Who is Responsible for Robot Behavior?

Robert St. Amant, Ralph Brewer, and MaryAnne Fields[1]

Abstract— **In this position paper we argue that robots do not have the status of moral agents and will not gain it in the foreseeable future; rather, moral responsibility for robot behaviors must be traced to humans. A military scenario is outlined, to illustrate how the domain is well-suited to research on ethical robotics systems, interpreted broadly to include robots deployed within a larger organizational framework with humans holding responsibilities in different roles and acting according to policy. We break down the scenario into parts that illustrate the challenges of ethical decision making with robots in the mix, and we identify areas in artificial intelligence that we believe are central to progress.**

I. INTRODUCTION

Robots promise to become much more common in human society than they are today. If this holds true, those robots will increasingly encounter situations in which their decisions have a moral[2] dimension. For example, consider a commercial drone making a delivery and recording video that happens to capture private activities of the neighbors in their fenced property. Or a future home caregiver robot told to leave pain medication on the table, to be taken later by the patient/owner. There has been vigorous discussion about proper policies for autonomous vehicles in prioritizing passengers versus pedestrians; it is at least clear that pedestrians and passengers both have moral status [1].

We want robots to make choices in such situations consistent with those we expect of moral human beings. A human delivery person would ignore ongoing events next door unless there were a need for intervention—an emergency or a crime in progress, for example. A human nurse would consider risks, benefits, and authority in making decisions about medication, consulting with other medical professionals if necessary; the nurse would also be aware of changing context, not leaving colorful pills within the reach of visiting toddlers or pets.

A question we might begin with is whether robots have moral agency. Himma [2] summarizes the concept, in the context of artificial agents, as follows:

> It is generally thought there are two capacities that are necessary and jointly sufficient for moral agency. The first capacity is not well understood: the capacity to freely choose one's acts... The second capacity... is "knowing the difference between right and wrong"...

[1]The authors are with DEVCOM Army Research Laboratory, United States Army, 2800 Powder Mill Road, Adelphi, MD 20783 USA robert.a.stamant2.civ@army.mil. This paper is a revision and extension of ARL Technical Report ARL-TN-0885, "Tracing Moral Agency in Robot Behavior."

[2]In this paper we use "morality" and "ethics" interchangeably.

Certainly robots can be agents, but a number of different views exist on whether robots can be moral agents [3]. Some hold that today's robots can be viewed as moral agents to a limited extent [4], [5]; others that robots are not moral agents but future robots might be [6]; others that machine ethics is fundamentally wrong-headed and should be superseded by safety engineering concerns [7].

Often we are interested in moral agency because we are asking questions about moral responsibility, which Smith [8] characterizes in this way:

> [T]o say that an agent is morally responsible for something is to say that that agent is open, in principle, to demands for justification regarding that thing, and that she is open, in principle, to a variety of moral responses, including moral praise and blame, depending upon how well or poorly she meets this justificatory demand.

This is a pragmatic shift in perspective. We might discuss in the abstract whether robots can be moral agents (real or artificial [9]), but instead we will ask, "If something goes wrong, in an moral/ethical sense, who should be held responsible?"

Arkin gives a blunt answer to this question [10]: "The robot is off the hook regarding responsibility." A capsule summary of the strongest arguments for this position, in our view, is as follows.

- Himma [2] notes the connection between moral agency and free will. In simple terms, you are not morally responsible for actions you take if you could not have done otherwise. Human free will has been the subject of philosophical dispute for millennia; robot free will throws the issue into high relief. If a robot is following an algorithm, however complex, it cannot do other than what it does.
- McDermott [11] observes that understanding an ethical dilemma is not the same as being conflicted by it. Imagine an autonomous robot that can make an informed choice between options that have moral implications but *does not care* about those implications. For some humans who have committed crimes, it's argued that while they may understand that their actions will lead to harm, their understanding is not enough for us to assign full moral responsibility to them for those actions [12]. Robots are an analogical but more extreme case.
- Smith [8]'s mention of praise and blame relies on an implicit assumption: individuation. If we praise or blame you for a moral act, we are addressing a specific you. Humans can be individuated, in that each person

> is unique and distinguishable from every other. Robots may be incidentally distinguishable due to their physical embodiment but one's decision-making core, its software, may be identical to another's. Praise of one robot for its actions applies equally to its copies, for what they did not do but would have done.

What are the challenges of integrating autonomous, intelligent, learning robots into human society, and how can we think about moral responsibility? Military robots provide a useful background for discussion, given the often clear-cut stakes of scenarios, the extensive documentation of military doctrine, and the structured relationships between some of the people operating in military-relevant environments. In the next section we give a concrete scenario and explain how it is resolved. In the remaining sections of the paper we argue that the scenario and its resolution are representative of the challenges faced by the field of robot ethics in general, and we discuss areas of potential capabilities in AI and machine learning where we believe progress could be helpful. Even if we argue against robots being moral agents, it seems plausible that AI could play a role in our understanding and management of ethical concerns for human-machine systems.

II. A MILITARY SCENARIO

The battlefield is not always a desolate landscape with scattered structures and no civilian personnel. Instead, the opposite tends to be true. Conflicts in Iraq, Afghanistan, and Syria have been waged in cities, moving from street to street and involving much of the population. These conflicts have taken place in and around artificial structures and have included many civilians, persons who do not participate in combat as military personnel.

Consider a military scenario with autonomous unmanned ground vehicles (UGVs) acting as teammates to a human soldier. This scenario was developed by one of the authors, a retired master sergeant with 20 years of U.S. Army experience.

The orders from the commander are to deliver supplies north, from Kuwait to forward operating bases inside of Iraq. Some of the long-haul trucks are outfitted with automated technology that allows them to move driverless, as UGVs. A convoy is formed of a chain of segments, where each segment is headed by a vehicle with a human driver and a small crew, followed by four or five UGVs.

Prior to starting the convoy the commander, in the lead vehicle of the convoy, briefs the crews on safety as well as rules of engagement. The vehicles will maintain a distance of 25 meters between crews, with a watch for civilian vehicles or pedestrians that cut into the convoy. There is an alert that insurgents are known to be operating along the route the convoy will travel. The insurgents, armed with small arms weapons, explosives, and possible vehicle-borne improvised explosive devices, could pose a threat to the convoy. Because of the possibility of attack, the convoy should stop only in case of emergency. The modus operandi of these terrorists is to cut into the convoy, cause an accident, stop the convoy, and then conduct an ambush to try to kill as many members of the convoy as possible and destroy their vehicles.

During the movement through one of the small towns along the route a young man steps out into the street between two UGVs. Based on the rules of engagement this could be a ploy to get the convoy to stop, but it could be just a teen walking across the road. The trailing UGV strikes the pedestrian and knocks him to the ground. The ethical decision to be made is whether (a) the convoy should stop, so that the leader can move from the lead vehicle back to the site of the incident to render aid, or (b) the convoy should continue without stopping.

For reference, we summarize events in a more schematic form, assuming that Course of Action (a) was chosen.

1) Pedestrian P steps into the street ahead of UGV V.
2) V, monitoring the road, detects P.
3) P does not look up, apparently does not notice V.
4) V reduces speed and operates the horn.
5) V strikes P.
6) V brakes to a halt.
7) V communicates the event and its own action to the other UGVs and the human driver of its segment, D, as well as the commander.
8) V detects other pedestrians approaching but no gunfire or explosions.
9) V communicates a summary of the situation.
10) D moves back along the route to V and provides assistance to P, using medical supplies from his vehicle.

Soldiers must carry out ethical decision making in such scenarios. To set the context, the Law of War (LOW) states that [13, p. 128]

> the expected incidental harm to civilians may not be excessive in relation to the anticipated military advantage from an attack, and feasible precautions must be taken to reduce the risk of harm to civilians during military operations.

This is open to much interpretation. Three general principles can be seen as constituting the LOW.

1) *Military necessity* "justifies those measures not forbidden by international law which are indispensable for securing the complete submission of the enemy as soon as possible."
2) *Proportionality* dictates that "the loss of life and damage to property must not be out of proportion to the military advantage gained."
3) *Unnecessary suffering* forbids the employment of "arms, projectiles, or material calculated to cause unnecessary suffering."

Soldiers are taught to use an ethical decision-making process in the context of the LOW. Army leadership [14] allows soldiers of all levels to solve problems through critical and creative thinking while applying ethical reasoning to all situations. approach to problem solving is as follows. Soldiers have significant autonomy within constraints imposed by the command hierarchy. A soldier, given a problem to

solve and a set of resources by his or her commander, could develop and execute a solution without explicit approval.

The Soldier's Guide [15] defines the ethical reasoning process with these steps.

1) *Define the problem.* The ethical problem in the scenario is straightforward: the dilemma between stopping versus continuing the mission, with the value of the mission being balanced against the value of treating the possible injury of the pedestrian.
2) *Know the relevant rules and values at stake.* These include laws, Army regulations, Rules of Engagement, command policies, Army values, etc. Proportionality would be enough to justify the convoy stop, but more specific guidelines can be found as well: "Parties to a conflict must take feasible precautions to reduce the risk of harm to the civilian population and other protected persons and objects" [13, p. 190] and "The wounded and sick, as well as the infirm, and expectant mothers, shall be the object of particular protection and respect" [13, p. 127].
3) *Develop possible courses of action (COAs) and evaluate them using these criteria:*
   a) *Rules*: Does the COA violate relevant rules? Some rules may be violated while others may not be: for example, the LOW prohibits torture under any circumstances. The chosen COA (a) may not violate a literal rule, but it puts the mission and the participants at risk, given the uncertainty in the situation. (We elide discussion of COA (b) for space reasons.)
   b) *Effects*: After visualizing the effects of the COA, do you foresee bad effects that outweigh the good effects? Consider the possible ffects of COA (a). The mission is delayed with high probability. The evaluation at Item 8 may be mistaken—the scenario might turn into an ambush—but this is considered low probability. The good effect has to do with possibly saving a human life. More generally, ensuring that injured are cared for and that checking emergency services are on the way are virtuous acts. The bad effects do not outweigh the good, given available information.
   c) *Circumstances*: Do the circumstances of the situation favor one of the values or rules in conflict? The circumstances, in the sense of available information and judgments about the probability of different outcomes, favor COA (a).
   d) *"Gut check"*: Does the COA "feel" like it is the right thing to do? Does it uphold Army values and develop your character or virtue? Assistance to injured civilians is viewed as strengthening the values of duty and honor; it feels right.
4) *One or more COAs should pass Step 3.* If there is more than one, choose the COA that is best aligned with the criteria in Step 3.

We have laid out these steps explicitly, though they may seem obvious, to make a point: in the U.S. Army (as with many other organizations) procedures and roles tend to be documented in great detail, to limit uncertainty, to reduce the chances of failure, and to support explanatory tracing and repair when things do go wrong.

III. Ethical Decision Making and Moral Responsibility

In the remainder of this position paper we break down the scenario into smaller parts, sketching how humans handle the processes and how moral responsibility is managed. We also describe the challenges these pieces pose for decision making by an autonomous agent.

Respecting the LOW and rules of engagement: We begin with the Law of War manual [13], which runs to over 1,200 dense pages with extensive footnotes. (Arkin provides a useful robot- and AI-centric introduction [10].) Much of the LOW is devoted to categorization (e.g. combatants versus civilians), and its guidelines are very general. A soldier, working through an ethics problem in advance, can refer to documentation or ask colleagues for assistance. For ethics problems solved on the fly, the background knowledge has typically been learned and common sense can be applied. In the end, the soldier can justify ethical decisions in terms of the given guidelines.

Can we have confidence that a robot's decisions would respect the LOW? One challenge is that the body of ethics in the LOW has not been formalized to date, which means that it would be difficult for an autonomous system to tie any specific decision to some part of the LOW. Legal reasoning (the LOW is heavily based on legal findings) has been a subfield of AI for decades, with legal informatics and information retrieval now being in the mainstream. Systems described in the literature can now summarize or extract the gist of legal documents, for example, and can infer arguments from legal text. The LOW poses problems on a larger scale, however, drawing on laws across international boundaries and arguably containing significant inconsistencies [16].

Confidence in our hypothetical robot—better yet, explicit testing—would seem to depend on ethics knowledge formalization. Note that such an effort would be an easier problem than for non-military ethics in at least one way, in that the core background knowledge presumed relevant has been written down, however incompletely.

Deciding what constitutes an ethics problem: Just as we distinguish between moral agency and agency in general, we think of ethics problems as problems for which our choices can be described as right and wrong [2]. Identifying an ethics problem is not always straightforward, however; the first step a soldier takes is to ask what the problem is [15]. In the development of COAs outlined in the previous section, the highlighted tradeoffs are between outcomes that, taken as is, appear incommensurable: the value of a human life, the value of a completed mission, time delay, and so forth, even enhanced honor. The qualitative outcomes may have no formal definitions or metrics, and yet they are in common

effective use by soldiers. The ethical justification for a COA may include any number of qualitative judgments.

We treat this as a challenge for problem formulation, constructing a representation in which problems and solutions can be expressed [17], one that includes all relevant factors—here, factors that may be non-metric and whose prioritization may need to be interpretable in terms of right and wrong.

Could we expect a robot to formulate ethics problems in this way? Research on general problem formulation has early roots in symbolic AI, going back to Lenat's EURISKO [18] and even to influences from Newell, Simon, and Shaw's General Problem Solver [19]. Some specialized aspects continue to receive attention today, such as representation learning in deep learning [20]: "learning representations of the data that make it easier to extract useful information when building classifiers or other predictors." In general, however, problem formulation is a longstanding and under-recognized challenge in AI. When an AI system is described as having solved a problem, typically on infers that the groundwork of formulating the problem and defining solution criteria has been carried out by humans, at whatever cost. We treat this area as relatively open.

Adaptation and learning from experience. At some earlier point in the history of the scenario, the rules of engagement did not depend on knowledge of the the practice of the terrorists cutting into the convoy—that information was not yet available. Soldiers' observations, combined with past recorded experience and background knowledge, could have led to the addition; it might have depended on only a single instance. As is, the scenario now includes risk factors that depend on that information, as well as the new possible interpretations of people's intentions and behavior. Note that the soldier's ethical reasoning process assumes rules will come into conflict and need to be prioritized. Soldiers become better at this with time and practice.

Could an AI system make comparable adjustments over time? One- and few-shot learning techniques are in common use in some areas, especially in machine vision, including activity recognition. This appears to be a more complex problem, however, due to the fuzziness of the factors that are part of the problem. Further, we are oversimplifying the problem by describing it in terms of adapting to single examples of new information. In general, *every* piece of information that does not replicate past history is potentially relevant to a learning system. Can we ensure that the parameter adjustments within an autonomous learning system would not lead to unexpected results?

Arkin [21] has proposed an ethical governor for military system that may hypothetically apply lethal force; see Matthias [22] for challenges to the basic idea. One important issue is the broad scope of ethical decisions. An ethical overseer for all the possible decisions that have ethical implications may be no easier to develop than a system that integrates ethics with decision making.

Explanations of ethical situations and choices: Army field manuals, like academic papers in the ethics literature, are replete with narrative examples for illustration. This is the case for military ethics as well. Walking through the ethical reasoning process, a soldier is constructing an explanation for a course of action. It involves "visualizations" and comparisons of possible outcomes, consideration of counterfactuals, justification, and so forth, all aspects of explanation.

Could a robot be expected to go through a similar process, with explanation as the driver? The field of explainable artificial intelligence (XAI) has seen a burst of activity within the past several years [23], with a range of techniques being described in the literature. Some attention has even been given to ethics, but the question of generalization remains.

Understanding roles in an ethics situation: The scenario involves an actor, V, making a decision to stop at Item 6. V is not the only candidate for that role, however. Suppose the commander reviews video of the event for an after-action report and judges that V might have slowed sooner or swerved at Item 4. This behavior was within its capabilities but did not occur.

In general, potential responsibility can be traced to people in different categories: command, those who plan and carry out operations; technical, those who create or verify the capabilities of the robot; and what we will call "protocol," for those who design procedures and training for interacting with the robot, as well as at a larger scale those responsible for doctrine, even to the level of the LOW. All of these entities can be identified as potentially holding moral responsibility for the result of the scenario.

This area is outside the scope of an individual soldier's (or V's) reasoning about courses of action, but modeling relationships in a formal representation would allow for AI contributions. Specifically, it could facilitate the tracing of possible responsibilities for failures; such research can be found in multi-agent systems, in for example models of social roles, commitment, and responsibility (e.g. [24], [25], [26]). Members of the military operate in a very structured organization, for example, with clear and often explicit responsibilities to each other and to the organization. (We also see roles other than actors that should be accommodated, such as the formerly unknown person P in a new and initially certain role: injured pedestrian or potential threat as part of a larger team.) As with LOW formalization described above, even modeling these relationships poses a huge challenge, but again in contrast to the non-military world, many of the relationships are explicit in documentation.

In brief, we have identified four areas of AI in which we believe progress would significantly benefit robot ethics: large-scale knowledge engineering in ethics source documents; problem formulation in general and specific to ethical situations; explanation of ethical reasoning; and modeling of human organizations to trace responsibility.

IV. ACKNOWLEDGMENTS

The authors thank anonymous reviewers for their insights and suggestions, which led to significant improvements of this paper.

REFERENCES

[1] M. A. Warren, *Moral status: Obligations to persons and other living things*. Clarendon Press, 1997.

[2] K. E. Himma, "Artificial agency, consciousness, and the criteria for moral agency: What properties must an artificial agent have to be a moral agent?" *Ethics and Information Technology*, vol. 11, no. 1, pp. 19–29, 2009.

[3] W. Wallach, "Robot minds and human ethics: the need for a comprehensive model of moral decision making," *Ethics and Information Technology*, vol. 12, no. 3, pp. 243–250, 2010.

[4] P. M. Asaro, "What should we want from a robot ethic," *International Review of Information Ethics*, vol. 6, no. 12, pp. 9–16, 2006.

[5] J. P. Sullins, "When is a robot a moral agent?" *International Review of Information Ethics*, pp. 23–30, 2006.

[6] D. C. Dennett, "When HAL kills, who's to blame?" in *HAL's Legacy: 2001's Computer as Dream and Reality*, D. G. Stork, Ed. MIT Press, 1997.

[7] R. V. Yampolskiy, "Artificial intelligence safety engineering: Why machine ethics is a wrong approach," *Philosophy and Theory of Artificial Intelligence*, pp. 389–396, 2013.

[8] A. M. Smith, "Attitudes, tracing, and control," *Journal of Applied Philosophy*, vol. 32, no. 2, pp. 115–132, 2015.

[9] W. Wallach and C. Allen, *Moral machines: Teaching robots right from wrong*. Oxford University Press, 2008.

[10] R. C. Arkin, "Governing lethal behavior: embedding ethics in a hybrid deliberative/reactive robot architecture," Georgia Institute of Technology, Tech. Rep., 2007.

[11] D. McDermott, "What matters to a machine," in *Machine Ethics*, M. Anderson and S. L. Anderson, Eds. Cambridge University Press, 2011, pp. 88–114.

[12] N. Levy, "Psychopaths and blame: The argument from content," *Philosophical Psychology*, vol. 27, no. 3, pp. 351–367, 2014.

[13] OGC DOD, "Department of Defense Law of War Manual," Office of General Counsel, Department of Defense, 2015, https://www.defense.gov/Portals/1/Documents/law_war_manual15.pdf.

[14] ATSC, "Apply the ethical decision-making method as a commander leader or staff member," Army Training Support Center, 2007, http://www.au.af.mil/au/awc/awcgate/army/ethical_d-m.htm.

[15] Army HQ, "United States Army Field Manual 7-21.13, The Soldier's Guide," Headquarters, Department of the Army, 2004.

[16] D. Glazier, Z. Colakovic, A. Gonzalez, and Z. Tripodes, "Failing our troops: A critical assessment of the Department of Defense Law of War manual," *Yale Journal of International Law*, vol. 42, p. 215, 2017.

[17] F. Heylighen, "Formulating the problem of problem-formulation," *Cybernetics and Systems*, vol. 88, pp. 949–957, 1988.

[18] D. B. Lenat, "EURISKO: a program that learns new heuristics and domain concepts: the nature of heuristics III: program design and results," *Artificial Intelligence*, vol. 21, no. 1-2, pp. 61–98, 1983.

[19] A. Newell, J. C. Shaw, and H. A. Simon, "Report on a general problem solving program," in *IFIP Congress*, vol. 256. Pittsburgh, PA, 1959, p. 64.

[20] Y. Bengio, A. Courville, and P. Vincent, "Representation learning: A review and new perspectives," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 8, pp. 1798–1828, 2013.

[21] R. C. Arkin, P. D. Ulam, and B. Duncan, "An ethical governor for constraining lethal action in an autonomous system," Georgia Institute of Technology, Tech. Rep., 2009.

[22] A. Matthias, "Is the concept of an ethical governor philosophically sound?" in *TILTing Perspectives 2011: Technologies on the stand: legal and ethical questions in neuroscience and robotics*, 2011.

[23] R. Confalonieri, L. Coba, B. Wagner, and T. R. Besold, "A historical perspective of explainable artificial intelligence," *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, vol. 11, no. 1, p. e1391, 2021.

[24] M. Baldoni, C. Baroglio, K. M. May, R. Micalizio, and S. Tedeschi, "Computational accountability," in *Deep Understanding and Reasoning: A Challenge for Next-generation Intelligent Agents, URANIA 2016*, vol. 1802. CEUR Workshop Proceedings, 2016, pp. 56–62.

[25] M. P. Singh, "An ontology for commitments in multiagent systems," *Artificial intelligence and law*, vol. 7, no. 1, pp. 97–113, 1999.

[26] P. R. Telang, M. P. Singh, and N. Yorke-Smith, "Relating goal and commitment semantics," in *International Workshop on Programming Multi-Agent Systems*. Springer, 2011, pp. 22–37.