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 counterarguments 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 counterattacks. 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.

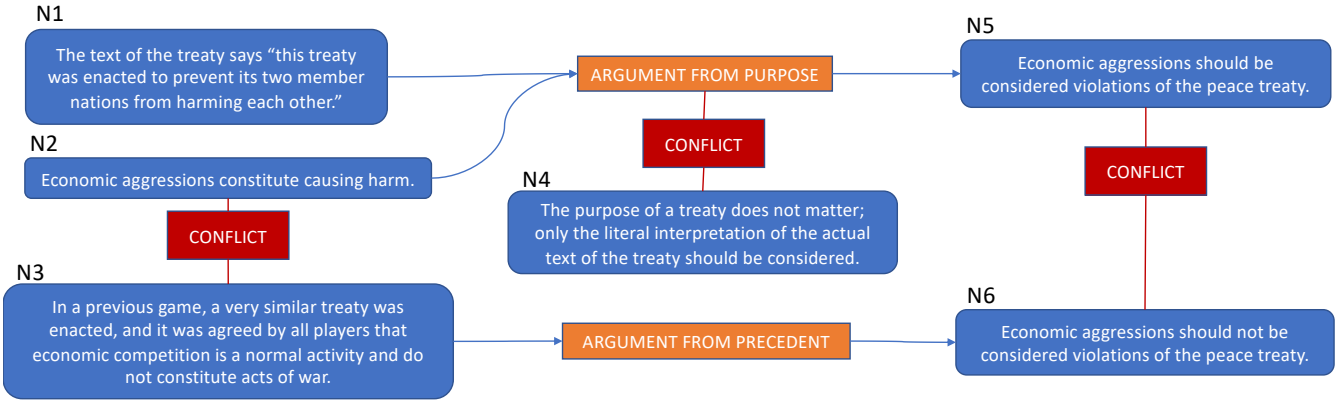
Dung's framework spawned a variety of approaches that extended it, most based on the argument interchange format [33]. Today, the amount of available tools for representing argumentation graphs is continuing to expand (see the review in [34]), particularly with the rate at which progress in natural language 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.

 ISBN: 978-1-68558-014-8

[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

 ISBN: 978-1-68558-014-8

[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.

 ISBN: 978-1-68558-014-8