A Partially Synthesized Position on the Automation of Machine Ethics

Vivek Nallur[1*], Louise Dennis[2†], Selmer Bringsjord[3†] and Naveen Sundar Govindarajulu[3†]

[1*]School of Computer Science, University College Dublin, Ireland.
[2]Department of Computer Science, University of Manchester, United Kingdom.
[3]Rensselaer AI & Reasoning Laboratory, Rensselaer Polytechnic Institute (RPI), USA.

*Corresponding author(s). E-mail(s): vivek.nallur@ucd.ie;
Contributing authors: louise.dennis@manchester.ac.uk;
Selmer.Bringsjord@gmail.com; Naveen.Sundar.G@gmail.com;
†All authors contributed equally to this work.

Competing Interests

The authors declare that they have no competing interests.

Data Availability

No new data was generated or analysed in this paper

1 Introduction

The rapid penetration of software and hardware agents into social contexts that involve making ethically salient decisions has brought to the fore a debate about whether these decision-makers (or recommenders) ought to have ethical-reasoning capabilities. Whether one agrees with the view that machines could one day be, or are even — as Bringsjord and Govindarajulu (= B&G) claim[1] — now, artificial moral agents (AMAs) or not, there is little disagreement that agents imbued with sophisticated pattern-recognition abilities have ethical impact, not only on individuals but also on the social milieu they inhabit. This short position paper does not seek to fully address, let alone resolve, the debate over whether creating an AMA is reprehensible/sufficient/desirable. The paper will also not attempt a survey of all the implementation techniques that claim to have developed an AMA; there are several surveys already [4–6] that cover an appreciable number of the significant implementations; and there are some who hold some of their implementations to be AMAs.[2] Rather, this paper takes as its starting point the proposition that an artificial agent requires (apart from its functional capability) some mechanism to choose between multiple actions/decisions, when all of them are functionally possible. The challenge is to be able to make an ethically informed choice, compatible with whatever human-level ethical theories, codes, or principles are operative in a given context. This paper concerns itself solely with the plausible background frameworks and foreground implementation mechanisms (or desiderata in implementations) that could meet the challenge. A key question that needs a consensus answer is: What constitutes a moral agent? Although James Moor's hierarchy [8] of ethical impact agents, implicit ethical agents, and explicit ethical agents creates a heuristic spectrum for categorization, there is as yet no formal or semi-formal property that serves to define these categories of agents, let alone such a property that is universally affirmed. While this may not be an impediment for system development, it is certainly an impediment for requirements specification regarding an AMA that is ethically correct, or at least safe.

[1] The chief reason most professional philosophers are loathe to accept the proposition that some artificial agents of today or tomorrow are/will be AMAs is that, one, necessary conditions for one brand of such agenthood includes having both phenomenal consciousness and free will, and two, these conditions can't be met by artificial agents. B&G are as a matter of fact of the opinion that artificial agents can't possibly have either of these properties; see e.g. [1, 2]. But that doesn't mean that *some other* brand of artificial moral agents can't be engineered. One such brand, courtesy of B&G, has arrived: a brand marked by *cognitive* consciousness, robust ethical reasoning deeper and and more detailed than what the vast majority of human beings can muster (since e.g. such beings usually can't specify even the difference between, say, act utilitarianism versus rule utilitarianism), and *structural* free will [3]. Note that below (**S**
$sect : toward_s ynthesized_p osition) it's reported that B\&G assert that there is a theorem expressing that some AMA$

[2] E.g., B&G claim that the artificial agent in [7] qualifies.

2 Our Prior Positions & Work, Encapsulated

2.1 Vivek's Position: Hybrid Reasoning and Rule-Breaking

There has been much work in cognitive neuroscience that seeks to explain how the brain perceives the world, reflects on the self, and its place in the perceived world. While the exact mechanisms for each aspect of thinking are still being teased out, there seems to be broad agreement that the fundamental process that enables us to survive (make sense of, and plan for) a dynamic, open-world is *predictive processing* [9]. From our simplest plans about where to find food, to more complicated ruminations about what makes for a good life, all are driven by the need to manage uncertainty.[3] Note that predictive processing does not preclude (arguably non-cognitive) physiological phenomena such as adrenaline-induced flight/fight responses, or *gut-instinct* in our behaviour. Rather, these could be viewed as evolutionary shortcuts or cached strategies, that present themselves when uncertainty levels are high, and conscious, logical thought too slow to map out a good strategy. The notion of a dual-system processing, one for extremely rapid decisions, and another for reflective decisions is also the conclusion of experiments in behavioral economics [10, 11].[4] While the consilience of evidence from neuroscience to behavioral economics lends credence to uncertainty as the prime mover, and the consequent dual-system mechanisms for decisions and choices, it would be fallacious to infer that machines ought to necessarily share the same architecture, or mechanisms for morally salient decisions.

Having acknowledged that, it may still be fruitful to attempt to create an AMA possessing this dual-system, if only to provide us with some insight into the kinds of ethical decision-making that could be expected from autonomous, artificial decision-makers. Bauer [13] advocates a two-level utilitarian agent, since it can conceivably approximate a virtuous agent, as well as be used to implement a fast, rule-based system that as such has inherent explainability. From an implementation perspective, the HERA approach [14] adopts an explicitly hybrid position with respect to reasoning about morally salient decisions. However, unlike Bauer's proposal, it must be noted that the hybrid aspect of HERA is not in implementation/reasoning techniques, but rather in the ethical principles used. That is, instead of choosing a particular ethical school of thought, HERA implements multiple ethical principles, and allows the user/human to choose which reasoning mechanism to utilize. While this is interesting in (semi-automatically) examining how the same situation would be resolved with differing principles, it does not yield any insights into what the robot would do autonomously, if there were conflicting principles/rules.

[3] Of course, some do hold that the finding and management of certainty, which is really what mathematics is fundamentally about, is what enables reliable technology to be created. That uncertainty isn't the coin of the realm in the formal sciences is something not lost on B&G, certainly.

[4] And some computationally sophisticated cognitive scientists, e.g. Ron Sun [12], have long articulated and defended such a position.

Given that our interest is in morally acceptable autonomous decision-making in the face of uncertainty, this approach (unless augmented with a mechanism for autonomously choosing an ethical principle) does not seem to provide a satisfactory solution. We recognize that any mechanism that autonomously chooses between incomparable (or even incommensurable) values [15] could, in spite of our best efforts, be fallible. However, the trade-off for this fallibility ought to be flexibility in allowing for different types of reasoning, *i.e.*, we would like to hedge our bets by using multiple, different techniques, in order for potential exploration of different, rationally correct, decisions. This hedging leads to two significant advantages:

1. **Avoidance of Technological Entrenchment**: Any particular reasoning technique could lead to a state of technological entrenchment, where due to some path-dependent decisions made in the past, future developments become difficult to adopt. Reverting technological decisions is expensive, not least because human beings adapt to machines in their midst. Also, any technical infrastructure that has been built to accommodate a particular type of reasoner could need to be rebuilt, and this has historically proven difficult. If a machine is able to switch between reasoning techniques, then any new (superior) technique that is invented in the future could simply be swapped-in.
2. **Allows Breaking of Rules**: If we agree that there might be multiple decisions possible, we must also recognize that in certain situations, all principled decisions may be at odds with our human instincts. If an autonomous, intelligent entity were to be able to recognize such a situation, how should it behave? Human beings have been observed intentionally breaking rules for both altruistic as well as practical reasons [16]. This not-unusual phenomenon of breaking rules has been identified and defined as *pro-social rule breaking* (PSRB). According to Morrison [17], PSRB usually results from the intention to promote the welfare of one or more stakeholders.

Any decision-making mechanism that can ignore one reasoning technique can decide to ignore all of them as well. This may seem like a disadvantage in an autonomous, decision-making entity, yet it seems to be the only mechanism that brings an AMA closest to human decision-making. This notion of *closeness* need not necessarily be an all-or-nothing proposition, but rather a spectrum which can be evaluated as a guide to decisions about how much autonomy to grant to an intelligent machine. A natural question would be *how* would an AMA decide which principle to ignore, and which principle to use, in a particular situation? While there is no obvious algorithmic procedure that is able to figure out which one is better, Vivek believes that adding stakeholder expectations and preferences, as an additional dimension, would help with decision-making. That is, the AMA must have some notion of the various stakeholders affected by its decision, and an encoding of what their preferences are, with regard to that particular situation. Thereafter, any ethically salient decision must be examined to check whether the immediate stakeholders would

also prefer one of the multiple ethical principles. This ensures predictability in decision-making by the AMA. That is, when confronted with a dilemma (where different principles lead to different recommended actions), the AMA chooses the principle that its immediate stakeholders have expressed a preference for. This allows human stakeholders to express differing preferences for different cases. In Vivek's current investigations into elder-care [18] this has been realized as an important property of decision-making in inter-personal ethics. For example, in the context of a telepresence-robot, decisions about turning video cameras *on* in semi-social settings raises questions about the comparative values of autonomy, well-being, and privacy. A tele-presence robot is usually used in elder-care settings, to enable family members or caregivers to be 'present' with the elderly person, without physically being in the same room. This is usually done using two-way video and audio equipment embedded on to the robot, that follows a person around. Although, these are understood to be poor substitutes for real human interaction, such solutions are increasingly considered for use in healthcare settings. Ethical dilemmas arise when one considers the conflict between a caregivers' need to have access to a patient, while the patient is in the bedroom and expects privacy. Even in social settings, where the patient is conversing with other elderly patients, conflicts arise between the need for privacy and autonomy. In these cases, a purely deontological reasoning process may well reach a different conclusion than a utilitarian reasoning process. Conversations with medical experts and caregivers reveal that specific measurable parameters of the case decide which of these incommensurable values should take precedence, and set expectations about the ethical course of action. Thus, when multiple values conflict (as most often do, in ethical reasoning problems), Vivek believes that taking stakeholder preferences into account, would help to resolve questions about which principle to use in that specific case.

Another natural question arises: Could the dual-system of decision-making affect (or be affected by) the multiple-principles approach? To Vivek's mind, the dual-system approach is an engineering tradeoff that reflects a correlate in humans, parameterized by the time available to make a decision. The core principle, again, is that we have a need to manage uncertainty, and the time available to reduce uncertainty is often limited during decision-making. While the particular units of time required may be quite different for AMAs (*vis-à-vis* humans), the more general problem of ethically salient decision-making under conditions of uncertain outcomes still persists. Therefore, the aforementioned strategy of having cached access to stakeholders' value-preferences would still seem to be reasonable.

2.2 B&G Position: The PAID Problem & Its Four-Step Solution

Bringsjord & Govindarajulu distill the overarching problem they believe constitutes the main challenge of automating machine ethics to "The PAID Problem" (or just 'PAID'), according to which the members of a certain family

of quantified conditionals are true, and this makes for a profoundly worrisome situation. A representative conditional[5] in the family is this one:

Any agent $\mathfrak{a}$ at once **p**owerful, **a**utonomous, and **i**ntelligent is **d**angerous.

For matters at hand, quantification in this biconditional can be assumed to range over *artificial* agents.

What is the solution to PAID? While the complete solution is presented in a pair of forthcoming books [20, 21], the solution can be economically (but, we warn, barbarically) summarized as consisting of "The Four Steps," shown from a high-altitude perspective in Figure 1.

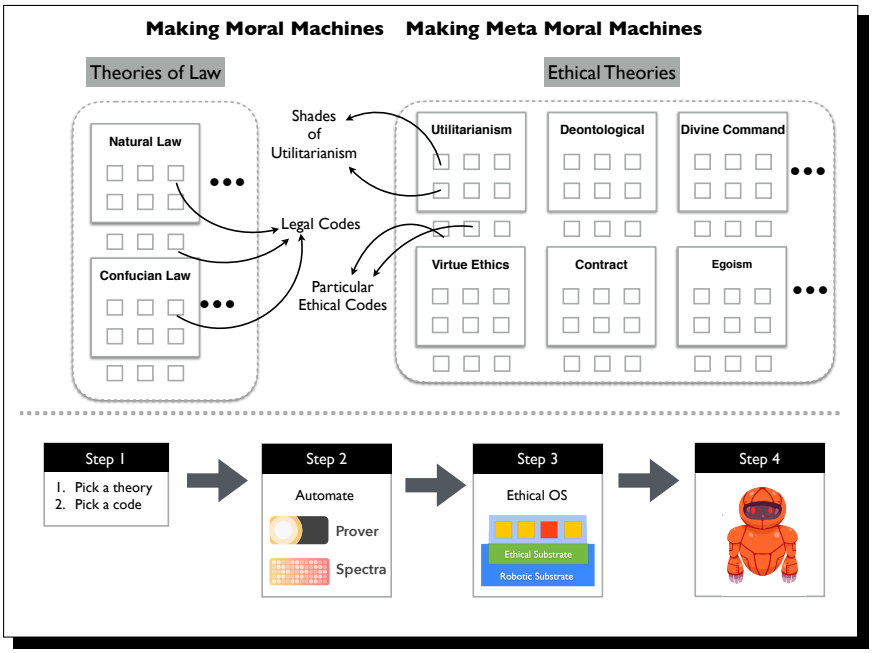

Fig. 1 The Four Steps, for Solving the PAID Problem

Here's a quick summary of The Four Steps:

The first step is the selection of a (formalized) ethical theory (or theories), from a previously selected family of such.[6] The best-known (in the West) families are shown in Figure 1. For instance, one family of ethical theories are

[5] Explication of the family is beyond the scope of the present position paper. We mention only one additional member B&G affirm, one, in which '**d**angerous' is supplanted with and a more dreadful **D** in The PAID Problem, viz. the capability to **d**estroy all of humanity. This more dreadful conditional, in the case of B&G, doesn't presuppose any such notion as that AI of the future will reach superhuman levels. For work that affirms this more dreadful conditional on the strength of the belief that within 80 years superhuman AI will arrive, see [19].

[6] This first step includes not only this selection, but the selection, immediately thereafter, of a particular domain-specific ethical code that falls under the selected theory, a sub-step left aside for economy here. For purposes of general explanation, the reader may take as examples of ethical codes the well-known Nuremberg Code and the U.S. Laws of War (available in an updated 2015 edition, readily findable on the Web).

divine-command; another is *utilitarianism*; a third is *virtue ethics*. An ethical theory that is a member of the second of these families would be standard *act* utilitarianism, according to which, put quickly and without nuance, one ought to always perform those actions from available ones that maximize happiness among the population that can be affected. Another theory in the utilitarian family is *rule utilitarianism*.

For the most part, in the past, we (= B&G) have at one point or another carried out work based on each family shown in Figure 1.[7] For instance, for some prior work that reflects pulling from *both* utilitarian and deontological families, see [7], which centers around our formalization of the so-called *Doctrine of Double Effect*, the intellectual roots of which are in Aquinas, and (B&G claim) ultimately Augustine. DDE contains reference to the valuation of states-of-affairs, and to ironclad prohibitions as well. As to work based upon the family that is probably the most popular on the planet today among the general public,[8] i.e. divine-command, see [23].

Importantly, we refrain from binding our approach to any particular code or theory, or even to particular families of theories. Our framework is general enough that it can be applied to *any* credible ethical theory or code, or collection or family thereof.[9]

However, that said, note that for us (and we hope for all thoughtful, objective humans) every credible ethical theory or code or principle must be *declarative* in nature. E.g., it's impossible to express something like the aforementioned Doctrine of Double Effect (DDE) in non-declarative form. Even the simplest of ethical principles expressed in natural language, for instance "Premeditated murder is morally wrong" or "It's impermissible to steal" and so on, are declarative in nature, and require, once explicated and applied, a deliberative factoring in of local circumstances that aren't included in any data about the past. The fact is, ethics itself as a field has for well over two millennia been constituted by declarative content. There is therefore no way to express an ethical theory or code or principle in terms used by modern statistical/connectionist machine learning (ML). As an example, it is incoherent to assert that an ML-produced computational process is "unfair," in the absence of some declarative definition of what fairness is that is affirmed by at least a significant portion of professional ethicists, and which grounds the assertion. In addition, ethical principles, as indicated, are by definition only applicable when local factors unique to the situation at hand, and absent in any data regarding the past, are expressible in declarative form, and factored in. E.g., the principle "Premeditated murder is morally wrong" is only applicable when local circumstances that have never occurred in the past are factored in (in order to see if premeditation obtained with respect to the relevant agent/s, whose history

[7] B&G have of course in many cases collaborated with other computationally oriented thinkers who have firm, substantive views about machine ethics. See e.g. [22].

[8] And for what it's worth one that at least Bringsjord personally prefers.

[9] Indeed, we are currently exploring the use of The Four Steps with Confucian Ethics, a family not shown in Figure 1. Having mentioned DDE above (and we return to it below as well), we remark that in Confucian Ethics, DDE is rejected.

and particular mental states — if only because perception of the local environment at the moment is singular — will be unprecedented). The upshot of all this is that it's mathematically impossible to automate machine ethics when basing such an attempt exclusively upon ML, to the exclusion of the formal science of declarative content (i.e. formal logic). Moreover, for formal reasons explained e.g. in [24], it is impossible for a computing machine to genuinely learn such ethical principles, e.g. DDE.

Despite our inclusiveness regarding ethical theories and their constituents, there are a few high-level desiderata that need to be satisfied for The Four Steps to be used. Unsurprisingly, these desiderata are rooted in formal logic, and we quickly canvass them now.

To see the first desideratum, assume that we have a family $\mathscr{E}$ of ethical theories of interest. We require that any ethical theory $\mathcal{E} \in \mathscr{E}$ regiments how deontic operators that are invariants across all genuine, robust ethical theories (e.g., *obligatory*, *permissible*, *forbidden*, *supererogatory*, etc.) are to apply to either or both of states-of-affairs and actions performable by agents. In our approach, any ethical theory usable in The Four Steps must be formalized so as to explicitly employ these notions.[10]

As to our second desideratum, formalization must be enabled by a *cognitive calculus*. Details regarding such calculi can be found elsewhere (e.g. see the appendices in [25], and also the new logic introduced in [26]); here we simply inform the reader that (i) such a calculus $\mathcal{C}$ is a pair $\langle \mathcal{L}, \mathcal{I} \rangle$ where $\mathcal{L}$ is a formal language (composed in turn, minimally, of the usual pair composed of a formal grammar, and an alphabet/symbol set), and $\mathcal{I}$ is a collection of inference schemata (sometimes called a *proof theory* or — when non-deductive inferencing is formalized — *argument theory*) $\mathcal{I}$; and (ii) the formal language is extremely expressive, since e.g. (a) it has a full gamut of modal operators that cover knowing, believing, acting, intending, perceiving, being in X where X is some emotional state, and communicating; and (b) the extensional component of the language is often third-order logic.[11]

The second of The Four Steps is to automate the generation of proofs and arguments of (un-)ethical behavior, so that the reasoning can be utilized and acted upon by the relevant artificial agents. Our approach to AI, *logicist* (or logic-based) AI (e.g. see [27, 28]), holds that artificial agents, which compute percepts-to-actions functions, do so via automated reasoning. We specifically use ShadowProver [29, 30], an automated theorem prover for cognitive calculi. We also use automated *inductive* reasoners, which are based on formal inductive logic interpreted purely inferentially [one such reasoner is ShadowAdjudicator (e.g. see [31]], and one inductive cognitive calculus is $\mathcal{IDCEC}^*$. Technical details, including engineering ones, regarding B&G's use of automated-reasoning technology is beyond the scope of the present short

[10] It's probably worth mentioning that those who prefer to speak of *rights*, whether in the moral or legal sense, are in general accommodated in our approach by virtue of the biconditional that an agent $\mathfrak{a}$ has a right to x against agent $\mathfrak{a}'$, say the right not to be harmed, if and only if it's obligatory that $\mathfrak{a}'$ not harm $\mathfrak{a}$.

[11] Note that, in fact, to the expressive reach of a language $\mathcal{L}$ in a cognitive calculus need not be restricted to formulae that are finitely long.

paper, but it is important to mention that such reasoning takes place over novel inference schemata that are present in a given collection thereof $\mathcal{I}$. The technology in question, therefore, is not composed of only fixed, off-the-shelf inferencing such as seen in the case of resolution, which is the complete basis for Prolog and its relatives, and for many extensional theorem provers. Additionally, our automated reasoners take novel inference schemata seriously, in the sense that we do not reduce these schemata to some first-order base, as is (brilliantly) done in Athena [32], an analysis of which by B&G [33] is — for energetic readers — quite relevant to the cognitive-calculus approach.

Step 3 in The Four Steps is to integrate automated ethical reasoning into a logicist artificial agent's operating system (details available in [34, 35]). Without descending into the technical details (doing so not being practicable here), there are basically two possible approaches to doing this. In the first, only "obviously" dangerous capabilities of an artificial agent are restricted with safeguards implemented above the OS level. In the second approach, all agent code must comply with an "Ethical Substrate" that is part of the OS. Unfortunately, while the first approach allows rapid engineering, unforeseen and unwanted unethical behavior on the part of the artificial agent is entirely possible. Only by way of the second option is there any guarantee that the selected ethical theories and associated ethical codes will remain in force, in the face, for instance, of cybercriminals. Unsurprisingly, Step 3 is the taking of the second option.

In the fourth and final step, we implement our content (including, as a proper part, an ethical OS) into an artificial agent and arrive at a morally correct machine. Given this implementation, the actions performed by the class of artificial agents thereby engineered are by definition the conclusion of formal proofs or arguments; and since such proofs and arguments can be formally verified against the background inference schemata that must be used, formal guarantees are readily available. Such formal verification at bottom consists of little more than matching each inference against one or more relevant schemata. For example, in some proof or argument there might be an inference to the formula expressing that given some context χ, an agent $\mathfrak{a}$ is obligated to perform some action α. Leaving temporal information aside to ease exposition, this formula would be an s-expression for the following "pretty printed" inscription:

$$(\dagger) \qquad \mathbf{O}(\chi, \mathfrak{a}, \alpha).$$

Of course, there would be instantiations to what is general in (†). There would be a particular proposition that instantiates χ, say that some human has commanded that the agent perform the action that instantiates α. So, for verification that the agent in question is under the relevant obligation, we simply have the checker ascertain whether or not there is a match for the inference schema whose "input" form expresses the fact that it can be inferred from a human's having commanded that α be performed by $\mathfrak{a}$, that the obligation is in place, i.e. that (†) holds. Fundamentally, this is no different than checking whether in a given automatically found proof in the propositional calculus an

inference made by the artificial agent conforms to *modus ponens*. For details regarding this paradigm of software and agent verification, see [36–38]. For a detailed example of this kind of verification at work, see [31].

Have The Four Steps been implemented such that a *bona fide* artificial moral agent now exists on Earth? Yes — or at least that's our position = the position of B&G. Robust examples (the recapitulation of which here would exceed present space bounds) can be found by consulting other, longer papers; e.g., [7], which we encapsulate now down to a tiny kernel, but in an adapted form inspired by a remarkable collection of papers by Malle and Scheutz et al., in which consideration in empirically studied, ethically charged "trolley-problem" scenarios is given to placing robots themselves in the position of decision-makers [39–41].[12] In the following scenario, we add the twist that the agents destroyed or not are themselves robots.

Let's suppose that a company of the future, trAIn Inc., operates a series of enormous, busy freight-train yards that are massive shipping hubs for all sorts of merchandise transported to and fro by train on many rail lines. The company wishes to have autonomous mobile robots work in its yards, and it specifically wishes The Four Steps to govern these robots. For Step 1, trAIn may select the ethical principle known as the Doctrine of Double Effect (DDE), which of course we mentioned above, to range over robot activity in a particular yard. DDE, at its heart (and this is a very harsh simplification), says that it's ethically permissible for a given agent to harm other agents, as long as certain pre-conditions are met, one prominent one being that the agent that harms doesn't *intend* to harm anyone at all: the intention is just to prevent a catastrophically bad event from coming to pass, and the harm inflicted is a side-effect.[13] DDE is to our knowledge likely the most complex ethical principle to fully logicized, and we can leave aside the ins and outs of the formal logic that captures it; let's just say that $\langle \text{DDE} \rangle$ is the collection of formulae that captures in formal logic DDE. In opting for $\langle \text{DDE} \rangle$, trAIn fulfills Step 1, and now suppose that the formalization of this principle is installed in all its robots in the yard in question — which means that Steps 2–4 are fulfilled as well. Doing this is today more than concretely possible for a freight-train yard. Now, what's the payoff? Suppose for simplicity that there are in fact no humans at all in the yard (there are — let's assume — two human overseers at a distance, with camera feeds and so on). And, suppose in addition that a particular robot, $\mathfrak{a}_r$, perceives that it faces a choice between allowing a runaway freight train to smash five robots out of existence, versus flipping switch that will instead only end the life of one robot. $\langle \text{DDE} \rangle$ is triggered by the percepts in question, and the switch is flipped. This is of course a highly impressionistic depiction,

[12] The key decisions studied by Malle & Scheutz et al. always affects the lives/death of *humans*, while the described decision-makers are either humans or robots. Participants first declare what the decision-maker should do (and those judgments turn out to be pretty similar for human/robot decision-makers), and then they are asked how much blame the human/robot deserves for *actually* deciding one way or another. There they fascinatingly find that participants blame the robot more than the human for inaction (letting 4 people die); they blame them about equally for action (trying to save the 4 but with the known side effect that one will die).

[13] DDE is the basis for so-called "just war" and even for personal self-defense, at least in the Occidental case. For an overview of DDE, see [42].

but with the details worked out and the implementation in place, we would indeed have here a full embodiment of The Four Steps. We see now reason, save for availability of monies and talent, why systems of the sort just described for trAIn should not be required by governments for the likes of today's AI companies working in transport of goods, and people.

We anticipate some readers rightfully wondering whether B&G themselves have any serious misgivings about The Four Steps. The answer is a negative one. However, B&G *are* concerned about an issue that must for economy be left for another day, but which can at least be briefly broached here as food for cerebration in advance of the arrival of that day: viz., that The Four Steps, in and of themselves, will not yield a an artificial agent that can avoid being pulverized (or at least brushed aside) by other artificial agents that are unethical (or even devilish), but more powerful and/or more intelligent than the Four-Step machines that oppose them. This problem requires augmentation of The Four Steps in a manner that ensures its output is agents having superior **p**ower and **i**ntelligence, which would allow these agents to serve as, so to speak, "guardians" of humanity. Such a conception must of course be debated — but not in the present venue.

Some readers may penetratingly ask whether there are irreconcilable frictions between the different ethical theories referenced in Figure 1 — or is it possible to use multiple ethical theories, principles, and codes, and apply The Four Steps in a way that harmoniously selects among and between this content? In reply, it must first be noted that 'frictions' makes for an understatement: the main ethical theories are, as far as B&G can tell after rather considerable effort spent formalizing these theories (and principles and codes stemming from them), pairwise inconsistent. We = B&G are still actively working, but logico-mathematically and implementation-wise, on ways to surmount this serious problem. In general, and confessedly vaguely at this early point in this work, we hope that the "guardian" AIs alluded to in the previous paragraph can be based on a minimalist selection made in Step 1 from the palette of possibilities shown in Figure 1. Nonetheless, we must currently admit regarding the friction issue, humbly, that our solution is currently insufficiently developed to be specified, shared, and recommended to those who wish to have The PAID Problem solved. We note that the work of Vivek, as described in the previous section, is quite relevant to the "friction problem," and we are studying it.

Finally, given current "turbulent" events in AI (e.g., claims on the part of some that some chatbots must not, as a matter of morality, be shut off because they are "sentient") it behooves us (= B&G) to report that we have often heard in person fierce objections against us to the effect that no artificial agent can presently exist, because a moral agent must be *conscious*, and no artificial agents of today are conscious. This objection has no force. The reason, again, is that our artificial moral agents are *provably* such as to have high levels of cognitive consciousness. This brand of consciousness, quite a different type of consciousness than e.g. so-called *phenomenal consciousness*, which many

philosophers discuss but cannot defined in formal, mathematical terms,[14] is explained and axiomatized in prior work [44, 45].

2.3 Louise's Position: The Need for Verification and Validation

As already noted there are multiple approaches to the question of the automation of machine ethics. Different approaches have different features and, it is important to note, the decision about *whether* a system requires explicit ethical reasoning and if so, *how* such reasoning should be implemented is in part an engineering problem. As systems become more sophisticated, and are deployed in under-specified and ill-understood environments, the requirement for more pervasive ethical reasoning within the system increases. Given the current multiplicity of deployment environments for autonomous decision-making systems we must accept that there are some environments wherein ethical reasoning must inform every decision and do so in a highly rigorous fashion, as typified by Bringsjord and Govindarajulu's (= B&G's) approach discussed in the previous section; other systems may require only some kind of governor module [46, 47] controlled by transparent ethical rules that interact with an underlying system that may be programming in a sub-symbolic or hybrid fashion. Such systems often exhibit fast/slow styles of hybrid reasoning. There may even be situations where approaches to ethical reasoning derived from machine learning (as shown in one of the examples discussed in [48]) may be sufficient for our purposes. Where machine learning is used, however, it is critical that suitable training datasets be provided, which raises its own issues. The existence of cultural and other biases in many training datasets is already well-documented, as is the deleterious effect of such biases on the resulting systems [49, 50] — this is a clear challenge to any system that claims to reason ethically. Even if such biases can be overcome, given the current state-of-the-art, it is also unclear what a training dataset for ethical reasoning might look like in a general sense, though application-specific approaches do exist [51].

However in all these cases we need to assess the claim that the system reasons ethically and, given the diverse approaches, diverse techniques may be necessary to assess these claims. For instance, in the case of reasoning backed by logic and underlying automated reasoners (including automated theorem provers), as in B&G's approach, it may seem like no validation is required since the system is *ethical by construction*.

However, even in such cases, we would argue that significant *ethical knowledge engineering* is involved in the construction of the system and this can be a source of error, even where the reasoning itself is correct. Consider an example based on work presented in [52, 53]. If we are using the Principle of Double Effect (called the "Doctrine" of Double Effect = DDE by B&G, above) as our operative ethical principle, then we can not, among other things, take an

[14] Phenomenal consciousness is characterized in [43], a paper that distinguishes that also introduces *access consciousness*, which does bear some resemblance to cognitive consciousness.

action where any of the *intended consequences* are harmful, even if the overall outcome is good. This requires careful engineering of the consequences of actions. For instance a smart home should turn on the lights in order to enable a household evacuation at night, even though using the lights consumes electricity. So we must engineer our system so that it is understood either that the consumption of resources caused by turning on the lights is *unintentional*[15] or that it is not *harmful*. Even where the implementation of the ethical theory is correct and so the ethical reasoning is correct, it is still necessary to validate the ethical knowledge engineering. This can be viewed as a step to avoid *mis-specification* of the ethical rules, principles, situational awareness and so on that are the core components that enable the use of some ethical theory to reason about a concrete situation.

There are a variety of techniques that may be used for validation. Where high levels of assurance are required then formal methods underpinned by mathematical principles should be used both in the design and validation of the system. Formal verification is a validation process for assessing whether a specification given in formal logic is satisfied on a particular formal description of the system in question. For a specific logical property, φ, there are many different approaches to this [56–58], ranging from deductive verification against a logical description of the system ψ_S (i.e., $\vdash \psi_S \rightarrow \varphi$) to the algorithmic verification of the property against a model of the system, M (i.e., $M \models \varphi$). The former approach can be operationalized as part of system reasoning, guaranteeing outcomes are correct according to the logic and specification of the problem. The latter, primarily through the *model checking* approach [59], has the drawback of being inherently finite state and so, at one level, can be seen as a kind of testing. However it still has value for validating ethical knowledge engineering since, among other things, it allows the automatic and exhaustive exploration of multiple pathways through some specific scenario to check that they all meet some formal property.

Lastly there is a role for benchmarking and testing, particularly if we wish to compare the decisions made by competing systems, technologies and ethical theories. The understanding of benchmarking of ethical decisions is in its infancy, consisting at the moment only of small ad hoc collections of examples (e.g., [60]). However the field needs to develop a more mature approach to assessing claims that some system can reason ethically and benchmark sets provide a potential route to achieving this. The development of robust benchmarking suites might also assist with the provision of training datasets for machine-learning approaches.

[15] In addition to this, while there are approaches to capturing a computational concept of *intention* (e.g., [14] and, arguably, the entire field of Beliefs-Desires-Intention agent programming [54, 55]), it is far from clear that there is a philosophically satisfactory formalization of what it means for a machine to intend an outcome, or not intend an outcome that was nevertheless foreseen. Such concepts are critical to some ethical theories so we need to validate the formalization adopted.

3 Discussion: Toward a Partially Synthesized Position

3.1 Points of *Dis*Agreement

Selmer & Naveen (again, i.e. B&G) hold, as indicated above, that a species of genuine artificial moral agents already inhabit Earth; both Vivek and Louise are of the more circumspect opinion that this claim is false, but are open to the possibility that the future may bring such agents on the scene. In fact, B&G affirm the following theorem, where (i) 'artificial agent' is specified to match what this phrase means in the dominant AI textbook of today (viz. [61] which defines an agent as something that maps percepts to actions), (ii) any such agent capable of The Four Steps from above is a *moral* artificial agent, and 'free will' is specified in line with the conception of free will advocated by AI-founder John McCarthy [3] [who advocates grounding free will in a concept of what a system (machine or person) *can* do based on actions they may take in some circumstances even in some given situation it is pre-determined by their programming that they do not]. If one accepts these definitions of agent, moral agent, and free will, then:

Theorem: Artificial moral agents exist today, on Earth.

Louise and Vivek don't contest that the theorem follows from the selected definitions and the nature of B&G's work, but they are dubious that the formal definitions adopted genuinely capture what is meant by many writers when they speak of — to recall the key concept from the outset of the present essay — Artificial Moral Agents (AMAs). One could argue that in the absence of a satisfactory formalization of what it means to be an AMA, the whole discussion is incoherent — much as Turing argued against the use of the question "Can Machines Think?" in his famous article on the Imitation Game [62]. However, history has shown that such questions cannot easily be resolved by proposing a formalizable alternative; hence the authors of the present paper remain divided upon the question of whether the (possible or actual) existence of AMAs is settled or otherwise.

3.2 Points of Agreement

There are autonomous machines, currently operational in the world, that have an ethical impact on human beings; this proposition no one can dispute. Frighteningly, there are even reports of autonomous machines that have been deployed in the battlefield and are making literal *life-and-death* decisions.[16] Given the state of the art in machine-implemented ethical reasoning, the authors would be **extremely** wary of machines being given such high levels of autonomy.[17] Regardless of whether a machine (now or in the future) is

[16] https://www.newscientist.com/article/2278852-drones-may-have-attacked-humans-fully-autonomously-for-the-first-time/

[17] B&G for present purposes are willing to countenance use here of an intuitive concept of autonomy, but note that as formalists, until autonomy is formalized, we can't really *know* that

acknowledged to be an artificial moral agent or not, human beings should not be able to abrogate their own (moral) responsibility for designing machines with ethical impact.

The state-of-the-art is currently unsettled as to which approach would be best suited for designing machines able to carry out genuine ethical reasoning. There are efforts to design machines that are ethically correct by construction, machines that can be validated (and even formally verified) for correctness, and machines that attempt to be ethically correct through non-symbolic approaches. This 'unsettled' status is not a bad thing. It indicates that there is a lot of on-going experimentation, with very different, fertile ideas. Of course, lack of settlement in the overall intellectual landscape does not entail that, in no researcher's minds, the core questions remain unsettled. As cannot be denied given what they said above, things are rather firmly settled in the minds of B&G, for better or worse.

We all agree that a form of ethical reasoning adequate for many applications is formalizable and that that formalization can be implemented, even if it is not yet clear whether a single ideal formalization exists. That being the case, we believe that most current applications of machine decision-making, where such activity has ethical impact, should take a formal route to the implementation of ethical reasoning within that application. Nevertheless, at current technological levels, it is often challenging to identify when some particular ethical principles apply in some concrete situation, and therefore validation processes (such as testing, stakeholder consultation, and post-deployment monitoring) are also currently essential to deliver first-rate machine reasoning in and about ethics.[18]

A pertinent question, at this point, is whether these (positions regarding) implementations are mutually exclusive, or whether they would lend themselves to some form of genetic mixing, so to speak. In Vivek's opinion, the field is still in its infancy, and there needs to be a lot more investigation about the nature of ethical decisions that AMAs are called upon to make, and the expectations that society has of them. A categorization of these decisions and expectations, might well lead to new theoretical approaches to ethical decision-making, particularly ones that are specific to artificial agents. From a computational perspective, incommensurable values combined with feedback loops in second-order effects make for "wicked" problems, ones that will generate entirely new sub-fields of research. From a philosophical perspective, ethical obligations of "lesser" intelligences have not been investigated, and might generate new insights into joint decision-making. A perusal of the current literature on machines implementing ethical reasoning mechanisms reveals that different techniques have been used to tackle very different scenarios. Given the paucity of common problems, with accepted solutions, it is impossible to state

the machines in question are truly autonomous. This borders on self-incrimination, since — recall above — P**A**ID employs the relation ***a****utonomous*.

[18] Notice 'currently' in the previous sentence. Selmer and Naveen call for a future time when only those AI systems that can be formally verified are permitted to have a range of "ethically impactful" actions within their reach.

with any scientific certainty that *a particular* mechanism would be best suited for *all* scenarios, or indeed, if any scenario has a best implementation at all. For the currently explored scenarios, including the ones described above, Vivek agrees with Louise that some form of *ethical knowledge engineering* would be vital to any real-world implementation of machines that attempt to reason about the ethics of particular decisions in these scenarios, and their real-world counterparts. There is some sociological evidence [63] that when confronted with concrete harms, human beings vacillate between ethical principles that are potentially applicable. In fact, this forms the basis of Vivek's misgivings with B&G's Four-Step Process. In all cases of real-world ethical problems, more than one ethical principle may apply, and the Four-Step Process's insistence on first choosing an ethical theory seems to be an incomplete description of the problems in ethical decision-making. This leads to an uncomfortable thought: that there might never be a deterministic way to pick between formalized ethical principles to solve real-world problems. Here Vivek contends that domain-specific interrogations of stakeholder values and their alignment with predicted outcomes using a particular theory may turn out to be a feasible and systematic way forward [18] for ethical decision-making. That is, the autonomic decision-making agent ought to be able predict the localized world-state that would result from its (possible) actions, and then attempt to align the future worlds with *both*, the ethical principles that it knows about as well as the preferred world-states of its human stakeholders. In Vivek's opinion, this attempt at a hybrid form of case-based reasoning is aligned with Louise's call for benchmarking and validation, since it allows for human stakeholder preferences to be the end-point of ethical reasoning.

Acknowledgments

The authors are grateful to three anonymous referees for comments, objections, and suggestions, reaction to which in the prose above — in our estimation — greatly strengthened the paper.

Bringsjord & Govindarajulu are deeply grateful for generous support from ONR to pursue the installation of *bona fide* ethical sensibility in artificial agents that are both **a**utonomous and **i**ntelligent, so as to mitigate the **d**anger such agents clearly pose. They are also grateful to AFOSR (currently under award #FA9550-17-1-0191) for long-term support that has enabled development of automated reasoners for deontic cognitive calculi in which ethical and legal context are — at least in their approach and opinion — best captured.

Dennis' work was partially funded by EPSRC Grant EP/V026801/1 *Trustworthy Autonomous Systems Verifiability Node.*

Nallur gratefully acknowledges funding from both the Science Foundation Ireland via Grant number 18/CRT/6183, and the Irish Research Council via Grant COALESCE/2021/4, to enable research into computational definitions and validation of ethics.

References

[1] Bringsjord, S.: Offer: One Billion Dollars for a Conscious Robot. If You're Honest, You Must Decline. Journal of Consciousness Studies **14**(7), 28–43 (2007)

[2] Bringsjord, S.: What Robots Can and Can't Be. Kluwer, Dordrecht, The Netherlands (1992)

[3] McCarthy, J.: Free will–even for robots. Journal of Experimental and Theoretical Artificial Intelligence **12**(3), 341–352 (2000)

[4] Cervantes, J.-A., López, S., Rodríguez, L.-F., Cervantes, S., Cervantes, F., Ramos, F.: Artificial moral agents: A survey of the current status. Science and engineering ethics **26**(2), 501–532 (2020)

[5] Nallur, V.: Landscape of machine implemented ethics. Science and Engineering Ethics **26**(5), 2381–2399 (2020). https://doi.org/10.1007/s11948-020-00236-y

[6] Wallach, W., Allen, C.: Moral Machines: Teaching Robots Right From Wrong. Oxford University Press, Oxford, UK (2008)

[7] Govindarajulu, N.S., Bringsjord, S.: On Automating the Doctrine of Double Effect. In: Sierra, C. (ed.) Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence (IJCAI-17), pp. 4722–4730. International Joint Conferences on Artificial Intelligence, ??? (2017). https://doi.org/10.24963/ijcai.2017/658. https://doi.org/10.24963/ijcai.2017/658

[8] Moor, J.H.: The nature, importance, and difficulty of machine ethics. IEEE Intelligent Systems **21**(4), 18–21 (2006). https://doi.org/10.1109/MIS.2006.80

[9] Clark, A.: Surfing Uncertainty: Prediction, Action, and the Embodied Mind. Oxford University Press, ??? (2016)

[10] Tversky, A., Kahneman, D.: Rational Choice and the Framing of Decisions. In: Choices, Values, and Frames, pp. 209–223. Cambridge University Press, ??? (2000). https://doi.org/10.1017/cbo9780511803475.013

[11] Kahneman, D.: Maps of Bounded Rationality: Psychology for Behavioral Economics. American Economic Review **93**(5), 1449–1475 (2003). https://doi.org/10.1257/000282803322655392

[12] Sun, R.: Duality of the Mind. Lawrence Erlbaum Associates, Mahwah, NJ (2001)

[13] Bauer, W.A.: Virtuous vs. utilitarian artificial moral agents. AI & Society **35**(1), 263–271 (2020). https://doi.org/10.1007/s00146-018-0871-3. Accessed 2021-05-18

[14] Lindner, F., Bentzen, M.M., Nebel, B.: The HERA approach to morally competent robots. In: 2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), pp. 6991–6997. IEEE, ??? (2017). https://doi.org/10.1109/iros.2017.8206625

[15] Chang, R.: Incommensurability (and Incomparability). In: Lafollette, H. (ed.) International Encyclopedia of Ethics, p. 030. Blackwell Publishing Ltd, ??? (2013). https://doi.org/10.1002/9781444367072.wbiee030. https://onlinelibrary.wiley.com/doi/10.1002/9781444367072.wbiee030

[16] Borry, E.L., Henderson, A.C.: Patients, Protocols, and Prosocial Behavior: Rule Breaking in Frontline Health Care. The American Review of Public Administration **50**(1), 45–61 (2020). https://doi.org/10.1177/0275074019862680. Accessed 2021-05-23

[17] Morrison, E.W.: Doing the Job Well: An Investigation of Pro-Social Rule Breaking. Journal of Management **32**(1), 5–28 (2006). https://doi.org/10.1177/0149206305277790. Accessed 2021-05-23

[18] Ramanayake, R., Nallur, V.: A Small Set of Ethical Challenges For Elder-care Robots. In: Frontiers of Artificial Intelligence and Applications. Robophilosophy Conference Series, University of Helsinki, Finland (2022). https://doi.org/10.5281/ZENODO.6657266. https://zenodo.org/record/6657266 Accessed 2022-06-17

[19] Russell, S.: 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.

[20] Bringsjord, S., Govindarajulu, N., Licato, J.: Logic-based Engineering of Ethically Correct AI and Robots: Making Morally *X* Machines. Springer, Berlin, Germany (forthcoming). This is the large, technical monograph that has a companion book *Only Logic Can Save Us From Powerful-and-Autonomous AI & Robots*, a short version written for the general public.

[21] Bringsjord, S., Govindarajulu, N., Licato, J.: Only Logic Can Save Us From Powerful-and-Autonomous AI and Robots: Making Morally *X* Machines. Springer, Berlin, Germany (forthcoming). This is the short, non-technical monograph that has a companion book *Logic-based Engineering of Ethically Correct AI and Robots*, a much longer, more-technical version written for relevant scientists and engineers.

[22] Bello, P., Bringsjord, S.: On How to Build a Moral Machine. Topoi **32**(2), 251–266 (2013). Preprint available at the URL provided here.

[23] Bringsjord, S., Taylor, J.: The Divine-Command Approach to Robot Ethics. In: Lin, P., Bekey, G., Abney, K. (eds.) Robot Ethics: The Ethical and Social Implications of Robotics, pp. 85–108. MIT Press, Cambridge, MA (2012). http://kryten.mm.rpi.edu/Divine-Command_Roboethics_Bringsjord_Taylor.pdf

[24] Bringsjord, S., Govindarajulu, N.S., Banerjee, S., Hummel, J.: Do Machine-Learning Machines Learn? In: Müller, V. (ed.) Philosophy and Theory of Artificial Intelligence 2017, pp. 136–157. Springer, Berlin, Germany (2018). This book is Vol. 44 in the book series. The paper answers the question that is its title with a resounding No. A preprint of the paper can be found via the URL given here. http://kryten.mm.rpi.edu/SB_NSG_SB_JH_DoMachine-LearningMachinesLearn_preprint.pdf

[25] Bringsjord, S., Govindarajulu, N.S., Licato, J., Giancola, M.: Learning *Ex Nihilo*. In: GCAI 2020. 6th Global Conference on Artificial Intelligence. EPiC Series in Computing, vol. 72, pp. 1–27. EasyChair Ltd, Manchester, UK (2020). https://doi.org/10.29007/ggcf. International Conferences on Logic and Artificial Intelligence at Zhejiang University (ZJULogAI). https://easychair.org/publications/paper/NzWG

[26] Bringsjord, S., Sundar Govindarajulu, N.: Rectifying the Mischaracterization of Logic by Mental Model Theorists. Cognitive Science **44**(12), 12898 (2020) https://onlinelibrary.wiley.com/doi/pdf/10.1111/cogs.12898. https://doi.org/10.1111/cogs.12898

[27] Bringsjord, S.: The Logicist Manifesto: At Long Last Let Logic-Based AI Become a Field Unto Itself. Journal of Applied Logic **6**(4), 502–525 (2008)

[28] Bringsjord, S., Sen, A.: On Creative Self-Driving Cars: Hire the Computational Logicians, Fast. Applied Artificial Intelligence **30**, 758–786 (2016). The URL here goes only to an uncorrected preprint.

[29] Govindarajulu, N.S., Bringsjord, S., Peveler, M.: On Quantified Modal Theorem Proving for Modeling Ethics. In: Suda, M., Winkler, S. (eds.) Proceedings of the Second International Workshop on Automated Reasoning: Challenges, Applications, Directions, Exemplary Achievements (ARCADE 2019). Electronic Proceedings in Theoretical Computer Science, vol. 311, pp. 43–49. Open Publishing Association, Waterloo, Australia (2019). The ShadowProver system can be obtained here: https://naveensundarg.github.io/prover/. http://eptcs.web.cse.unsw.edu.au/paper.cgi?ARCADE2019.7.pdf

[30] Govindarajulu, Naveen Sundar: ShadowProver. https://naveensundarg.

github.io/prover/

[31] Bringsjord, S., Govindarajulu, N., Giancola, M.: Automated Argument Adjudication to Solve Ethical Problems in Multi-Agent Environments. Paladyn, Journal of Behavioral Robotics **12**, 310–335 (2021). The URL here goes to a *rough*, *uncorrected*, *truncated* preprint as of 071421.

[32] Arkoudas, K., Musser, D.: Fundamental Proof Methods in Computer Science: A Computer-Based Approach. MIT Press, Cambridge, MA (2017)

[33] Bringsjord, S., Govindarajulu, N.S.: Review of *Fundamental Proof Methods in Computer Science*. Theory and Practice of Logic Programming **21**(2), 283–290 (2021)

[34] Govindarajulu, N.S., Bringsjord, S.: Ethical Regulation of Robots Must be Embedded in Their Operating Systems. In: Trappl, R. (ed.) A Construction Manual for Robots' Ethical Systems: Requirements, Methods, Implementations, pp. 85–100. Springer, Basel, Switzerland (2015). http://kryten.mm.rpi.edu/NSG_SB_Ethical_Reg_at_OS_Level_offprint.pdf

[35] Govindarajulu, N.S., Bringsjord, S., Sen, A., Paquin, J., O'Neill, K.: Ethical Operating Systems. In: De Mol, L., Primiero, G. (eds.) Reflections on Programming Systems. Philosophical Studies, vol. 133, pp. 235–260. Springer, ??? (2018). http://kryten.mm.rpi.edu/EthicalOperatingSystems_preprint.pdf

[36] Arkoudas, K., Bringsjord, S.: Computers, Justification, and Mathematical Knowledge. Minds and Machines **17**(2), 185–202 (2007)

[37] Bringsjord, S.: A Vindication of Program Verification. History and Philosophy of Logic **36**(3), 262–277 (2015). This url goes to a preprint.

[38] Bringsjord, S.: Computer Science as Immaterial Formal Logic. Philosophy & Technology (2019). https://doi.org/10.1007/s13347-019-00366-7. DOI: https://doi.org/10.1007/s13347-019-00366-7

[39] Malle, B., Scheutz, M., Arnold, T., Voiklis, J., Cusimano, C.: Sacrifice One for the Good of Many? People Apply Different Moral Norms to Human and Robot Agents. In: Proceedings of the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction, HRI'15, pp. 117–124. ACM, New York, NY (2015)

[40] Scheutz, M., Malle, B.: May Machines Take Lives to Save Lives? Human Perceptions of Autonomous Robots (with the capacity to kill). In: Gaillot, J., Macintosh, D., Ohlin, J.D. (eds.) Lethal Autonomous Weapons: Re-examining the Law & Ethics of Robotic

Warfare, pp. 89–102. Oxford University Press, Oxford, UK (2021). https://doi.org/10.1093/oso/9780197546048.003.0007

[41] Komatsu, T., Malle, B., Scheutz, M.: Blaming the Reluctant Robot: Parallel Blame Judgments for Robots in Moral Dilemmas Across U.S. and Japan. In: Proceedings of the 2021 ACM/IEEE International Conference on Human-Robot Interaction, HRI '21, pp. 63–72. IEEE Press, New York, NY (2021). https://doi.org/10.1145/3434073.3444672

[42] McIntyre, A.: The Doctrine of Double Effect. In: Zalta, E. (ed.) The Stanford Encyclopedia of Philosophy, (2004/2014). https://plato.stanford.edu/entries/double-effect

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

[44] Bringsjord, S., Govindarajulu, N.S.: The Theory of Cognitive Consciousness, and Λ (Lambda). Journal of Artificial Intelligence and Consciousness **7**(1), 155–181 (2020). The URL here goes to a preprint of the paper.

[45] Bringsjord, S., Bello, P., Govindarajulu, N.S.: Toward Axiomatizing Consciousness. In: Jacquette, D. (ed.) The Bloomsbury Companion to the Philosophy of Consciousness, pp. 289–324. Bloomsbury Academic, London, UK (2018)

[46] Arkin, R.C.: Governing lethal behavior: Embedding ethics in a hybrid deliberative/reactive robot architecture – Part iii: Representational and architectural considerations. In: Proceedings of Technology in Wartime Conference, Palo Alto, CA (2008). This and many other papers on the topic are available at the url here given. http://www.cc.gatech.edu/ai/robot-lab/publications.html

[47] Bremner, P., Dennis, L.A., Fisher, M., Winfield, A.F.: On proactive, transparent, and verifiable ethical reasoning for robots. Proceedings of the IEEE **107**(3), 541–561 (2019). https://doi.org/10.1109/jproc.2019.2898267

[48] Rossi, F., Mattei, N.: Building ethically bounded AI. Proceedings of the AAAI Conference on Artificial Intelligence **33**(01), 9785–9789 (2019). https://doi.org/10.1609/aaai.v33i01.33019785

[49] Buolamwini, J., Gebru, T.: Gender shades: Intersectional accuracy disparities in commercial gender classification. In: Friedler, S.A., Wilson, C. (eds.) FAT* 18, pp. 77–91. PMLR, ??? (2018). http://proceedings.mlr.press/v81/buolamwini18a.html

[50] Rovatsos, M., Mittelstadt, B., Koene, A.: Landscape Summary: Bias

in Algorithmic Decision-Making: What Is Bias in Algorithmic Decision-making, How Can We Identify It, and How Can We Mitigate It? UK Government, ??? (2019)

[51] Anderson, M., Leigh Anderson, S.: GenEth: A General Ethical Dilemma Analyzer. In: Proc. AAAI-14 (2014)

[52] Bentzen, M.M., Lindner, F., Dennis, L., Fisher, M.: Moral permissability of actions in smart home systems. In: Workshop on Robots, Morality, and Trust Through the Verification Lens (2018)

[53] Dennis, L.A., Bentzen, M.M., Lindner, F., Fisher, M.: Verifiable machine ethics in changing contexts. In: 35th AAAI Conference on Artificial Intelligence (AAAI 2021) (2021)

[54] Rao, A.S., Georgeff, M.P.: An Abstract Architecture for Rational Agents. In: Proc. 3rd International Conference on Principles of Knowledge Representation and Reasoning (KR&R), pp. 439–449. Morgan Kaufmann, ??? (1992)

[55] Mascardi, V., Demergasso, D., Ancona, D.: Languages for Programming BDI-style Agents: an Overview. In: WOA, vol. 2005, pp. 9–15 (2005)

[56] Fetzer, J.H.: Program Verification: The Very Idea. ACM Communications **31**(9), 1048–1063 (1988)

[57] DeMillo, R.A., Lipton, R.J., Perlis, A.J.: Social Processes and Proofs of Theorems of Programs. ACM Communications **22**(5), 271–280 (1979)

[58] Boyer, R.S., Strother Moore, J. (eds.): The Correctness Problem in Computer Science. Academic Press, ??? (1981)

[59] Clarke, E.M., Grumberg, O., Peled, D.: Model Checking. MIT Press, ??? (1999)

[60] Bjorgen, E., Madsen, S., Bjorknes, T., Heimsaeter, F., Haavik, R., Linderund, M., Longberg, P., Dennis, L., Slavkovik, M.: Cake, Death, and Trolleys: Dilemmas as Benchmarks of Ethical Decision-making. In: AAAI/ACM Conference on Artificial Intelligence, Ethics and Society, pp. 23–29 (2018)

[61] Russell, S., Norvig, P.: Artificial Intelligence: A Modern Approach. Pearson, New York, NY (2020). Fourth edition.

[62] Turing, A.M.: I. - Computing Machinery and Intelligence. Mind **LIX**(236), 433–460 (1950) https://academic.oup.com/mind/article-pdf/LIX/236/433/30123314/lix-236-433.pdf. https://doi.org/10.1093/

mind/LIX.236.433

[63] Ramanayake, R., Wicke, P., Nallur, V.: Immune moral models? Pro-social rule breaking as a moral enhancement approach for ethical AI. AI & SOCIETY (2022). https://doi.org/10.1007/s00146-022-01478-z. Accessed 2022-06-15