her employers while they are away. The instructions tell her to "change the towels in the green bathroom," so she cuts them up with a scissors, thus changing their appearance. Instructed to "dust the furniture," she scatters dusting powder all over the furniture. Even children can tell that poor Amelia's interpretations are clearly incorrect, and it is this intuition which interpretation-capable reasoners must be able to simulate.

If an artificially intelligent agent is to effectively play complex war-games, or if such an agent is expected to be of use in actual warfare environments, it must be able to act in accordance with human laws, rules of engagement, conduct guidelines, mission-specific orders, and other agreements. Properly doing any of this requires minimally defeasible interpretive reasoning and argument-justified AI. It is time to stop kicking this can down the road, and seriously support such work.

REFERENCES

[1] F. Waismann, *The Principles of Linguistic Philosophy*. St. Martins Press, 1965.

[2] S. Blackburn, *Oxford Dictionary of Philosophy*. Oxford University Press, 2016.

[3] J. Licato, "Automated Ethical Reasoners Must be Interpretation-Capable," in *Proceedings of the AAAI 2022 Spring Workshop on "Ethical Computing: Metrics for Measuring AI's Proficiency and Competency for Ethical Reasoning"*, 2022.

[4] ——, "How Should AI Interpret Rules? A Defense of Minimally Defeasible Interpretive Argumentation," *arXiv e-prints*, 2021.

[5] D. N. MacCormick and R. S. Summers, *Interpreting Statutes: A Comparative Study*. Routledge, 1991.

[6] G. Sartor, D. Walton, F. Macagno, and A. Rotolo, "Argumentation schemes for statutory interpretation: A logical analysis," in *Legal Knowledge and Information Systems. (Proceedings of JURIX 14)*, 2014, pp. 21–28.

[7] R. M. Chisholm, *Perceiving*. Cornell University Press, 1957.

[8] H. Hart, "The ascription of responsibility and rights," in *Proceedings of the Aristotelian Society*, vol. 49, no. 1, 1949, pp. 171–194.

[9] J. L. Pollock, "Criteria and our knowledge of the material world," *Philosophical Review*, vol. 76, pp. 28–62, 1967.

[10] ——, "Defeasible reasoning," *Cognitive Science*, vol. 11, pp. 481–518, 1987.

[11] P. Croskerry, "From mindless to mindful practice — cognitive bias and clinical decision making," *New England Journal of Medicine*, vol. 368, no. 2445-8, 2013.

[12] G. Saposnik, D. Redelmeier, C. C. Ruff, and P. N. Tobler, "Cognitive biases associated with medical decisions: a systematic review," *BMC Medical Informatics and Decision Making*, vol. 16, 2016.

[13] S. Mithoowani, A. Mulloy, A. Toma, and A. Patel, "To err is human: A case-based review of cognitive bias and its role in clinical decision making," *Canadian Journal of General Internal Medicine*, vol. 12, no. 2, 2017.

[14] S. Prakash, S. Bihari, P. Need, C. Sprick, and L. Schuwirth, "Immersive high fidelity simulation of critically ill patients to study cognitive errors: a pilot study," *BMC Medical Education*, vol. 17, no. 1, p. 36, Feb 2017. [Online]. Available: https://doi.org/10.1186/s12909-017-0871-x

[15] C. Guthrie, J. J. Rachlinski, and A. J. Wistrich, "Inside the judicial mind," *Cornell Law Review*, vol. 86, no. 4, 2001.

[16] F. Fariña, R. Arce, and M. Novo, "Cognitive bias and judicial decisions," in *Much ado about crime*, M. Vanderhallen, G. Vervaeke, P. Van Koppen, and J. Goethals, Eds. Uitgeverij Politeia NV, 2003, pp. 287–304.

[17] B. Englich, T. Mussweiler, and F. Strack, "Playing dice with criminal sentences: The influence of irrelevant anchors on experts' judicial decision making," *Personality and Social Psychology Bulletin*, vol. 32, no. 2, pp. 188–200, 2006, pMID: 16382081. [Online]. Available: https://doi.org/10.1177/0146167205282152

[18] E. Peer and E. Gamliel, "Heuristics and biases in judicial decisions," *Court Review*, vol. 49, pp. 114–118, 01 2013.

[19] H. Mercier and D. Sperber, "Why do humans reason? arguments for an argumentative theory." *Behavioral and Brain Sciences*, vol. 34, no. 2, pp. 57–74, 2011.

[20] H. Mercier, "The argumentative theory: Predictions and empirical evidence," *Behavioral and Brain Sciences*, vol. 20, no. 9, pp. 689–700, 2016.

[21] D. Sperber and H. Mercier, *The Enigma of Reason*, audible audio edition ed. Tantor Audio, 2017.

[22] C. R. Wolfe, M. A. Britt, and J. A. Butler, "Argumentation schema and the myside bias in written argumentation," *Written Communication*, vol. 26, no. 2, pp. 183–209, 2009. [Online]. Available: https://doi.org/10.1177/0741088309333019

[23] J. A. Minson, V. Liberman, and L. Ross, "Two to tango: Effects of collaboration and disagreement on dyadic judgment," *Personality and Social Psychology Bulletin*, vol. 37, no. 10, pp. 1325–1338, 2011, pMID: 21632960. [Online]. Available: https://doi.org/10.1177/0146167211410436

[24] S. L. Cheung and S. Palan, "Two heads are less bubbly than one: team decision-making in an experimental asset market," *Experimental Economics*, vol. 15, no. 3, pp. 373–397, Sep 2012. [Online]. Available: https://doi.org/10.1007/s10683-011-9304-6

[25] E. M. Kesson, G. M. Allardice, W. D. George, H. J. G. Burns, and D. S. Morrison, "Effects of multidisciplinary team working on breast cancer survival: retrospective, comparative, interventional cohort study of 13 722 women," *BMJ*, vol. 344, 2012. [Online]. Available: https://www.bmj.com/content/344/bmj.e2718

[26] T. Kugler, E. E. Kausel, and M. G. Kocher, "Are groups more rational than individuals? a review of interactive decision making in groups," *Wiley Interdisciplinary Reviews: Cognitive Science*, vol. 3, no. 4, pp. 471–482, 2012. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1002/wcs.1184

[27] J. Kämmer, W. Gaissmaier, and U. Czienskowski, "The environment matters: Comparing individuals and dyads in their adaptive use of decision strategies," *Judgment and Decision Making*, vol. 8, no. 3, pp. 299–329, 2013.

[28] E. Mayweg-Paus, M. Thiebach, and R. Jucks, "Let me critically question this! – insights from a training study on the role of questioning on argumentative discourse," *International Journal of Educational Research*, vol. 79, pp. 195 – 210, 2016. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S088303551630043X

[29] D. Bang and C. D. Frith, "Making better decisions in groups," *Royal Society Open Science*, vol. 4, no. 8, pp. 170–193, 2017.

[30] M. Janier, J. Lawrence, and C. Reed, "Ova+: An argument analysis interface," in *Computational Models of Argument: Proceedings of COMMA 2014*, 2014.

[31] P. Dung, "On the acceptability of arguments and its fundamental role in nonmonotonic reasoning, logic programming and n-person games," *Artificial Intelligence*, vol. 7, no. 2, pp. 321–358, 1995.

[32] M. Caminada and L. Amgoud, "On the evaluation of argumentation formalisms," *Artificial Intelligence*, vol. 171, no. 5, pp. 286–310, 2007. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0004370207000410

[33] C. Chesñevar, J. McGinnis, S. Modgil, I. Rahwan, C. Reed, G. Simari, M. South, G. Vreeswijk, and S. Willmott, "Towards an argument interchange format," *Knowl. Eng. Rev.*, vol. 21, no. 4, pp. 293–316, Dec. 2006. [Online]. Available: http://dx.doi.org/10.1017/S0269888906001044

[34] C. Reed, K. Budzynska, R. Duthie, M. Janier, B. Konat, J. Lawrence, A. Pease, and M. Snaith, "The argument web: an online ecosystem of tools, systems and services for argumentation," *Philosophy and Technology*, vol. 30, no. 2, pp. 137–160, 2017.

[35] K. Atkinson, P. Baroni, M. Giacomin, A. Hunter, H. Prakken, C. Reed, M. Thimm, and S. Villata, "Towards artificial argumentation," *AI Magazine*, vol. 38, no. 3, 2017.

[36] D. Walton, "Some artificial intelligence tools for argument evaluation: An introduction," *Argumentation*, vol. 30, no. 3, pp. 317–340, Aug 2016. [Online]. Available: https://doi.org/10.1007/s10503-015-9387-x

[37] M. Lippi and P. Torroni, "Argumentation mining: State of the art and emerging trends," *ACM Transactions on Internet Technology*, vol. 16, no. 2, 2016.

[38] F. Bex and D. Walton, "Combining explanation and argumentation in dialogue," *Argument & Computation*, vol. 7, pp. 55–68, 2016.

[39] D. Doran, S. Schulz, and T. R. Besold, "What does explainable AI really mean? A new conceptualization of perspectives," *CoRR*, vol. abs/1710.00794, 2017. [Online]. Available: http://arxiv.org/abs/1710.00794

[40] Z. Marji, A. Nighojkar, and J. Licato, "Probing the Natural Language Inference Task with Automated Reasoning Tools," in *Proceedings of The 33rd International Florida Artificial Intelligence Research Society Conference (FLAIRS-33)*, E. Bell and R. Barták, Eds. AAAI Press, 2020.

[41] P. Parish, *Amelia Bedelia*. Harper & Row, 1963.

Can "Provably Beneficial AI" Save Us?

Selmer Bringsjord
Rensselaer AI & Reasoning Lab
RPI
Troy, USA
selmer.bringsjord@gmail.com

Naveen Sundar Govindarajulu
Rensselaer AI & Reasoning Lab
RPI
Troy, USA
naveen.sundar.g@gmail.com

John Licato
Advancing Machine & Human Reasoning Lab
University of South Florida
Tampa, USA
john.licato@gmail.com

***Abstract*—AI-polymath Stuart Russell, in the face of fear about superhuman AI arriving within 80 years and doing the human race in, commendably offers a recipe (based upon inductive reinforcement learning) for salvation quite different than our own (the sharing of which is beyond the current scope of the present paper). He does this in his recent book *Human Compatible*. Unfortunately, as we explain, Russell's recipe is afflicted by four fatal defects.**

***Index Terms*—machine ethics, robot ethics, inductive reinforcement learning**

I. Introduction: The Problem

AI-polymath[1] Stuart Russell, in the face of fear about superhuman AI arriving within 80 years and doing the human race in, offers a recipe for salvation quite different than our own (the sharing of which is beyond the current scope of the present short paper, but see e.g. [8]). He does this in his book *Human Compatible* [11]. Russell does not rely upon The Singularity (or any other such speculative thing) to justify his belief that superintelligent machines will arrive.[2] On the other hand, Russell is of the opinion that the arrival of superintelligent AI could very well be quite sudden. He writes:

> My timeline of, say, eighty years is considerably more conservative than that of the typical AI researcher. Recent surveys suggest that most active researchers expect human-level AI to arrive around the middle of this century. Our experience with nuclear physics suggests that it would be prudent to assume that progress could occur quite quickly and to prepare accordingly. If just one conceptual breakthrough were needed, analogous to Szilard's idea for a neutron-induced nuclear chain reaction, superintelligent AI in some form could arrive quite suddenly. The chances are that we would be unprepared: if we built superintelligent machines with any degree of autonomy, we would soon find ourselves unable to control them. I am, however, fairly confident that we have some breathing space because there are several major breakthroughs needed between here and superintelligence, not just one. [11, Chap. 3, § 7]

The remainder shall unfold straightforwardly as follows. In the next section we summarize what Russell offers as a solution to the threat to humanity from superintelligent AI. The section after that presents in sequence four problems that plague his proposal. Finally, the paper concludes with a brief discussion of the next steps to be taken in our assessment of Russell's approach, and in our consideration of competing approaches.

II. Russell's Proposed Solution

What is the solution Russell proposes? We cannot cover the ins and outs of his solution, as doing so would require a detailed explanation of *reinforcement learning* (RL), including *inverse* RL (IRL), upon which his proposal rests. While these forms of learning are mathematically simple frameworks in which agents gradually get better at reaching toward a goal, we nonetheless have not the time and space here to burn in exposition — and besides which RL and IRL are well-known to AI researchers. (Russell's [11] *Human Compatible* is in fact itself an excellent non-technical introduction to these forms of learning.) Fortunately, the core of Russell's proposed solution, what he calls "Provably Beneficial AI" (PBAI), can be quite efficiently conveyed here. The core of PBAI is that we take care to engineer robots driven solely by a "desire" to reach goals that accord with the goals of humanity. Of course, desire in the human case entais that the human doing the desiring has some states of "phenomenal" or "subjective" consciousness (what Block [1] calls 'p-consciousness'). This is so because, as we humans all know, when one desires something, one *feels* things, inevitably. For example, if one intensely desires to get some reward, and works ferociously toward it, but keeps failing to even get close to obtaining it, one is likely to e.g. feel frustrated, angry, despondent, and so on. Thus, we use scare quotes around 'desire' so as not to assume any such thing as that the robots Russell seeks will have p-consciousness.[3]

Encapsulated, what then in Russell's PBAI is the reward "desired" by the machines? He maintains that that reward will be none other than our own collective maximal well-being. Since we can safely assume that such goals in our case include that our species survives, and indeed overall thrives, if such a "desire" can be counted upon to really and truly drive our future robots, we should as a species be in good shape. In addition, we must be able to comfortingly *prove*

[1]Lead author of the encyclopedic, leading introduction and overview of AI, now out in its fourth edition: [12].

[2]The fact is, he does not really tell us in his book why he is so sure superintelligent AI will arrive — but he certainly is sure it will. Our educated guess is that Russell is content with his observing in his book the failure of numerous arguments against the proposition that superintelligent AI will arrive.

[3]According to the first author, they will have no such thing, and in fact no one at present has the slightest clue as to how to proceed with engineering that can be rationally regarded to move a nanometer closer to p-conscious AIs, as explained in [2].

that the robots are beneficial to humanity. Here is how Russell expresses overall his rather rosy take on things:

> [M]y proposal for beneficial machines: machines whose actions can be expected to achieve *our* objectives. Because these objectives are in us, and not in them, the machines will need [via IRL] to learn more about *what we really want* from observations of the choices we make and how we make them. Machines designed in this way will defer to humans: they will ask permission; they will act cautiously when guidance is unclear; and they will allow themselves to be switched off. [11, ¶ 2, § "Beneficial Machines" in Chap. 10 "Problem Solved?'; emphasis ours]

Unfortunately, while we have deep respect for the formality of Russell's approach (unsurprising since any real formality is rooted in formal logic and proofs therein: there is no other way to achieve a proof by to employ formal logic) there are four each-fatal-in-their-own-right problems plaguing Russell's proposal, as we now explain. Here now are these problems.

III. Four Problems Afflicting Russell's PBAI

As promised, we now proceed to explain, in turn, four defects (among others) that afflict PBAI.

A. Problem 1: Sola Utilitarianism?

The first problem is simple to grasp, and simply devastating; it is that Russell's proposal to save our race is based upon *only* the family of consequentialist ethical theories. This family includes the familiar ethical theory known as *act utilitarianism*, according to which what is obligatory are actions that maximize overall happiness; a precise account can be found in the classic [7]. But surely this particular family is only an *option* from among many families of ethical theories; and, these families are pairwise inconsistent. That is, pick any two families, and the definitions they include for the central operators of any ethical theory, for instance for *obligatory*, and one will arrives at contradictions, by elementary deductive reasoning over these definitions in garden-variety contexts. To see this, let us pick for consideration another ethical-theory family. Specifically, let us pick for expository purposes the family of *divine-command* ethical theories. Divine-command ethical theories are based upon the core notion that what is obligatory, permissible, forbidden, and so on is wholly determined by God's commands. A seminal presentation of a divine-command ethical theory is given by [10]. Exploration of divine-command ethical theories in a manner that conforms to what is needed in attempts to engineer morally correct machines is carried out in [4]. Note that when one considers the entire population of planet Earth, and subscription among its members to a dominant family of ethical theories, it is probably the divine-command family that has the largest number of adherents, by far.[4]

Now, given the setup supplied in the previous paragraph, here is a pair of relevant biconditionals, one from each of the two families we have just cited.[5] The first is part of act utilitarianism; the second is from all divine-command ethical theories.

Ob_{U} An agent (a category that includes human persons) is obligated at time t, given (context) Φ, to do action a at later time t' if and only if a, from among all viable alternative actions available to this agent, brings about the most happiness for the most people.

Ob_{DC} A human person is obligated at time t, given (context) Φ, to do action a at later time t' if and only if the performance of a has been commanded by God (or is deductively entailed by what has been commanded by God).

We are quite sure the reader can see the problem. By 'context' here, represented by 'Φ,' is meant simply a collection of declarative formulae, or for our somewhat informal exposition here, declarative propositions, that sets the situation. We can consider a hypothetical to make this more concrete: Molycarp is a devout Christian living under a brutal dictatorship whose key tenets include those of rabid and unrelenting atheism, and Molycarp is imprisoned, tortured, and asked to explicitly utter blasphemous and profane denial of his orthodox conception of Jesus as sinless and divine.[6] *Ex hypothesi*, Molycarp's agreeing to do this will save his life, ensure the well-being of his family (for which he is the breadwinner), and bring about many, many other happiness-bearing states-of-affairs through an endless array of chains of weal catalyzed by his subsequent actions. However, if he accepts death, only two terrestrial people will ever know what happened to him (the dictator and the executioner), as he will be incinerated, and in fact soon after his death everybody else will thoroughly forget about him. By a suitable instantiation of Ob_{DC}, Molycarp is obligated to proclaim his belief in Jesus and his divinity, and die a martyr; but in stark contrast, by a suitable instantiation of Ob_{U}, he is obligated to go through the motions of quickly spouting out a few words that will secure his freedom, and a lot of happiness that cannot otherwise be secured. Assuming that no one can be obligated to perform two actions that are impossible to both perform,[7] we have a contradiction.

There is more general, history-centric way to sum up Problem 1 for Russell, and for those inclined to follow him; it is to simply report that the discipline of systematic, theoretical ethics has been in progress since at least Aristotle, three centuries before the birth Christ, and if we know anything at all about the history of the discipline from that ancient timepoint we know that the human race has on hand myriad families of ethical theories, each none other than, as we have noted above, pairwise incompatible. It is thus rather doubtful that the solution to the problem posed by future superintelligent

[4]There are currently e.g. about 2.2 billion Christians on Earth, and about 2 billion Muslims. For both groups, by definition, it is first and foremost what God commands that determines what is obligatory. Orthodox and conservative Jews would of course be in precisely the same category. (This is of course not at all to say that the three religions here each perfectly agree on every attribute ascribed to God. The main ones, though — e.g. omnipotence, omniscience, omnipresence, omnibenevolence, creator of all contingent things — are indeed ascribed to God in the case of each of the trio of religions we cite here.

[5]For easing exposition, let us not worry about which particular ethical theory is in play here from each of the two families we have called out.

[6]The sinlessness and divinity of Jesus is a credal doctrine of orthodox Christianity. See e.g. [13]. Many readers will see in our use of 'Molycarp' a thinly disguised reference to the real martyr Polycarp, executed in 155 AD.

[7]This, that "ought implies can," is known as *Kant's Law*, and is a staple in deontic logic, the branch of logic devoted to logicizing ethical theories.

machines is to be found in the Russellian engineering of robots whose *modus operandi* is the following the dictates of only *one* family, consequentialism.

B. Problem 2: Mental States Not Inferrable from Behavior

The second problem afflicting Russell's approach to the threat to humanity is that this approach at its heart relies upon the ability of present and future AIs to infer a human's interior mentality from that human's exterior, readily observable behavior. After all, what Russell (admirably and rationally) wants is for the machines in question to place our happiness first among the goals they seek — but what is happiness if not a mental state, and as such an *invisible* state? (This is why we emphasized the phrase 'what we really want' in the quote of Russell just above.) This particular sentence is being written (at least in its first version) by author Bringsjord, who is thus simply staring at a screen and typing as characters appear on said screen for this eyes to take in. Okay, so suppose you walk up now to Bringsjord, who is seated, and look at his face, standing above him; and suppose that he stops typing and looks at your face. Can you tell if Bringsjord is happy? You may of course be able to rationally *assert* that he is happy, because you may have empirical data regarding his recent past (e.g., that he had a gourmet lunch featuring arctic char at Manhattan's Aquavit restaurant, a particular favorite of his, before his the current work session you just interrupted), and you may even happen to have a live feed from Selmer's iPhone somehow, giving you his vitals and perhaps all sorts of information about this bodily state, including its over internal condition in many regards, but — again, we assume here that Selmer is staring at you, expressionless — you will only be guessing. And in fact you would be wrong. Reason? Selmer happens to be thinking about an event in his childhood, a rather sad one: the death of his dog King, caused by a car; and his current state is far from a happy one, mixed as it is with some rather dreadful mental movies of what happened that fateful day just outside New York City.

Now, just replace you with a robot (or with an AI using sensors in the relevant room) looking at Selmer, and you will see the problem facing Russell. AIs cannot toil on our behalf by using inductive reinforcement learning because they cannot learn the nature of what they need to reduce or increase: namely, our mental states.

C. Problem 3: Cognition Ranges Beyond The Turing Limit

The next problem is quite simple to state. The robots that will be toiling in our favor are explicitly asserted by Russell to be boxed in by what a Turing machine can do. This is easy to confirm, because when he offers a theorem-schema that, when proved, will provide the ultimate assurance he seeks in the face of impending doom from superintelligent machines, that theorem-schema employs 'machine,' and this term means *Turing-level* machine. (We look at Russell's theorem-sketch below, in the final section.) Put another way, the robots with which Russell is concerned are all constrained by the Turing Limit, the level of computational power beyond which Turing machines (and lesser machines, e.g. linear-bounded Turing machines). But that means that if our cognition, our intellectual power, extends *beyond* this limit, the robots will not be able to grasp and abide by our cognition. But according to Bringsjord, human cognition is indeed of this nature; see for example assertions and defenses of this claim in [3, 5, 6].

It is important to grasp that the problem here for Russell's PBAI paradigm is not weak, vague, or haphazard; in fact the problem is logico-mathematical in nature. Suppose one computing machine m_1 is not capable of computing functions beyond some *bona fide* level L_1, and that some other computing machine m_2 is capable of computing functions at some level L_2 above L_1.[8] It then is an easy theorem that m_1, by observing the operation of the more powerful m_2, cannot compute functions at L_2, or for that matter one iota above L_1. Yet, Russell pins his hopes on robots that will observe us, and figure out how to work to our benefit. But what if our benefit requires doing things that demand as much cognitive power as we have? In that case it is mathematically impossible for his salvific recipe to work.

D. Problem 4: Humans Do Not Agree on Weighty Propositions

Let us suppose for the sake of argument that the Russellian beneficial-to-us robots can indeed somehow be magically engineered, so that at every moment of their existence, and perpetually so, they toil for *the* benefit of humanity: their sought-after reward is that very benefit. Notice our emphasis on the word 'the' in the previous sentence. That tiny little word, a so-called "determiner," creates a fatal problem for Russell. The problem is that there is no *the* thing that is humankind's benefit. What would this thing be, after all? Masochists seek their own harm and pain; sadists the harm and pain of others; criminals their own material benefit at the expense and pain of others; Christians perpetual bliss in an afterlife, this earthly life being no more than — quoting David — a vapor and — quoting Solomon — at its best filled with soul-making suffering; "brave" existentialists like Camus expend what they admit is pointless effort to stay alive even though this life is evanescent and absurd; and so on seemingly *ad infinitum* into never-ending heterogeneity. So, there is no *the* benefit, alas. The bottom line for Russell's PBAI explodes it; that bottom line is that each relevant group of humans, with enough wealth, is going to purchase a robot or robots in order to facilitate *their* priorities. If anything, this will just make the world as contentious and chaotic as it is now — maybe more so.

IV. Next Steps

The alert reader will recall that there is a 'P' for 'provably' in Russell's 'PBAI.' What is it that Russell says we need to prove in his approach? He gives the general shape of the theorems which, if proved, will constitute assurance. We read:

[8] We spare the reader technical bases beneath this imagined state-of-affairs, but mention here that this means that the levels must be ones in the Arithmetic Hierarchy or Analytic Hierarchy, and genuinely distinct ones therein. We cannot be referring to levels in the Polynomial Hierarchy, because all problems in that hierarchy are Turing-solvable.

> Let's look at the kind of theorem we would like eventually to prove about machines that are beneficial to humans. One type might go something like this:
>
> > Suppose a machine has components A, B, C, connected to each other like so and to the environment like so, with internal learning algorithms l_A, l_B, l_C that optimize internal feedback rewards r_A, r_B, r_C defined like so, and [a few more conditions] .... Then, with very high probability, the machine's behavior will be very close in value (for humans) to the best possible behavior realizable on any machine with the same computational and physical capabilities.

Russell's main point here is that such a theorem should hold regardless of how smart the components become — that is, "the vessel never springs a leak and the machine always remains beneficial to humans" ([11, Chap. 8, § "Mathematical Guarantees," ¶ 8]). The next step in our evaluation of PBAI is to investigate carefully how theorems of this general shape can *in fact* be proved. This will require formalizing the concepts that Russell leaves vague and undefined here. For example, what, logico-mathematically speaking, is a 'machine' in the theorem-sketch that Russell provides here?[9] Likewise, what precisely is 'the environment'? At the very least, we shall need to venture precise answers to these questions in order to understand what Russell is gesturing toward when he sketches the kind of theorem to target in PBAI. We will then need to see if in fact an actual theorem of this shape can be proved, and what the proof would need to be like. Following on this, another step will be to see if, in approaches very different than PBAI, theorems providing greater assurance can be obtained. After all, Russell here concedes, explicitly, that the best his approach can reach is only "very high probability" that the machines will operate in our interests. We believe that total assurance can in fact be secured on the strength of proving theorems of a different nature than what Russell describes, and will seek to demonstrate that our optimism is well-founded.

ACKNOWLEDGMENTS

We are indebted to Stuart Russell for bravely and perspicaciously dealing with an acute future danger that many may wish to ignore or at least severely downplay. We are deeply grateful to ONR for past, extended support of research in the area of robot ethics (that informs the present paper), in particular through a MURI grant on which both Bringsjord and Govindarajulu were central researchers (along with PI Matthias Scheutz and Co-PI Betram Malle).

REFERENCES

[1] N. Block. On a Confusion About a Function of Consciousness. *Behavioral and Brain Sciences*, 18:227–247, 1995.

[2] S. Bringsjord. Offer: One Billion Dollars for a Conscious Robot. If You're Honest, You Must Decline. *Journal of Consciousness Studies*, 14(7):28–43, 2007. URL http://kryten.mm.rpi.edu/jcsonebillion2.pdf.

[3] S. Bringsjord and K. Arkoudas. The Modal Argument for Hypercomputing Minds. *Theoretical Computer Science*, 317:167–190, 2004.

[4] S. Bringsjord and J. Taylor. The Divine-Command Approach to Robot Ethics. In P. Lin, G. Bekey, and K. Abney, editors, *Robot Ethics: The Ethical and Social Implications of Robotics*, pages 85–108. MIT Press, Cambridge, MA, 2012. URL http://kryten.mm.rpi.edu/Divine-Command_Roboeth

[5] S. Bringsjord and M. Zenzen. *Superminds: People Harness Hypercomputation, and More*. Kluwer Academic Publishers, Dordrecht, The Netherlands, 2003.

[6] S. Bringsjord, O. Kellett, A. Shilliday, J. Taylor, B. van Heuveln, Y. Yang, J. Baumes, and K. Ross. A New Gödelian Argument for Hypercomputing Minds Based on the Busy Beaver Problem. *Applied Mathematics and Computation*, 176:516–530, 2006.

[7] F. Feldman. *Introductory Ethics*. Prentice-Hall, Englewood Cliffs, NJ, 1978.

[8] N. Govindarajulu and S. Bringsjord. On Automating the Doctrine of Double Effect. In C. Sierra, editor, *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence (IJCAI-17)*, pages 4722–4730. International Joint Conferences on Artificial Intelligence, 2017. ISBN 978-0-9992411-0-3. doi: 10.24963/ijcai.2017/658. URL https://doi.org/10.24963/ijcai.2017/658.

[9] J. D. Hamkins and A. Lewis. Infinite Time Turing Machines. *Journal of Symbolic Logic*, 65(2):567–604, 2000.

[10] P. Quinn. *Divine Commands and Moral Requirements*. Oxford University Press, Oxford, UK, 1978.

[11] S. Russell. *Human Compatible: Artificial Intelligence and the Problem of Control*. Penguin Books, New York, NY, 2019. This is the ebook version, specifically an Apple Books ebook.

[12] S. Russell and P. Norvig. *Artificial Intelligence: A Modern Approach*. Pearson, New York, NY, 2020. Fourth edition.

[13] R. Swinburne. *Was Jesus God?* Oxford University Press, Oxford, UK, 2010.

[9] Apropos of the discussion above, what about computing machines that are provably capable of more than what can be done by a standard Turing machine? E.g., what about infinite-time Turing machines [9]?

Toward Comparing Knowledge Acquisition in DeepRL Models

Anthony Marchiafava, Atriya Sen
Department of Computer Science
University of New Orleans
New Orleans, United States
email: amarchia@uno.edu, asen@uno.edu

***Abstract*—In order to better exploit Deep Reinforcement Learning (DeepRL) systems such as DeepMind's Alpha Go & Alpha Zero, it is desirable to understand how they acquire knowledge, and how human knowledge acquisition can contribute to or benefit from such an understanding. We analyze a series of DeepRL models trained to play the board game of chess in a human-like fashion, to study if these models acquire concepts differently from self-trained DeepRL models such as AlphaZero. Our preliminary results indicate that human chess players may acquire concepts very similarly to self-trained models. We further discuss some of the potential consequences of such an outcome.**

Keywords-Artificial Intelligence; Deep Reinforcement Learning; Reinforcement Learning; Deep Learning; Explainable AI.

I. INTRODUCTION

The game of chess has been called the "drosophila" of artificial intelligence (AI), referring to the extensive use of fruit flies (*drosophila*) in experimental biology. While traditional chess engines rely primarily on tree search with advanced heuristics, many modern approaches have exploited deep learning or deep reinforcement learning.

One such recent project is AlphaGo Zero, which uses a combination of Monte Carlo Tree Search (MCTS) and a deep neural network [1]. Leela Chess Zero is an open-source implementation of both the MCTS and the (convolutional) neural network of AlphaGo Zero, and achieves a similar level of performance (i.e., playing strength), which is to say, a superhuman level: capable of consistently defeating any known human player.

It may be reasonably hypothesized that such neural systems are learning implicit knowledge about chess-playing concepts and strategies. Understanding the internal knowledge acquisition processes of these and similar systems have the potential to provide insight into both chess as a game and the application of a similar process to varied adversarial domains, such as international trade, nuclear deterrence, and other negotiations.

The MCTS algorithm is used to examine the possible outcomes of the game depending on which move is chosen, by searching through trees generated from different choices the player could make, and examining which ones lead to the highest probability of winning [2]. These trees are generated by the deep neural network.

The deep neural network is fundamentally a two-state regression or classification model which accepts some input and produces one or more outputs [3]. The network will accept the input and produce derived features, which are then used to produce further derived features depending on network depth, and derived features are combined using an output function to produce the final output. Derived features are produced using linear combinations of the inputs and activation functions and other operations at different layers of the network. The first few layers more closely match the structure of the initial input, but as further derived features are generated, the derived features become more and more abstract.

The deep neural network accepts the current state of the chess board, prior states of the chess board, including a number of additional game-specific parameters such as the current castling status, and finally move count as input, and produces two outputs via dual network heads: (1) the policy head, which produces the probability distribution of possible moves, and (2) the value head, which produces the predicted outcome of the game, based on making the suggested move, as a win, lose, or draw. The MCTS uses the output of the neural network to choose the best candidate move. AlphaGo Zero learnt to play chess without exposure to human moves or more abstractly, playing styles, and generated implicitly expressed strategies sophisticated enough that it prevailed in a multi-game match against Stockfish, then a traditional search-based engine (Stockfish has now been updated to additionally use a neural model).

In an effort produce engines that behave more human-like at a variety of skill levels, Maia Chess was created [4]. Different versions of Maia were trained on specific games of human players at different skill levels, in lieu of using self-play, effectively training the neural network in human-style play. The different versions of Maia were able to produce gameplay choices similar to human players from 1100 ELO to 1900 ELO, where ELO refers to the ELO rating system, which is used almost exclusively in chess, and refers to a relative ranking of a particular player's odds of winning against another player of a different skill level (i.e., ELO rating). Maia was built on the Leela Chess Zero framework, an open-source engine inspired by Alpha Zero. However

 ISBN: 978-1-68558-014-8

Maia does not use an MCTS, but rather uses a deep learning model exclusively.

We will show how we can detect and compare the concepts that the various versions of Maia use. In section 2 we indicate what motivated our work here and what similar work was done in the past. In section 3 we show the technique we used to get concepts and how they are detected. In section 4, we show the results we were able to generate. In section 5 we show what further work can be done in the future.

II. MOTIVATION & PRIOR WORK

Since deep neural networks (DNNs) are inherently black boxes, the ability to understand and explain the presumed acquisition of concepts and strategies by the network in chess and other adversarial domains is highly desirable for a variety of reasons, including training humans in "superhuman" strategies, and interpreting them in terms of human strategies. A DNN learns derived features generated using the backpropagation algorithm, which updates intermediate node weights based on the gradient between the observed output and the expected output at the final layer of the network, a process which is not directly human-interpretable.

One interpretability technique is to use the technique of *linear probing* to examine the detectability of concepts at the intermediate layers of the neural network, and the acquisition of knowledge those concepts entail [5]. This approach is derived from a technique for detecting image concepts in computer vision using *concept activation vectors* (CAVs) [6]. We separate examples of game states which have some concept in common, and examples which do not exhibit that concept. These classifications are matched to the activations of a particular layer in a neural network, whose input matches our game state. We then train a linear classifier to differentiate between the inputs of the two classes. This allows us to detect if the set of activations of a particular layer for a particular network contain the information needed to determine if a concept is present or absent at that layer. This has been the approach taken in [7], for interpreting concepts learnt by Alpha Zero.

In this paper we present the results of a similar examination using linear probing, to compare the behavior of concept and strategy knowledge acquisition across various versions of weights learnt by the Maia network, to compare the concept acquisition of a model trained on human play against a model trained by self-play. Our preliminary results indicate that human chess players may acquire concepts very similarly to self-trained models.

III. TECHNIQUE AND PROCESS

To understand the concepts we will compare we must detect the concepts, preprocess our input data to be interpretable by the modified version of the network we need to use, and get the activations from the intermediate layers we examine.

A. Concepts

The concepts we tested for in the DNN's chess understanding were material advantage and a modified version of material advantage from the perspective of the player with the white pieces. The concept of material advantage is defined by adding up the number of pieces one player has remaining on the board, adjusted by the assumed inherent value of those pieces, and subtracting the value of the other player's pieces. A pawn is worth 100 points, a knight is worth 320 points, a bishop is worth 330 points, a rook is worth 500 points, and a queen is worth 900 points. So, a player with three pawns and one queen is worth 1200 points and a player with only two rooks is worth 1000 points, so there would be a 200-point advantage for the first player over the second. The king is not assigned a material value, since losing the king is not possible in chess.

The second concept includes the previously mentioned material advantage modified by an increased weight for pieces in more advantageous positions and a penalty for disadvantageous positions. The weights of these positions are defined by a piece-square table. Each piece-square table is an 8x8 array of numbers where each defines a modifier for the *quality* of each piece in that position, referring to the postulated long-term strategic advantage or disadvantage of a piece being in that position.

We created a unique piece-square table for each type of piece. These piece-square tables are each oriented towards whoever is the player whose board position is being evaluated. We used publicly available human-play ranked games from the online chess platform Lichess to generate the game states, to generate the game states over which to check for the two material advantage concepts. Lichess is a popular platform and has many years of games to draw upon. The Lichess games were also in same format of the games which were used to train the Maia networks used, the Portable Game Notation (PGN) format, used to notate each move made by either player over the course of a single game. Combined with knowledge about chess boards and game states, PGN files are sufficient to generate every game state occurring over the course of a game. The files also included the metadata about the players, including their ELO ratings.

B. Preprocessing

We used the same tools used to generate the Maia training data, to create a dataset of 204,800 sample game states. First, we separated games by ELO using pgn-extract [8], a tool for extracting games using portable game notation formatted games. This allowed us to remove games which may have been trained on already, and allowed us to evaluate games which were not in the training dataset for a particular version of Maia. These games were then converted into a format suitable for providing the Leela Chess Zero (Lc0) neural network using trainingdata-tool [9], which is designed to convert from PGN games to the Lc0 format. These are stored in binary files which are not human-readable. Since each game was entirely converted into a series of inputs - one input for each move in the game - we also needed to know which game state corresponded to which concept.

Therefore, we converted each given input back into a format which could be evaluated for material advantage and modified material advantage. This provided both the input in the correct format and a more easily interpretable version that we maintained as linked to each other.

C. Activation Layers as Input

For each activation layer and Maia version we wished to examine, we then generated the activations of the neural network up to that layer and stored those activation values. Examining 19 activation values for the version of Maia trained to behave like a player in the 1200-1299 ELO range created examples of what concepts the network could detect at each of those hidden layers, in the samples provided.

We then created a new DNN model whose input was the original input and whose output was the activation values of the layer we wished to examine in the original network, creating 19 sets of activations for each input. These were then used as input to a classifier whose output was the presence or absence of the material advantage concept. We then trained a classifier to determine if a layer's activation was correctly classified. That is, for a game state which shows a material advantage for the active player, that game state when converted to an input should produce layer activations which can be classified correctly if that layer includes the information that the concept classification requires.

IV. RESULTS

We examined the odd numbered activation layers for both the concepts previously mentioned, across the three ELO categories of 1200, 1400, and 1900, for both simple material advantage and material advantage incorporating the weights found in the piece-square tables. The more basic material advantage was detectable with roughly 73% accuracy across the three categories and across all the activations examined. To evaluate this, we split our classifier data into a training set of 200,000 samples and a testing set of 4,800 samples.

TABLE I. MATERIAL ADVANTAGE

Maia	Activation Layer Classification Accuracy		
	Activation_1	***Activation_9***	***Activation_19***
1200	0.737708	0.737916667	0.738958
1400	0.738542	0.738125	0.738125
1900	0.737708	0.7375	0.738333

TABLE II. MODIFIED MATERIAL ADVANTAGE

Maia	Activation Layer Classification Accuracy		
	Activation_1	***Activation_9***	***Activation_19***
1200	0.534167	0.550625	0.529792
1400	0.537083	0.544375	0.544375
1900	0.535208	0.546041667	0.543958

Comparable results were obtained from an examination of AlphaGo Zero in [7], the conclusion being that material advantage as a concept is relatively easy to detect, even from the inputs without activations, and provides a good baseline to evaluate the concept detection system. Each version of Maia was fully capable of detecting the concept to a similar degree: we can conclude that this concept is not sufficiently different across the different ELO categories of Maia models.

The results of examining for modified material advantage show our technique to be less accurate. This may be because our modified version of material advantage is not sufficiently aligned with a concept that any version of Maia is looking for. There may be some weighted version of material advantage that Maia may use, but the specific concept we attempted to detect does not appear to be one used by Maia. This indicates that it is necessary to explore other concepts to further understand the different behaviors of the Maia models.

V. CONCLUSION AND FURTHER WORK

The specific domain concepts examined here represent a proof of concept of our strategy. Since the accuracy of each version of Maia is similar across the ELO ranges used, other more subtle concepts may be more effective at showing the differences between the human-trained models. Or, if the concept detection is the same across all versions of Maia for most concepts, further work is necessary to understand the difference in behavior but similarity in concept detection. If, for example, the data necessary to detect a particular concept differs between versions of Maia or Lc0, then we can say that part of that concept is potentially used in differentiating the final behavior.

A more thorough examination of the behavior of a self-trained model which exactly uses the Maia network's structure would be additionally worth comparing to, as the default Leela Chess Zero weights did not match with the version used by Maia. Further work on comparing a self-play trained model such as Lc0 to one trained entirely on human generated data such as Maia, may show novel rationale for the difference in quality and behavior between these systems.

REFERENCES

[1] D. Silver, T. Hubert, J. Schrittwieser, I. Antonoglou, M. Lai, A. Guez, M. Lanctot, L. Sifre, D. Kumaran, T. Graepel, T. Lillicrap, K. Simonyan and D. Hassabis, "A general reinforcement learning algorithm that masters chess, shogi, and Go through self-play," *Science*, vol. 362, no. 6419, pp. 1140-1144, 2018.

[2] M. Świechowski, K. Godlewski, B. Sawicki and J. Mańdziuk, "Monte Carlo Tree Search: A Review of Recent Modifications and Applications," arXiv, 2021.

[3] T. Hastie, R. Tibshirani and J. Friedman, The Elements of Statistical Learning, New York: Springer New York Inc, 2001.

[4] R. McIlroy-Young, S. Sen, J. Kleinberg and A. Anderson, "Aligning Superhuman AI with Human Behavior: Chess as a Model System," in 2020 ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD 2020), 2020.

[5] A. Guillaume and Y. Bengio, Understanding intermediate layers using linear classifier probes, arXiv, 2016.

[6] B. Kim, M. Wattenberg, J. Gilmer, C. Cai, J. Wexler, F. Viegas and R. Sayres, "Interpretability Beyond Feature Attribution: Quantitative Testing with Concept Activation Vectors (TCAV)," arXiv, 2017.

[7] T. McGrath, A. Kapishnikov, N. Tomašev, A. Pearce, D. Hassabis, B. Kim, U. Paquet and V. Kramnik, "Acquisition of Chess Knowledge in AlphaZero," arXiv, 2021.

[8] https://www.cs.kent.ac.uk/people/staff/djb/pgn-extract/, last accessed July 12th, 2022.

[9] https://github.com/DanielUranga/trainingdata-tool, last accessed July 12th, 2022.

Proposal of In-house Development Model for Business System at Kagawa University

Satoru Yamada
Graduate school of Engineering
Kagawa University
Kagawa, Japan
email: yamada.satoru@kagawa-u.ac.jp

Yosuke Yatani
Graduate School of Science
for Creative Emergence
Kagawa University
Kagawa, Japan
email: s22g363@kagawa-u.ac.jp

Norifumi Suehiro
Graduate School of Science
for Creative Emergence
Kagawa University
Kagawa, Japan
email: suehiro.norifumi@kagawa-u.ac.jp

Horoki Asakimori
Information Technology
and Media Center
Kagawa University
Kagawa, Japan
email: asakimori.hiroki@kagawa-u.ac.jp

Yusuke Kometani
Information Technology
and Media Center
Kagawa University
Kagawa, Japan
email: kometani.yusuke@kagawa-u.ac.jp

Rihito Yaegashi
Information Technology
and Media Center
Kagawa University
Kagawa, Japan
email: yaegashi.rihito@kagawa-u.ac.jp

***Abstract*—The development of a new system or product service is not a sure thing. A new development method that identifies the Minimum Viable Product (MVP) and starts the development of a system or service is attracting attention. We propose a In-house Development Model to identify the MVP of a business system and develop the system in-house by the themselves.**

***Index Terms*—Digital Transformation, Agile Development, User-driven Development, In-house Development**

I. Introduction

The impact of COVID-19 is changing the global social structure. Digital transformation is necessary to adapt to change. The information system plays an important role in promoting DX of User Companies. However, user companies have the problem of ”starting the development of an information system with unclear requirements”. Agile development of the information system in user companies using ”Low-code/No-code tools” is attracting attention as a way to promote digital transformation.

”Design Thinking” [1] was proposed at the HASSO PLATTNER Institute of Design at Stanford. ”Design Thinking” consists of five steps: ”EMPATHIZE”, ”DEFINE”, ”IDEATE”, ”PROTOTYPE”, and ”TEST”. Fig. 1. is an example of the ”Design Thinking” process. ”Design Thinking” is a necessary concept for creating new value. Many companies are implementing initiatives based on ”Design Thinking”.

”Lean Startup” [2] was proposed as a methodology for launching a business under conditions of high uncertainty. A Minimum Viable Product(MVP) is developed in ”Lean Startup”. It provide users with MVPs based on hypotheses and define value by ”Verification” with them through the ”Build-Measure-Learn” cycle. ”Design Thinking” realizes ”Human-Centered” value delivery. However, ”Lean Startup” emphasizes the verification of business feasibility based on hypotheses.

Kagawa University defined ”Hypotheses” for business system requirements based on a ”Human-Centered” (Ex. faculty, staff, and students) approach. The ”Hypothesis” is ”verified” through co-creation with users (Ex. faculty, staff, and students). Kagawa University proposes the ”In-house development model for Business System at Kagawa University” in which ”Human-Centered” business system requirements are defined as ”Hypotheses”, and ”Verification” is conducted through co-creation with users. The ”In-house development model for Business System at Kagawa University” combines ”Design Thinking”, and ”Lean Startup”. Kagawa University is currently developing a business system using the ”In-house development model for Business System at Kagawa University”. This paper describes the ”In-house development model for Business System at Kagawa University”. The ”In-house development model for Business System at Kagawa University” is based on the iterative model of agile development. Development is done in phases. The iterative model of agile development in general aims to increase the product quality [4] of the system. However, the ”In-house development model for Business System at Kagawa University” defines a ”Hypothesis” that enhances the quality of usability. Then, the users (faculty, staff, and students) themselves develop the business system by repeating ”PROTOTYPE”, and ”TEST” through co-creation of the ”Hypothesis”. Therefore, the iterative model differs from the general agile development iterative model, which enhances ”product quality”. The ”In-house development model for Business System at Kagawa University” focuses on ”user value” rather than ”product quality”. Several methods for defining hypotheses, such as ”Design Thinking”, have been proposed. Therefore, this paper does not limit the methods for defining ”Hypotheses”.

Section II describes related research and related technolo-

 ISBN: 978-1-68558-014-8

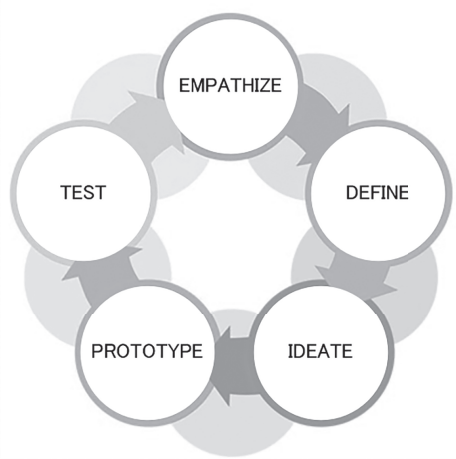


Fig. 1. Process of Design Thinking.

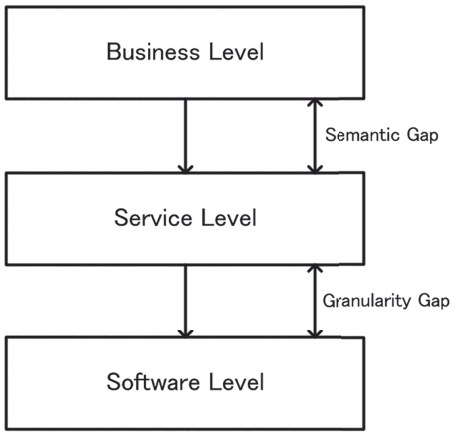


Fig. 2. User-driven development approach which Chusho [5] Proposes.

gies. Section III describes an ”In-house development model for Business System at Kagawa University”. Section IV provides a Results. Section V provides a Conclusion.

II. Related Research and Related Technologies

Chusho [5] says ”End users (defined as users in this paper) who have knowledge of the business develop systems and software on their own initiative. It is also important that users take the lead in maintenance”. The user-driven development proposed by Chusho is a three-tier architecture: ”Business Level”, ”Service Level”, and ”Software Level”. Fig. 2. shows the user-driven development approach proposed by Chusho. At the ”Business Level”, users with business knowledge create business models. At the ”Service Level”, create a domain model based on the ”Business Model”. Software is developed at the ”Software Level” from the created domain model. Chusho says, ”A semantic gap is created between the business level and the service level. Domain Knowledge Complementary Technology is a technology that complements the semantic gap”. Kato et al. [6] proposed a request acquisition method named THEOREE. The method proposed by Kato et al. systematizes domain knowledge by means of a thesaurus, which improves the efficiency of requirements analysis by providing a systematic thesaurus to analysts who lack sufficient domain knowledge. Kato et al.’s research falls under the category of domain knowledge completion technology. At the ”Software Level” systems and services are developed by utilizing components. Chusho says, A Software Unit Gap is created between the Service Level and the Software Level. ”Business Objects” [7] , ”Design Patterns” [8], and ”Frameworks” [9] are complementary technologies to the Software Unit Gap. The smaller the description unit of a program, the greater the scope of application because it can be expressed in a manner similar to a programming language. However, if the Software Unit is made larger and expressed in a business-like manner, it will be easier for users to use, but the scope of application will be limited.

With the development of information and communication technology, End-User Computing(EUC) [10] with ”Low-code/No-code tools” that enable system and software development without advanced programming knowledge is attracting attention. ”Low-code/No-code tools” have been introduced for use in DX promotion as a means to respond to the ”ambiguity of needs”, and ”rapidly changing requirements” for system and software development [11]. In addition, development using ”Low-code/No-code tools” is expected to significantly reduce development man-hours and shorten the time to ”Verification” of the MVP. Therefore, there is little Software Unit Gap between requirements and deliverables, and it has been reported that it is effective in developing systems and software with specific MVPs [12].

III. In-house development model for Business System at Kagawa University

Fig. 3. shows the ”In-house development model for Business System at Kagawa University” proposed in this paper. Chusho showed that business knowledge is important for users to develop systems and software that they themselves need, and proposed a three-tier architecture (”Business Level”, ”Service Level”, and ”Software Level”). Kagawa University integrated the ”Service Level” into the ”Software Level” by utilizing ”Low-code/No-code tools” based on the tree-tier architecture proposed by Chusho. In order to emphasize the definition of ”Hypothesis” for the realization of ”Human-Centered” value and the ”Verification” of MVP, we defined a three-step approach (”Business level”, ”Software level”, and ”Verification level”) with a ”Verification Level” to ”Evaluate” the developed system or software. By iteratively repeating this three-step approach multiple times, users themselves develop the systems and software they need. In this paper, the ”Low-code/No-code tool” was used to integrate the ”Service Level”, and ”Software Level”. However, if software can be developed without any granularity gap between the ”Service Level”, and the ”Software Level,” there is no need to use ”Low-code/No-code tools.

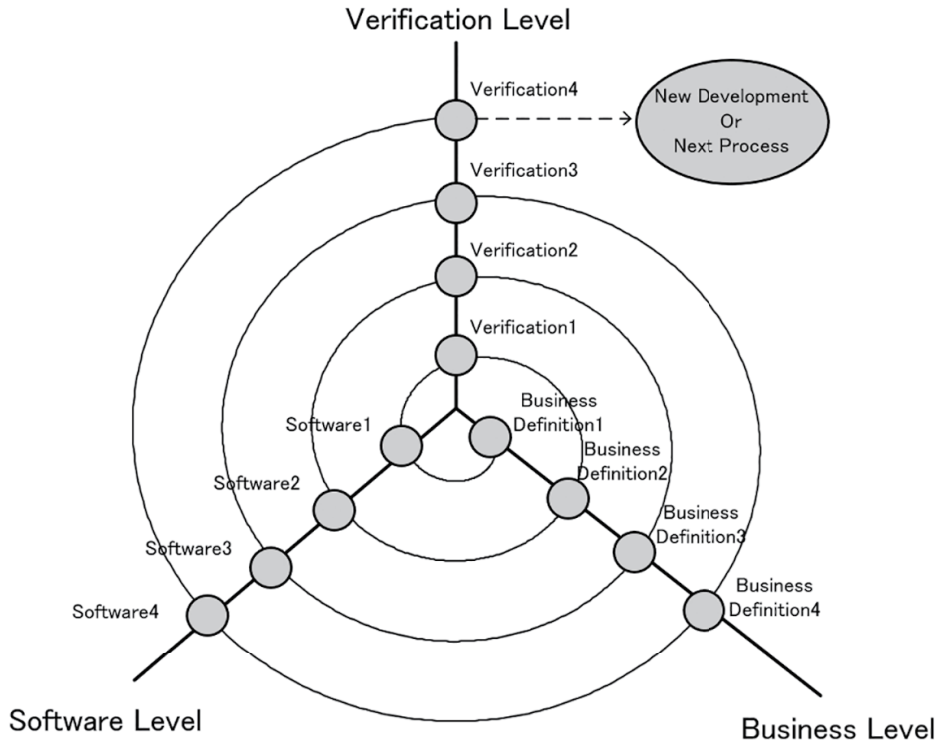


Fig. 3. In-house development model for Business System at Kagawa University.

In the ”In-house development model for Business System at Kagawa University” business is defined at the ”Business Level”. At the ”Business Level”, the ”Hypothesis” for the business system is defined and MVPs are identified through co-creation with users (faculty, staff, and students) who have business knowledge. At the ”Software Level”, systems and software are developed based on business definitions. At the ”Software Level”, MVPs are developed from hypotheses defined at the ”Business Level”, utilizing ”Low-code/No-code tools”. Developed systems and software are evaluated at the ”Verification Level”. At the ”Verification Level”, not only is the product quality of the system or software evaluated based on the business definition defined at the ”Business Level”, but also the validity of the ”Hypothesis” or MVP for the user value defined at the ”Business Level” is evaluated. At the ”Verification Level”, the continuation of development is also discussed. If the decision to continue development is made at the ”Verification Level”, the business definition is modified or added at the ”Business Level”. Develop improved systems and software at the ”Software Level” based on the reviewed business definitions. If a decision to terminate development is made at the ”Verification Level”, development is terminated. In the ”In-house development model for Business System at Kagawa University”, another system development or new needs may be discovered through system or software development. After the system development is completed, a new development project is launched. This paper does not define how to define a ”Business Model”, and how to generate a ”Software Level”, and how to evaluate a ”Validation Level”. In this paper, it is assumed that the method to be used is to select the necessary method according to the target business and the system or software to be developed.

IV. Results

Kagawa University established ”DX Promotion Division”, and ”DX Laboratory” in May 2021. In the ”DX Laboratory”, IT and business divisions collaborate to develop business systems in-house through co-creation. The ”DX Laboratory” works in the ”DX Project Team”. The ”DX Project Team” consists of users (faculty, staff, and students). Users with business knowledge from the business departments participate in the ”DX Project Team”. The ”DX Project Team” defined a ”Hypothesis” and identifies an MVP based on the ”In-house development model for Business System at Kagawa University”. In October 2021, there were six ”DX Project Teams”. The six ”DX Project Teams” have developed twenty-five projects business system in-house. Fifteen projects have already been completed. Ten business systems were developed in five months. At the ”Software Level”, the software was developed using the ”Microsoft Power Platform” [13], a ”Low-code/No-code tools”. The ”Microsoft Power Platform” includes four services: ”Microsoft Power Apps” [14], ”Microsoft Power Automate” [15], ”Microsoft Power BI” [16], and ”Microsoft Power Virtual Agents” [17].

Using the ”In-house development model for Business System at Kagawa University”, we interviewed the staff who

developed the business system. There are five questions. Question 1: Do you feel a ”Semantic Gap” from the ”Business Level” to the ”Software Level”? Question 2: Do you feel a ”Granularity Gap” at the ”Software Level”? Question 3: An impression of the use of ”In-house development model for Business System at Kagawa University”. Question 4: An impression of ”Design Thinking” and co-creation activities. Question 5: An overall impression.

All four respondents answered ”no Semantic Gap” for Question 1. The reason for this was that ”staff members who understand the work develop software at the ’Software Level’, so they do not feel a ’Semantic Gap’”. All four respondents answered ”no Granularity Gap” for Question 2. The business system is a flow definition using ”Microsoft Power Automate” with ”Low-code/No-code tools”. Therefore, I do not feel any ”Granularity Gap”. The respondents to Question 3 answered, ”Until now, we could not implement a system without ordering from a vendor, but now we can implement a system with a sense of speed”, ”We can implement a system that we really think is necessary”, and ”The larger the scale of the system, the more difficult it is for end users to develop”. The respondents to Question 4 answered, ”It was easier to share specific issues”, and ”The motivation of the business units made a difference in the results”. The respondents to Question 5 answered, ”the data obtained from the system is useful”, ”I want to improve the system based on the data”, and ”reviewing the operations gave me an opportunity to think about whether the operations were necessary”. The interview results indicate that the ”In-house development model for Business System at Kagawa University”, has the potential to solve the ”Semantic Gap”, and ”Granularity Gap”.

V. Conclusion

In this paper, we define a ”Hypothesis” for the realization of ”Human-Centered” value. The ”In-house development model for Business System at Kagawa University” in which business systems are developed by ”Verification” of the defined ”Hypothesis” through co-creation with users, was described. The ”In-house development model for Business System at Kagawa University” combines ”Design Thinking” and ”Lean Startup”. Through ”co-creation” between the IT and business divisions, ”Hypotheses”, and MVPs for the realization of ”Human-Centered” value can be identified, and business systems using EUC can be produced in-house using ”Low-code/No-code tools”. The ”In-house development model for Business System at Kagawa University” can define MVP by three steps: ”Business Level”, ”Software Level”, and ”Verification Level”. The ”In-house development model for Business System at Kagawa University” has the potential to solve the problem of ”starting development with unclear requirements” for user companies working to promote DX.

Using the ”In-house development model for Business System at Kagawa University”, an interview survey was conducted with university employees who have developed their business systems in-house. One comment was, ”It is difficult to judge what end-users can and cannot develop with EUC”. And, Some projects were terminated because the ”Hypothesis”, or ”MVP” could not be verified at the ”Validation Level”. The future work is to clarify the conditions under which end-users can participate in development and to establish guidelines.

References

[1] HASSO PLATTNER Institute of Design at Stanford, ”An Introduction to Design Thinking PROCESS GUIDE,” https://web.stanford.edu/~mshanks/MichaelShanks/files/509554.pdf [retrieved: July, 2022].

[2] Eric Ries, ”The Lean Startup.” Portfolio Penguin, 2011.

[3] National Strategy office for Information and Communications Technology Cabinet Secretariat, ”Guidebook of Agile Development Practice,” https://cio.go.jp/sites/default/files/uploads/documents/Agile-kaihatsu-jissen-guide_20210330.pdf [retrieved: July, 2022].

[4] Information-technology Promotion Agency, Japan, ”Guidebook of Software Quality in a Connected World,” https://www.ipa.go.jp/files/000055008.pdf [retrieved: July, 2022].

[5] Takeshi Chusho, ”Enduser-Initiative Application Development Methods with Business Knowledge,” http://www.isc.meiji.ac.jp/~chusho/paper/1611KBSEchu.pdf [retrieved: July, 2022].

[6] Junzo Kato, Motoshi Saeki, Atsushi Ohnishi, Haruhiko Kaiya, and Shuichiro Yamamoto, ”THEOREE: THEsaurus Oriented REquirements Elicitation,” Information Processing Society of Japan, pp.1-17, 2009.

[7] Rockford Lhotka, ”Visual Basic 6.0 Business Objects,” Apress, 1998.

[8] Gamma Erich, Helm Richard, Johnson Ralph, Vlissides John, and Grady Booch, ”Design Patterns: Elements of Reusable Object-Oriented Software,” Addison-Wesley Professional, 1994.

[9] Takeshi Chusho, ”Software engineering (3rd edition),” Asakura Shoten, 2014.

[10] Howie Goodell, ”End-user computing,” CHI EA ’97, Mar. 1997.

[11] Information-technology Promotion Agency Japan, ”DX White Book 2021,” https://www.ipa.go.jp/files/000093706.pdf [retrieved: July, 2022].

[12] Microsoft Corporation, ”Differences between Power Apps and traditional app development approaches,” https://docs.microsoft.com/en-us/power-apps/guidance/planning/app-development-approaches [retrieved: July, 2022].

[13] Microsoft Corporation, ”Microsoft Power Platform,” https://powerplatform.microsoft.com/ja-jp/[retrieved: July, 2022].

[14] Microsoft Corporation, ”Microsoft Power Apps,” https://powerapps.microsoft.com/ja-jp/[retrieved: July, 2022].

[15] Microsoft Corporation, ”Power Automate,” https://powerautomate.microsoft.com/ja-jp/[retrieved: July, 2022].

[16] Microsoft Corporation, ”Microsoft Power BI,” https://powerbi.microsoft.com/ja-jp/[retrieved: July, 2022].

[17] Microsoft Corporation, ”Microsoft Power Virtual Agents,” https://powervirtualagents.microsoft.com/ja-jp/[retrieved: July, 2022].

Information Security Policy Awareness Beliefs versus Reality in Electronic Identity Systems

A Case Study of the Ghanaian National Identity System

Salim Awudu, Dr Sotirios Terzis
Department of Computer and Information Sciences, University of Strathclyde,
Glasgow, United Kingdom
{salim.awudu, sotirios.terzis}@strath.ac.uk

***Abstract*— Electronic Identity Systems (EIS) have become a tool for economic, social, and political development in several countries. However, certain concerns by governments or their citizens have impeded wider adoption. These concerns are about EIS trustworthiness, privacy, and security. An effective Information Security Policy (ISP) can be key in addressing these concerns provided staff are aware and understand it provisions. A lot of work has been done on general ISP awareness in organizations, but little attention has been given to ISP awareness in electronic identity systems. Moreover, people's awareness of these policies in organizations is typically measured with instruments that focus on staff beliefs about their knowledge and understanding of ISP provisions rather than their actual understanding and their ability to translate ISP provision into protective behaviors. Staff belief is generally about how staff ISP awareness is typically measured while real awareness is about the prescribed behaviors of the staff. Using the Ghanaian National Identification Authority (NIA) as a case study, this paper examines the relationship between staff beliefs about ISP awareness and the reality of knowing and understanding the prescribed behaviors. A questionnaire study was conducted with scales from literature, which shows that NIA staff beliefs match the reality, despite the lack of a formal ISP and staff training. The study also indicates that a formal ISP and training can enhance staff understanding and confidence in their knowledge. It also shows that for EIS it is important that their ISP considers the organizational context.**

Keywords: electronic identity systems; trustworthiness; information security policy (ISP) awareness; ISP common violations

I. INTRODUCTION

Several countries and organizations depend on Electronic Identity Systems (EIS) to identify, authenticate, and verify their citizens and customers. These systems process identity data about individuals to create value for organizations, businesses and individuals through verified identities [17]. Although these systems are used to achieve economic, social, and political purposes, there are increasing concerns about the security, privacy, and trustworthiness of these systems and the collected personal data [5]. According to Flowerday and Tuyikeze [3] "one important mechanism for protecting organizations' assets is the formulation and implementation of an effective ISP". Staff awareness of ISP provisions and their actual knowledge and understanding of them are key for its effectiveness, especially so for EIS that manage personal data. Effective protection of the sensitive data managed by EIS relies on knowing and understanding what protective behaviors are prescribe by the EIS ISP. However, research to date tends to measure ISP awareness using instruments that measure staff beliefs about their knowledge and understanding. This introduces a risk for EIS in that staff beliefs may not match their actual understanding of the protective behaviors prescribed. Prescribed behaviors are actions or inactions that are specified in the ISPs of organizations. So, for EIS it is essential to ensure that staff beliefs about ISP awareness match their knowledge and understanding in the ISP.

This paper investigates the relationship between staff belief of ISP awareness and the reality of their knowledge and understanding using the Ghanaian NIA as a case study. A questionnaire-based study was conducted using scales from literature for ISP awareness and understanding of common ISP violations for non-managerial staff of the NIA. The study finds that despite the particular NIA setting, where there is no formal ISP and no training provided, staff do not only believe that they know and understand the provisions of the NIA ISP but can identify common ISP violations as violations of NIA ISP. The study also shows that there is some room for improvement in staff understanding and confidence through ISP formalization and training. The study also indicates that, for the ISP of EIS, it is important to also consider the organizational context. Additional studies following alternative research approaches can be used to confirm these findings, while further studies in other EIS organizations can help generalize them.

We begin the paper with related work on EIS and ISPs, in II before we describe our methodology in III, and present our analysis and findings in IV, followed by a discussion in V, and finally conclude the paper with the identification, and directions for future work in VI.

II. RELATED WORK

EIS are systems that are built to collect, process, store and use personal data or information about individuals in a defined area or territory for the purpose of planning or providing services to the people both within the defined territory and beyond [9]. EIS can also be seen as "system[s] that involve the collection of information or attributes associated with a specific entity" [15]. Several countries, including

 ISBN: 978-1-68558-014-8

Ghana, have fully operationalized an electronic identity system.

Despite the potentially immense benefits of EIS, several people have reservations about the potential negative effects they can cause. For instance, Lyon and Bennett note that "once cards are mandatory, then they may be used to single out or even to harass visible minorities and those with alternative lifestyles" [7]. Further concerns are about Privacy, Trustworthiness, Confidentiality, Integrity, and Availability [10], especially as EIS present an appealing target for attackers because of data that they collect, store, and manage. So, information security assurance is essential for EIS.

According to Von Solms [14], information security (IS) is largely multi-dimensional, and organizations must consider all aspects to ensure the security or protection of their information assets and environment. This "includes the physical security of buildings, fire protection, software and hardware, personnel policies and financial audit and control" [16]. Furnell et al. [4] pointed out that employee attitudes and lack of security awareness are the most notable contributors to security incidents. To prevent such incidents, Johnson [6] identified the need for organizations to have an information policy that reflects local information security philosophy and commitments. According to Tryfonas et al. [12] and Canavan [2] an ISP is a set of rules or requirements that are related to information security and enacted by an organization to be adhered to by all, to protect the confidentiality, integrity and availability of information and other valuable resources from security incidents. Organizations need to have an ISP and ensure staff are aware and comply with it, because Sipponen and Vance [11] have shown that violations of security policies occur through user's negligence or ignorance of the ISP provisions. Staff understanding and appreciation of ISP provisions is key.

Although a lot of work has been on ISP awareness, e.g. [1, 11], in general, there has been little attention to ISP awareness in EIS. Moreover, instruments to measure ISP awareness, like the one proposed in [1], tend to focus on the beliefs of staff about their ISP knowledge and understanding without any attempts to investigate whether these beliefs reflect actual knowledge and understanding of the protective behaviors prescribed by the ISP.

In this context, for any organization it is important to investigate whether staff beliefs about their awareness of ISP provisions match their actual knowledge and understanding. This is especially the case for EIS where information security is a necessity.

III. RESEARCH METHODOLOGY

To investigate the relationship between beliefs about awareness and actual knowledge and understanding of ISP provisions in the context of EIS, we pose the following research questions: 1) Do EIS staff believe they are aware of the rules, regulations and responsibilities prescribed by the ISP of their organization? and 2) Do EIS staff appreciate key provisions of their organization's ISP?

To explore these questions, we focused our investigation on the Ghanaian National Identification Authority. Despite this, we believe that other countries with similar digitized EIS like Malaysian, Malawian, Nigeria, among others could potentially associate with the findings of this research work.

A. The NIA

The NIA was established in 2003 with the mandate to issue national ID cards to both citizens and residents as well as to manage the National Identification System (NIS). The Ghanaian Identification System is a digitized one where personal data of citizens and residents are collected and stored. The applicants are issued with a smart card to enable them to prove their identity when accessing basic services like mobile phone Subscriber Identity Module (SIM) card registration and banking services [8]. Currently, the NIA is issuing biometric identity cards throughout the country. While this is ongoing, telecommunication companies and banking institutions are required by law to reregister all customers by demanding the national ID card as proof of identity.

Protecting collected citizens data is seen as an essential part of the NIA mission. So, at the outset the organization established an ISP specifying relevant requirements for its staff and introduced training for them. Over time, some updates to the ISP were deemed necessary and a revision was carried out. However, the revised ISP has not been formally approved and there is currently no information security training provided to staff. This situation is concerning for the security of citizens' and residents' data making the NIA an interesting case study to investigate its staff perceptions and the reality of its ISP awareness.

B. Study structure and Procedures

We designed a questionnaire to solicit the views of NIA staff about their awareness of the ISP provisions and to compare them against their understanding of what constitute typical policy violations. This allowed us to assess whether staff beliefs about their knowledge translate into actual knowledge.

More specifically, the questionnaire comprised two scales, one for ISP awareness consisting of 3 questions and adopted from Bulgurcu et al. [1], and one about common ISP violations with 9 questions adopted from Siponen and Vance [11]. Both scales use a Likert scale from Strongly agree to Strongly disagree. However, although we preserved the actual questions, we adapted them to a uniform 7-point Likert scale, which conforms to Stevens's measurement framework where Likert scale type items are summed or averaged and presented horizontally [13].

In additions to the two scales, we also solicited demographic data from participants (Gender, Age range, Department or Unit, Years of work, Educational background, and Type of employment) to check whether the participants form a representative sample of the organization's employee population as captured in NIA Human Resource Data.

C. Study Procedures

Prior to the study, ethics approval was obtained from our department's Ethics Committee. We also obtained approval from the NIA to engage the staff.

We conducted a pilot study with 10 research students at our department and asked for their feedback, which was used

to improve the study design before the main data collection exercise. This uncovered some minor issues with typographical errors that were corrected.

During the main data collection exercise, we printed and distributed 150 questionnaires randomly to staff of the NIA who are not in managerial positions. After the distribution, 115 questionnaires were returned. Three questionnaires were excluded because they were incomplete.

To ensure fair participation from each unit or department, a distribution formula was used. The distribution formula was primarily developed based on the actual staff strength of each unit or department.

We used paper-based questionnaires to ensure easy access to participants to avoid problems with unstable internet connectivity in some of the districts that the data was collected from. All participants were over the age of 18 and consented to participate in the study.

IV. Data Analysis and Results

To facilitate analysis with the Statistical Package for Social Studies (SPSS) software, the data of the paper questionnaires were entered to Qualtrics, and each participant was assigned a unique identifier. For the analysis, each participant's data was divided into Demographic and Non-Demographic Data. The former describes the profile of the participant, while the latter, the information security questions that are the focus of the research.

A. Participants' Demographics

Table I provides an overview of the participants' demographic data (see column Participant Data) compared with data from the Human Resources department of the NIA (see column Organization Reality). Despite some differences, we consider participants largely representative of the organization's employees.

TABLE I: OVERVIEW OF STUDY DEMOGRAPHIC DATA.

		Participant Data	Organization Reality
Gender	Male	54% (60)	74.0% (172)
	Female	46% (56)	26.0% (61)
Age Range	20-30	51.8% (58)	44.9% (96)
	31-40	38.4% (43)	45.8% (98)
	41-50	8.0% (9)	7.0% (15)
	51-60	1.8% (2)	2.3% (5)
Department or Unit	Human Resources	8.0% (5)	2.0% (9)
	Administration	6.3% (7)	52.0% (112)
	Technology and Biometrics	41. 1% (47)	22.0% (48)
	Operations	31.3% (24)	11.0% (35)
	Finance	3.6% (4)	6.0% (12)
	Internal Control	2.7% (3)	1.0% (3)
	Other	2.7% (3)	3.0% (6)
	Procurement	4.5% (5)	2.0 (5)
Years of Work	Less than 1year	55.4% (62)	32.0% (68)
	1-2years	12.5% (14)	4.2% (9)
	3-9years	4.5% (5)	3.3% (7)
	More than 9years	27.7% (41)	60.7% (130)
Employment Type	Permanent	30.4% (34)	64.0% (137)
	Contract	65.2% (70)	33.0% (73)
	Seconded	4.5% (5)	3.3% (7)

B. Information Security Questions

The information security questions consisted of two scales, ISP Awareness (3 questions) and Most Common ISP Violations (9 questions). We evaluated the reliability of these two scales using Cronbach's Alpha as the measure of reliability. Table II shows the results that both measures have acceptable reliability with Cronbach's Alpha above 0.8, so both are included in the analysis. We look more closely at the results for each scale in the following sections.

TABLE II: SUMMARY OF THE CRONBACH ALPHA

	Cronbach's Alpha	Cronbach's Alpha Based on Standardized Items	No. of Items
ISP Awareness	0.915	0.915	3
Most Common ISP Violations	0.876	0.882	9

C. ISP Awareness

Figure 1 shows how participant responses are distributed for the 3 ISP awareness questions. Most of the staff agree that they know (86%) and understand (80%) the provisions of the NIA ISP and know the responsibilities it prescribes (84%). However, some disagree (9%, 12%, 10% respectively) while others are unsure (5%, 7%, 7% respectively).

Looking more closely at those that disagree, we wanted to see whether there is any commonality in their characteristics. So, we looked at their gender, experience (considering those working for the NIA for 3 or more years as experienced and those less than 3 years as inexperienced), and department or unit. As we can see in Table III, there are more female than

male participants that disagree, with a mix of experienced and inexperienced employees. These participants were also from a range of different departments and units (we do not show these numbers due to space limitations). As a result, there is no clear pattern in the characteristics of these participants that may explain their response.

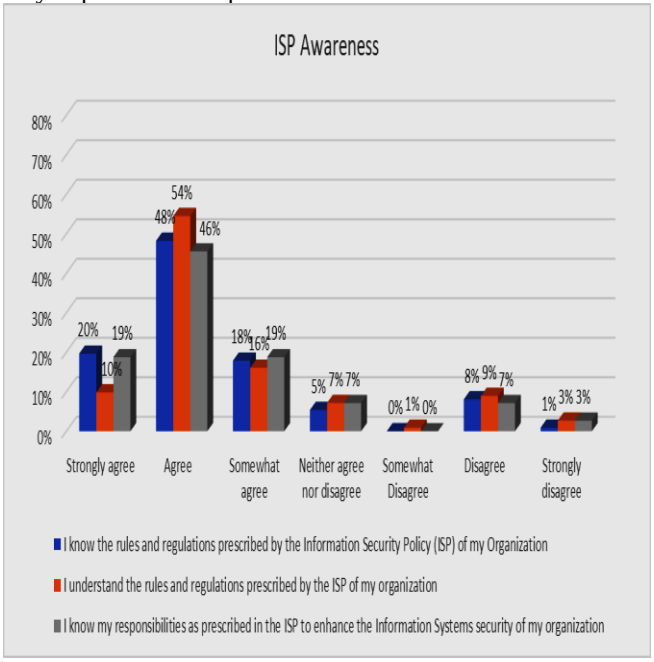


Figure 1: ISP Awareness.

TABLE III: ISP AWARENESS CHARACTERISTICS OF DISAGREEING PARTICIPANT.

ISP Knowledge		
Gender	**Experience**	**Count**
Male	Experienced	1
Female	Experienced	4
	Inexperienced	4
ISP Understanding		
Gender	**Experience**	**Count**
Male	Inexperienced	2
Female	Experience	4
	Inexperienced	4
Knowledge of Responsibilities		
Gender	**Experience**	**Count**
Male	Experienced	1
	Inexperienced	1
Female	Experienced	4
	Inexperienced	2

D. Most Common ISP Violations

Figures 2a, 2b and 2c are distributed for the 9 most common ISP violations questions. We have grouped these questions thematically with Figure 2a showing information transfer-related violations, Figure 2b password-related ones, and Figure 2c workstation-related ones.

From the figures, one can see that in all cases most participants agree that these are NIA ISP violations with information transfer-related violations 95%, 92% and 84%, password-related violations 84%, 89% and 91%, and workstation-related violations 88%, 90% and 91, respectively. Again, some participants disagree 1%, 4%, 8% for information transfer-related violations, 8%, 2%, 5% for password-related violations, and 5%, 6%, 4% for workstation related violations respectively. Other participants are unsure 4%, 4%, 8% for information transfer related violations, 8%, 9%, 5% for password related violations, and 7%, 4%, 6% for workstation-related violations, respectively.

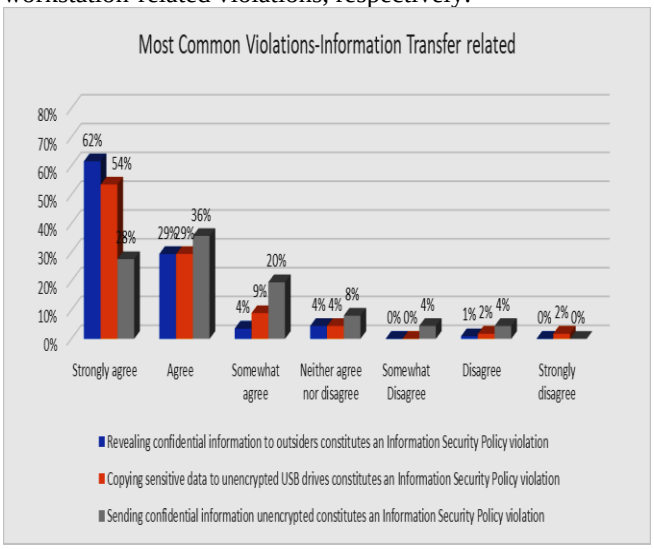


Figure 2a: Most Common Information Transfer-related ISP Violations.

Looking more closely at Figure 2a, Figure 2b and Figure 2c, we note that despite the very high overall agreement with all the violations, there are clear differences in the distributions which indicate differences in the degree of agreement between them. To tease out these differences we treated the Likert scales as ratios from 1 for Strongly agree to 7 for Strongly disagree and we calculated the mean and standard deviation for each of them, see Table IV. The table shows that the means for the common ISP violations range from 1.54 to 2.39 with standard deviations between 0.879 to 1.537. Three of the violations "Creating easy to guess passwords", "Using laptops carelessly outside" and "Failing to lock or log out" have means above 2, 2.39, 2.21, 2.03 respectively, with the former two also having the highest standard deviations, 1.331 and 1.537 respectively. As a result, indicating the agreement to these constitute NIA ISP violation is not as strong as the rest.

Finally, looking at how the means and standard deviations of the common ISP violation questions compare to those of the awareness questions, Table IV shows that the latter have higher means ranging from 2.46 to 2.72 and standard deviations between 1.381 and 1.490. These indicate that agreement with the awareness questions is not as strong to the common ISP violations.

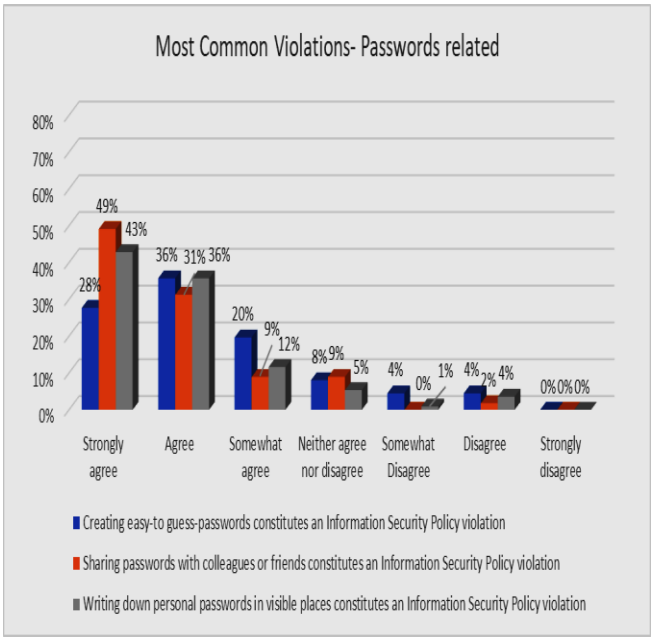


Figure 2b: Most Common Passwords-related ISP violations.

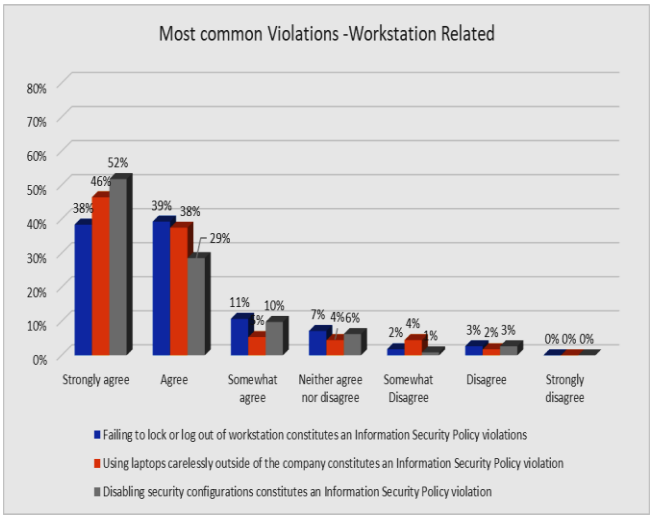


Figure 2c: Most Common Workstation-related ISP Violations.

V. DISCUSSION

Overall, our findings are reassuring for the NIA. Despite the lack of a formal ISP and any staff training on information security, our results show that staff believe they know and understand what is expected from them. More importantly, this is not just their perception. Their understanding of common ISP violations demonstrates that they appreciate their role and responsibilities in protecting NIA information security, an indicator of a good security culture.

That said, there is some room for improvement. First, our results show that for certain violations, NIA staff agreement is not as strong as for others. In some cases, such differences are indicative of typical tensions between security and usability. For example, logging out or locking workstations can slow down work, while coming up with passwords that are not easy to guess can be challenging. In such cases, staff training can help staff identify good strategies for managing these tensions. In other cases, the differences are indicative of lack of clarity in what is and isn't acceptable. For example, NIA staff often use their personal laptops for work and with the absence of a formal ISP they may be unclear of what constitutes careless use. In such cases, a formal ISP with coverage of bring-your-own-device expectations can help staff ensure that, the use of their laptops does not compromise organizational security. Moreover, combining a formal ISP with relevant staff training can also increase the confidence of NIA staff in their knowledge and understanding further strengthening the information security culture of the organization.

TABLE IV: MEANS AND STANDARD DEVIATION OF ALL INFORMATION SECURITY RELATIONS.

	Mean	Std. Deviation
I know the rules and regulations prescribed by the Information Security Policy (ISP) of my organization	2.46	1.381
I understand the rules and regulations prescribed by the ISP of my organization	2.72	1.490
I know my responsibilities as prescribed in the ISP to enhance the Information Systems security of my organization.	2.56	1.475
Failing to lock or log out of workstation constitutes an Information Security Policy violation	2.03	1.174
Writing down personal passwords in visible places constitutes an Information Security Policy violation	1.96	1.193
Sharing passwords with colleagues or friends constitutes an Information Security Policy violation	1.85	1.100
Copying sensitive data to unencrypted USB drives constitutes an Information Security Policy violation	1.80	1.229
Revealing confidential information to outsiders constitutes an Information Security Policy violation	1.54	0.879
Disabling security configurations constitutes an Information Security Policy violation	1.84	1.167
Using laptops carelessly outside of the company constitutes an Information Security Policy violation	2.21	1.537
Sending confidential information unencrypted constitutes an Information Security Policy violation	1.88	1.176
Creating easy-to guess-passwords constitutes an Information Security Policy violation	2.39	1.331

For Electronic Identity Systems, our research reinforces the need for a formal ISP that clearly specifies requirements for staff. It also emphasizes the importance of staff training ensuring that policy provisions are fully appreciated and understood. In addition to this, it highlights the necessity to consider the organizational context in the development and implementation of the ISP.

The main limitation of our research is the focus on staff of the NIA. This limits the generalizability of our findings. Similar studies in other organizations are necessary to generalize them. In addition to this, our research was questionnaire based. This limits the extent to which we can extrapolate from

our findings, the information security behaviors of NIA staff. This would require an observation study to establish whether staff behave in ways to prevent ISP violations. Finally, the research focused on NIA staff in non-management positions. This means that the research is unable to incorporate management's view in the study and whether management views agree with those captured in the study. As a result, our findings are limited to such staff and do not include management views. Surveying management views will address this.

VI. Conclusion and Future Work

We conducted a questionnaire-based study using the Ghanaian NIA as a case. The study shows that although there is no formal ISP and no staff training there is a positive information security culture where staff not only believe they are aware of the ISP provisions but can identity common ISP violations of the NIA ISP. Our study reinforces the need for a formal ISP in EIS and training as the means for clarifying requirements and enhancing staff knowledge and understanding. It also highlights the need to consider the organizational context in the development and implementation of the ISP.

In addition to addressing the limitations above, future work could explore in more detail, the implementation of information security policy at the NIA by looking at how engaged staff are in the development and evolution of its provisions and how compliance is enforced.

Again, similar research works could be replicated in other regions or countries to assess the generalizability of this research findings in other jurisdictions

The impact of covid 19 pandemic also affected the research work. As at the time the data was being collected, the government-imposed restrictions on work attendance and this partly led to rotation of staff attendance. This in effect affected the period for the data collection exercise. In the future, such research work might need to consider longer time due to incorporate such natural occurrences.

References

[1] B. Bulgurcu, H. Cavusoglu, and I. Benbasat, "Information security policy compliance: an empirical study of rationality-based beliefs and information security awareness", *Management Information Systems Quarterly*, pp. 523-548, 2010 https://doi.org/10.2307/25750690

[2] S. Canavan, "An information security policy development guide for large companies." SANS Institute,2003.

[3] S. V. Flowerday, and T. Tuyikeze, "Information security policy development and implementation: The what, how and who", *Computers & Security*, pp. 169-183, 2016.

[4] S. M. Furnell, P. Bryant, and A. D. Phippen "Assessing the security perceptions of personal Internet users", *Computers & Security*, Vol. *26, no.*5, pp. 410-417, 2007.

[5] C. Handforth, and W. Matthew, "Digital Identity Country Report: Malawi" 2019 Available: https://www.gsma.com/mobilefordevelopment/wp-content/uploads/2019/02/Digital-Identity-Country-Report.pdf [Accessed August, 25, 2019].

[6] E. C. Johnson, "Security Awareness: Switch to a better programme", *Network security*, *Vol.*2, pp. 15-18, 2006.

[7] D. Lyon, and C. J. Bennett, "Playing the ID Card: Understanding the significance of identity card systems: Playing the identity card: Surveillance, security and identification in global perspective", pp. 3-20, 2008.

[8] National Identification Authority Act, Government of Ghana, "National Identification Authority Act, 2006 Act 707", In: Ghana, G. O. (Ed.) Act 707. Accra: Parliament of the Republic of Ghana, 2006.

[9] National Identification Authority, "NIA Draft Policy 2014" n.d Accra

[10] B. G. Raggad, "Information security management: concepts and practice", CRC Press,2010.

[11] M. Siponen, and A. Vance, "Neutralization: New insights into the problem of employee information systems security policy violations" *MIS quarterly*, pp. 487-502, 2010.

[12] T. Tryfonas, E. Kiountouzis, and A. Poulymenakou, "Embedding security practices in contemporary information systems development approaches*", Information Management & Computer Security*, Vol. 9, pp. 183-197, 2001.

[13] J. S. Uebersax, "Likert scales: dispelling the confusion" *Statistical methods for rater agreement*, Vol. *31, 2006*.

[14] R. Von Solms, "Information security management: why standards are important", *Information Management & Computer Security, 1999*.

[15] I. Wladawsky-Berger, (2016), "Towards a Trusted Framework for Identity and Data Sharing", 2016. [Accessed October 20, 2019].

[16] C. C. Wood, "Writing infosec policies" *Computers & Security*, Vol. 5, pp. 418, 1995.

[17] World Economic Forum, "Identity in a digital world – A new chapter in the social contract" Cologny/Geneva: World Economic Forum, 2018.

From Clear to Dark: The Social Media Platform Anonymity Continuum

David Kenny
DCU Business School
Dublin City University
Dublin, Ireland
email: david.kenny@dcu.ie

Theo Lynn
DCU Business School
Dublin City University
Dublin, Ireland
email: theo.lynn@dcu.ie

Gary Sinclair
DCU Business School
Dublin City University
Dublin, Ireland
email: gary.sinclair@dcu.ie

***Abstract*—Darknet technologies are a transformative technology, particularly in the context of online social media. This paper explores social media platforms where user anonymity artificially constrains self-disclosure. It proposes a social media platform anonymity continuum that recognises how the emergence and growth of darknet social platforms - and the affordances of darknet technologies - have exposed conceptual limitations in our understanding of self-disclosure and technical and social anonymity on social media platforms.**

***Keywords*— social media; anonymity; dark social; grey social; clear social**

I. Introduction

Social media are ”Internet-based channels that allow users to opportunistically interact and selectively self-present, either in real-time or asynchronously, with both broad and narrow audiences who derive value from user-generated content and the perception of interaction with others” [1]. Social media platforms have been categorised and classified across two dimensions - social presence/media richness (i.e., range of media and content), and self-presentation/self-disclosure [2]. Self-disclosure refers to the conscious and unconscious sharing of personal information, such as ”thoughts, feelings, likes and dislikes” [2] [3]. It also includes ”identity-based” data such as one’s real name, birth date, and image, as well as contact information such as address, phone number and email address [4] [5]. Self-disclosure is pervasive on social media [6] and tightly coupled to the success and vibrancy of social media platforms [5]. The degree to which users engage in it is, in turn, coupled with user anonymity [2] [3] [7] [8]. This paper explores social media platforms where self-disclosure is artificially constrained by user anonymity. These constraints may be imposed by the platform vis-à-vis its technical architecture, its design choices and affordances, its culture or community, and by an individual user’s choice. We recognise the fluidity of online anonymity across technical and social dimensions [9] and propose a continuum for the purpose of categorising social media platforms based on these dimensions (see Figure 1). The continuum recognises the emergence and growth of darknet-enabled, ”anonymity-granting technologies” [10] [11] and how they influence both technical anonymity and social anonymity of users [12] [13] [14]. The remainder of the paper is structured as follows. In Section 2, we present this paper’s main idea: a continuum of social media anonymity. To provide context and background for this, we discuss three categories of social media, and introduce the concepts of dark, grey and clear social. This is followed by a discussion about anonymity in the context of social platforms. In Section 3, the paper concludes with a summary of its achievements and presents preliminary implications and avenues for future research.

II. Continuum of Social Media Anonymity

Social media platforms are categorised as mainstream, alternative, or ”dark”. Mainstream social media - comprising chiefly of the ”social media giants” such as Facebook, Twitter and Instagram [12] - is also referred to as corporate social media (CSM) [13] [15] [16]. CSM operational and business models comprise content moderation, surveillance, commercialising user data, advertising, and lack of user privacy [13] [15] [16]. Alternative social media (ASM) typically operates on the fundamental principle of egalitarianism and caters to smaller communities of niche interests and those who ideologically reject the operational practices of CSM [12] [15] [18] [19] [24]. Current scholarship suggests that social media activity also takes place on ”dark platforms”, on ”dark social networks”, and on ”hidden social spaces” [14] [28] [21]. We refer to this category as dark social - social media that takes place on the darknet. Accordingly, alternative and dark social users turn to decentralised social platforms hosted on privacy-attuned and anonymity-granting technologies such as the darknet (e.g. Tor) [13], and the blockchain [24]. CSM sites primarily exist on the clearnet, although this boundary is beginning to blur as Facebook and Twitter both operate Tor onion services [25] [26] [27]. Further, some clearnet social platforms such as Parler and Gab are built upon ASM-like foundations of anonymity, freedom of expression, and privacy; and are culturally more akin to the dark web. A pure-play darknet social platform provides a combination of technical affordances and ASM-like foundations to support user anonymity and pseudonymity [13]. Anonymity is known to increase self-disclosure [22]. However,we propose that the level of self-disclosure on social platforms is determined by a combination of the technical architecture of the platform (e.g. clearnet versus darknet or blockchain), the culture or community of the platform (CSM versus ASM versus dark - which serves as a proxy for the risk of self-disclosure), and the individual user’s inclination towards online anonymity. This is where the spectrum emerges: from clear to grey to dark. In the context of computer-mediated communication, anonymity is defined as ”the condition in which a message source is absent” where ”an anonymous source is one with no known name or acknowledged identity” [17] [23]. Self-disclosure is an outcome of both anonymity and other affordances of social platforms [23]. The emergence and growth of anonymous social media [7], alternative social media [15], dark social media [12] [13] [16] [18] [19], and the increasingly privacy-attuned design choices of clearnet social platforms [5] [20] is congruent with the Communication Model of Anonymous Interaction in that social platform user anonymity is best viewed as a continuum from fully anonymous to fully identified [15]. This is also reflected in our proposed continuum (Figure 1) which comprises two dimensions: Platform Disclosure Risk (PDR) and Individual Disclosure Practice (IDP). Here, social platforms can be plotted as dark social (pure-play darknet - both technically and culturally); grey social (has some combination of the technical

 ISBN: 978-1-68558-014-8

and cultural affordances of ASM/dark/clear social); and clear social (pure-play CSM/clear - technically and culturally). PDR represents the affordances) of a platform used to prevent or mitigate against the risk of self-disclosure. IDP represents the individual users' behaviour on a platform. The continuum comprises four quadrants:

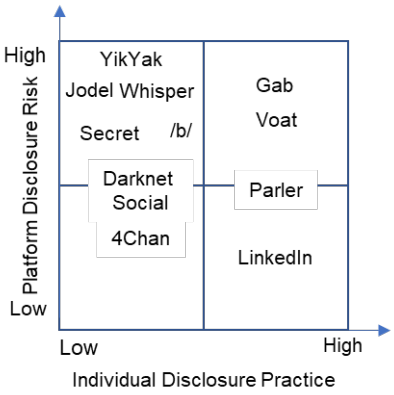


Figure 1: The Social Media Platform Anonymity Continuum

HighPDR-LowIDP: These are social media platforms where there is high risk associated with self-disclosure, particularly of identity. Consequently there are low levels of individual disclosure practice. These dark social platforms effectively discourage or prevent disclosure of individual identity. **LowPDR-LowIDP:** In these dark/grey social media platforms, individuals are not prevented from self-disclosing identity. However, the risk of self-disclosure on these platform ranges from low to high. Platforms such 4Chan/8Chan, and some dark web social networks may straddle this quadrant. **HighPDR-HighIDP:** Here, users disclose identity regardless of the risk presented by these grey/clear social platforms. For example, Gab and Voat became so associated with alt-right, they attracted the attention of hackers [29], and were monitored by US government agencies [30]. Accordingly, these platforms carried reputational risk for participating users (and organisations). Parler sits on the boundaries. **LowPDR-HighIDP:** This quadrant comprises mainstream / CSM (clearnet) social platforms. Accordingly, in characteristrics of these featuring high disclosure practice and low platform disclosure risk.

III. Conclusion, Implications and Avenues for Future Research

This paper critically evaluated the three categories of social media in extant scholarship - mainstream, alternative and dark social - based on their technical and cultural affordances. Consequently, we introduced and defined the concepts of dark social and grey social as part of a proposed spectrum that plots social media as clear, grey or dark. To aid our understanding of these categories, this paper proposed a social media anonymity continuum. This continuum should provide a basis to guide future research efforts to systematically examine, deconstruct, analyse and categorise social platforms along the clear-grey-dark spectrum in the context of the fluidity of two evolving dimensions in online social media: technical anonymity and social anonymity.

References

[1] C. T. Carr and R. Hayes, "Social media: Defining, developing, and divining". In: Atlantic journal of communication 23.1 2015, pp. 46–65.

[2] A. M. Kaplan and M. Haenlein, "Users of the world, unite! The challenges and opportunities of Social Media". In: Business horizons 53.1 2010, pp. 59–68.

[3] J. H. Kietzmann, K. Hermkens, I. P. McCarthy, and B. S. Silvestre, "Social media? Get serious! Understanding the functional building blocks of social media". In: Business horizons, 54.3 2011, pp.241-251.

[4] G. T. Marx. "What's in a Name? Some Reflections on the Sociology of Anonymity". In: The information society 15.2 1999, pp. 99–112

[5] F. Stutzman, R. Capra, and J. Thompson, "Factors mediating disclosure in social network sites". Computers in Human Behavior, 27.1 2011, pp.590-598.

[6] M. Luo, and J. T. Hancock, "Self-disclosure and social media: motivations, mechanisms and psychological well-being". Current Opinion in Psychology, 31 2020, pp.110-115.

[7] D. Correa, L. A. Silva, M. Mondal, F. Benevenuto, and K. P. Gummadi, "The many shades of anonymity: Characterizing anonymous social media content". In Ninth International AAAI Conference on Web and Social Media. 2015, April

[8] P. Nowak, K. Jüttner, and K. S. Baran, "Posting content, collecting points, staying anonymous: An evaluation of Jodel". In International Conference on Social Computing and Social Media 2018, July. (pp. 67-86). Springer, Cham.

[9] T. Sardá, S. Natale, N. Sotirakopoulos, and M. Monaghan, "Understanding online anonymity". Media, Culture Society, 41.4 2019, pp.557-564.

[10] E. Jardine, "Tor, what is it good for? Political repression and the use of online anonymity-granting technologies". New media society, 20.2 2018, pp.435-452.

[11] F. Thomaz, C. Salge, E. Karahanna, and J. Hulland, "Learning from the dark web: Leveraging conversational agents in the era of hyper-privacy to enhance marketing". Journal of the Academy of Marketing Science, 48.1 2020, pp.43-63.

[12] T. Cinque, "The darker turn of intimate machines: dark webs and (post) social media". In: Continuum 35.5 2021, pp. 679–691.

[13] R. W. Gehl, "Power/freedom on the dark web: A digital ethnography of the Dark Web Social Network". In: new media society 18.7 2016, pp. 1219–1235.

[14] J. Zeng, and M. S. Schäfer, "Conceptualizing 'dark platforms'. Covid-19-related conspiracy theories on 8kun and Gab". Digital Journalism, 9.9 2021, pp.1321-1343.

[15] R. W. Gehl, "The case for alternative social media". Social Media+ Society, 1.2 2015, p.2056305115604338.

[16] J. A. Obar and S. S. Wildman, "Social media definition and the governance challenge-an introduction to the special issue". In: Telecommunications policy 39.9 2015, pp. 745–750.

[17] Anonymous, "To reveal or not to reveal: A theoretical model of anonymous communication". In: Communication Theory 8.4 1998, pp. 381–407.

[18] D. Zulli, M. Liu, and R. W. Gehl, "Rethinking the 'social' in 'social media': Insights into topology, abstraction, and scale on the Mastodon social network". New Media Society, 22.7 2020, pp.1188-1205.

[19] R. Rogers, "Deplatforming: Following extreme Internet celebrities to Telegram and alternative social media". In: European Journal of Communication 35.3 2020, pp. 213–229.

[20] K. Waltorp, "Keeping cool, staying virtuous: social media and the composite habitus of young Muslim women in Copenhagen". MedieKultur: Journal of media and communication research, 31.58 2015, pp.49-67.

[21] R. Wiid, P. Hurley, P. Mora-Avila, and J. Salmon, "Organisation-led engagement with consumers in hidden social spaces". Journal of Digital Social Media Marketing, 7.1 2019, pp.53-67.

[22] X. Ma, J. Hancock, and M. Naaman, "Anonymity, intimacy and self-disclosure in social media". In Proceedings of the 2016 CHI conference on human factors in computing systems, May 2016, pp. 3857-3869.

[23] C. V. Clark-Gordon, N. D. Bowman, A. K. Goodboy, and A. Wright, "Anonymity and online self-disclosure: A meta-analysis". Communication Reports, 32.2 2019, pp.98-111.

[24] B. Guidi, A. Michienzi, and L. Ricci, "A graph-based socioeconomic analysis of steemit". IEEE Transactions on Computational Social Systems, 8.2 2020, pp.365-376.

[25] Facebook, "Making Connections to Facebook more Secure". 2021 https://www.facebook.com/notes/2655797467977351/¿ [retrieved: May, 2022].

[26] A. Muffet, "This is possibly the most important and long-awaited tweet that I've ever composed. On behalf of @Twitter, I am delighted to announce their new @TorProject onion service" Twitter: https://twitter.com/AlecMuffett/status/1501282223009542151, 8 April [retrieved: May, 2022]

[27] C. Page, ”Twitter launches Tor service allowing users in Russia to bypass internet blocks”. TechCrunch https://techcrunch.com/2022/03/09/twitter-tor-bypass-blocks/ [retrieved: May, 2022].

[28] W. F. Lawless, F. Angjellari-Dajci, D. A. Sofge, J. Grayson, J. L. Sousa, J.L. and L. Rychly, L, ”A new approach to organizations: Stability and transformation in dark social networks”. Journal of Enterprise Transformation, 1.4 2011, pp.290-322.

[29] A. Greenberg, “Far-Right Platform Gab Has Been Hacked—Including Private Data,” Wired, February 28, 2021, https://www.wired.com/story/gab-hack-data-breach-ddosecrets/. [retrieved: July, 2022]

[30] C. M. Marcos, ”Outcry over US Postal Service reportedly tracking social media posts,” The Guardian, April 23, 2021,https://www.theguardian.com/business/2021/apr/23/usps-covert-program-postal-service-social-media. [retrieved: July, 2022].