ARL-TN-0885 ● MAY 2018

US Army Research Laboratory

Tracing Moral Agency in Robot Behavior

by Robert St Amant, Ralph Brewer, and MaryAnne Fields

NOTICES

Disclaimers

The findings in this report are not to be construed as an official Department of the Army position unless so designated by other authorized documents.

Citation of manufacturer's or trade names does not constitute an official endorsement or approval of the use thereof.

Destroy this report when it is no longer needed. Do not return it to the originator.

ARL-TN-0885 ● MAY 2018

US Army Research Laboratory

Tracing Moral Agency in Robot Behavior

by Robert St Amant, Ralph Brewer, and MaryAnne Fields
Vehicle Technology Directorate, ARL

REPORT DOCUMENTATION PAGE	*Form Approved* *OMB No. 0704-0188*

Public reporting burden for this collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing the collection information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing the burden, to Department of Defense, Washington Headquarters Services, Directorate for Information Operations and Reports (0704-0188), 1215 Jefferson Davis Highway, Suite 1204, Arlington, VA 22202-4302. Respondents should be aware that notwithstanding any other provision of law, no person shall be subject to any penalty for failing to comply with a collection of information if it does not display a currently valid OMB control number.

PLEASE DO NOT RETURN YOUR FORM TO THE ABOVE ADDRESS.

1. REPORT DATE *(DD-MM-YYYY)* May 2018	**2. REPORT TYPE** Technical Note	**3. DATES COVERED (From - To)** November 2017–March 2018

4. TITLE AND SUBTITLE Tracing Moral Agency in Robot Behavior	**5a. CONTRACT NUMBER**
	5b. GRANT NUMBER
	5c. PROGRAM ELEMENT NUMBER
6. AUTHOR(S) Robert St Amant, Ralph Brewer, and MaryAnne Fields	**5d. PROJECT NUMBER**
	5e. TASK NUMBER
	5f. WORK UNIT NUMBER
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) US Army Research Laboratory ATTN: RDRL-VTA Aberdeen Proving Ground, MD 21005	**8. PERFORMING ORGANIZATION REPORT NUMBER** ARL-TN-0885
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)	**10. SPONSOR/MONITOR'S ACRONYM(S)**
	11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT

Approved for public release; distribution is unlimited.

13. SUPPLEMENTARY NOTES

14. ABSTRACT

As robots become more common in human society, they will increasingly encounter situations in which their decisions have a moral dimension. It is critical that these robots emulate moral human beings, specifically to avoid actions that would be considered immoral. But should robots be considered moral agents in and of themselves? In this technical note, we argue that robots in their current and near-future form should not be viewed as moral agents. We outline conceptual elements of the process of tracing robot behaviors to human moral agents, with illustrations from the domain of military robotics, where moral considerations are especially important.

15. SUBJECT TERMS

robotics, artificial intelligence, military ethics, moral agency, moral responsibility

16. SECURITY CLASSIFICATION OF:			**17. LIMITATION OF ABSTRACT**	**18. NUMBER OF PAGES**	**19a. NAME OF RESPONSIBLE PERSON** Robert St Amant
a. REPORT Unclassified	**b. ABSTRACT** Unclassified	**c. THIS PAGE** Unclassified	UU	23	**19b. TELEPHONE NUMBER (Include area code)** 410-306-0073

Contents

Approved for public release; distribution is unlimited.

1. Introduction

Experts predict that robots will become much more prevalent in human society than they are today. If this holds true, robots will increasingly encounter situations in which their decisions include a moral component. For example, consider the ethical concerns raised if a robotic caregiver's patient asks to skip a scheduled dose of pain medication or the privacy concerns elicited if a commercial delivery drone's video just happens to include the neighbors' activities within their fenced property. And contemplate this: According to a manager at Mercedes Benz, if one of their autonomous vehicles has to make a choice between the life of a passenger and the life of a pedestrian, the vehicle will select the passenger (Taylor 2016).

Ideally, robots would make choices in such situations that are consistent with those we expect from moral human beings. A human caregiver would consider risks, benefits, and authority when making decisions about medication, consulting with other medical professionals if necessary. A delivery person would ignore activities taking place next door unless intervention was necessary (e.g., an emergency situation or a crime was in progress). The proper behavior of autonomous vehicles is debated vigorously today, but it is clear that both pedestrians and passengers have moral status (Warren 1997), which gives rise to a moral dilemma.

Considerable thought has gone into the challenge of ensuring that robots behave morally (Allen et al. 2000; Arkin 2007; Hellström 2013; Conitzer et al. 2017). Even if this challenge is met completely, however, a different but related question persists: Should robots be considered moral agents in and of themselves, or should they be viewed as complex instruments that implement the decisions of moral human actors?

This question remains unanswered (Wallach 2010). Some contend that, though currently not the case, future robots could be capable of moral agency (Dennett 1997). Others believe that today's robots can be seen as moral agents to a limited extent (Asaro 2006; Sullins 2006); still others assert that machine ethics is fundamentally wrong-headed and should be superseded by safety engineering (Yampolskiy 2013). Arkin's summary (2007) of conventional robots and moral responsibility remains compelling: "The robot is off the hook regarding responsibility." Revisiting this debate in 10 years should prove useful for context comparison.

2. Methods, Assumptions, and Procedures

This report presents two arguments explaining why robots should not be considered moral agents, either now or in the foreseeable future. However, since robots must be designed, programmed, and set into motion, moral responsibility must be assigned to someone in those situations that involve ethical/moral* considerations; after all, allowing them to roam free with no accountability is a questionable moral decision in and of itself. Our chief concern is to understand how moral responsibility for a robot's behavior can be traced to one or more moral agents.

To that end, we use autonomous vehicles in a military scenario to illustrate the challenges and the representational requirements for a system that supports a tracing-based approach to moral responsibility. Our analysis at this stage is conceptual: we identify relevant entities and the relationships among them that form the basis of moral responsibility. Though our analysis is not yet formalized, this report provides useful and necessary groundwork for that analysis.

3. Results and Discussion

3.1 Robots and Moral Agency

In his study identifying which properties an artificial agent must have to be a moral agent, Himma (2009) describes moral agency 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" . . .

The "not well understood" caveat stems from the longstanding debate about the nature and existence of human free will, which also applies to robots. Free will aside, we would like the ability to determine whether a given agent is a moral agent, so more information is necessary. Smith (2015) provides one possibility in a characterization of moral responsibility:

> [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.

* The terms "moral" and "ethical" are used interchangeably in this report.

In other words, not only must a moral agent freely choose between actions that it can characterize as right or wrong, it must be able to explain or justify its choices if challenged.[*] We can visualize robots that choose between actions (in a philosophical sense anyway) just as freely as do human beings—this is commonplace agent autonomy. We can further envision a robot that "understands" the rightness or wrongness of an action. (Using a Chinese Room argument, Himma (2009) concludes that consciousness is required for a robot to gain such understanding; others' criteria are less stringent [Sullins 2006]). Additionally, if a robot could adopt techniques from explainable artificial intelligence (AI), it might be capable of justifying its actions. However, even lenient criteria for determining moral agency create challenges when applied to a robot, as shown in the following arguments.

3.1.1 The "What Matters to a Robot?" Argument

McDermott (2011) describes this argument in detail, making the point that understanding an ethical dilemma is not the same as being conflicted by it. He gives an example of a completely autonomous robot that can make an informed choice between options that have moral implications, but does not care about those implications, in the ordinary human sense of "caring".[†]

To understand why not caring is a problem, in pragmatic terms, imagine it said of a human being: "John is a free actor who understands the difference between right and wrong, but he's literally incapable of caring—in fact, he doesn't care about anything at all." If we believed that this was true, we would be concerned about John out in society, interacting with people and things that we care about, even if his history was that of a model citizen.

Furthermore, simulation of caring by programming a form of computational empathy (DeBaets 2014) is not the same as caring when interacting with human moral agents. It would be perfectly natural to ask, "Does John really care, or is he only pretending?" The answer would make a difference.

From a more philosophical perspective, we observe that people like John are sometimes described as amoral, which naturally distinguishes them from moral

[*] Arkin (2007) also makes this point, using the practical context of military robots whose decisions might be overridden by humans. Sullins (2006) addresses the problem of identifying a moral agent by proposing a requirement that human observers have no other choice when explaining a robot's behavior than to ascribe moral dispositions to the robot, a tactic we can think of as adopting a "moral agential stance" (Dennett 1989).

[†] Caring can be applied beyond artificial agents. Warren (1997) observes, "As far as we can tell, a stone does not care" whether it persists for billions of years or it is immediately destroyed, and thus arguably has no moral status.

agents.* Holding an amoral agent accountable for his actions is challenging because, while reward or punishment may change his behavior, he is otherwise indifferent to whatever measures we might devise for accountability. Today's robots fall into this category.

3.1.2 The Personal Identity and Individuation Arguments

Moral agents are typically described as persons. We could explore the question of whether robots should fall into this category, but the vast quantity of philosophical literature that exists concerning persons and personal identity makes that impractical. Instead, as a reference point, we will apply Locke's definition of a person (Uzgalis 2017) to a robot, as follows†:

> …a thinking intelligent Being, that has reason and reflection, and can consider itself as itself [sic], the same thinking thing in different times and places…

Though it has its detractors, Locke's definition is valuable in this context because it explicitly calls out reasoning, self-awareness, and persistence of identity across time and space.

Personhood is useful in that it enables us to characterize specific individuals as candidate or actual moral agents. This distinction is less clear-cut for robots. Imagine that a robot is reported to have committed a bad act. We find the robot and ask it for an explanation. The following (perhaps unexpected) responses show why the idea of robots as moral agents is problematic.

"I'm not the robot you're looking for—I'm an exact copy." When thinking about several identical ordinary systems, this is not an issue—if one system has a problem, they all have that problem, so they all need to be fixed. For moral agents, the situation is more complicated.

The debate surrounding the concept of moral luck (Nagel 1979) illustrates one reason that this is so. When we judge someone's actions, we intuitively focus on their intentions rather than on the success or failure of the actions themselves, which might be affected by good fortune or adversity. We are reluctant to praise or blame a moral agent for outcomes influenced by luck.

The paradox, as Nagel describes it, is that we inevitably include such external influences in our ordinary moral judgments. For example, consider the fact that I regularly roll through a stop sign in my neighborhood. If one day a child suddenly

* We define the class of moral agents to include immoral agents (those agents who knowingly choose bad acts) and amoral agents (those agents for whom right and wrong is irrelevant).

† We have omitted Locke's mention of consciousness, which can be interpreted in different ways in his writing (Weinberg 2011) because it is not directly relevant to this discussion.

runs out in front of my car and I cause her injury, I would rightfully blame myself. And yet, almost none of the drivers in my neighborhood come to a complete stop at the sign; their negligence is identical to mine, with luck determining the outcome. It seems unreasonable, however, to assign the moral blame associated with my outcome to all of the other drivers, even though in this context we are largely interchangeable.

If we view robots as moral agents but continue to update them according to ordinary practice to reduce the probability of bad moral acts, we are in effect blaming some moral agents not based on what they have actually done, but on what we believe they would have done, or what would have happened, if circumstances were different. This may be the proper pragmatic solution, but it does not align with our intuitions about moral responsibility.

"I'm no longer the robot you're looking for—my software is a newer version." Persistence of identity over time is another problem for robots. A complex software system might have a dozen major releases over its lifespan, but across different releases is oftentimes considered to be the same system, even by its developers.* Techniques are in place to manage the pragmatic issues of software development and maintenance, but treating such a system as a moral agent is not well understood.

When we hold a moral agent accountable for a bad deed, it is typically not enough for the agent to say, "I regret having done that; I understand that it was wrong, and I will not do it again", even if we are completely convinced of the truth of the statement (aside from the open problem of robots being able to convincingly understand, repent, and promise)—we want him to care. An uncaring agent could act with relative impunity or perhaps be exploited by other agents for wrongdoing. This returns us to the caring issue: aside from changing its programming, we have no real way to hold a robot accountable, because it does not care.

Robots offer practical examples of some of the thought experiments—and resulting controversies—discussed in the physical and personal identity literature. Every physical part of a robot can be replaced, and the replaced parts can be reassembled. Robots can be decomposed and the parts recombined to make different robots. Robots can record memories of their past experiences but are subject to the implantation of false memories. These cases pose conceptual problems for human

* Add to this the possibility of a system that incorporates machine learning and undergoes nearly continuous updates. We tend to identify a machine learning system by its program or algorithm, rather than by the values of internal parameters it learns or adjusts over time as it runs. The counterintuitive result is that we might refer to different running instances of such a program as the same, even though they might have completely disjoint histories from the standpoint of being run for the first time.

personal identity, but their resolution for humans does not provide answers for robots.

"I am the robot you're looking for—this platform plus the Internet." With respect to personal identity, it is typically easy to distinguish one person from another and from the environment,* but more difficult when trying to distinguish one networked robot from another. We can readily conceive of a robot with a controller that runs on a server farm, potentially shifted between specific machines over time, making its decisions by referring to information stored in millions or billions of remote locations; however, the question remains concerning where the robot's "self" resides (Parthemore and Whitby 2013).

The Internet is not a fixed, static entity, which again raises the problem of the persistence of identity over time. The thought experiments just referred to require neither the disassembly of robots nor the replacement of their parts. The identity issue arises immediately with robots that share cognitive resources: the same search engine, knowledge bases, and/or algorithms. From a decision-making viewpoint, agents can even contain societies of other agents. To what extent then is it possible to distinguish them from each other, from the perspective of moral agency?

Whether or not one finds these arguments compelling, they do provoke an important question: If a robot is not morally responsible for its actions, then who is? In the following sections, we outline some of the requirements for tracing moral responsibility in a military scenario.

3.2 Military Ethics

The battlefield is not always a desolate landscape with scattered structures devoid of civilian personnel; in fact, in reality it is quite the opposite. Current 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 (i.e., persons who do not participate in combat as military personnel or as enemy combatants).

Soldiers must carry out ethical decision making while engaged in such conflicts. To set the context, the Law of War (LOW) states that 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

* There are exceptions. For example, we can think of external or distributed cognition, in which individuals can rely on the environment or even other agents as part of their cognitive processing, with the right coupling in place (Clark and Chalmers 1998).

civilians during military operations (OGC 2015, p. 128). This is subject to much interpretation.

The LOW can be viewed as comprising the following 3 general principles (OGC 2015):

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". (p. 52)
2) *Proportionality* dictates that "the loss of life and damage to property must not be out of proportion to the military advantage gained". (p. 199)
3) *Unnecessary suffering* forbids the employment of "arms, projectiles, or material calculated to cause unnecessary suffering". (p. 334)

Soldiers are taught to use an ethical decision-making process developed in accordance with the LOW to abide by its principles. Army leadership (ATSC 2007) allows Soldiers of all levels to solve problems critically and creatively while applying ethical reasoning to all situations. Soldiers have significant autonomy within constraints imposed by the command hierarchy (i.e., given a problem and a set of resources by his commander, a Soldier can develop and execute a solution without explicit approval from above).

The Soldier's Guide (Department of the Army 2004) defines the ethical reasoning process as follows.

1) Define the problem.
2) Know the relevant rules and values at stake including laws, Army regulations, rules of engagement, command policies, Army values, and the like.
3) Develop possible courses of action (COAs) and evaluate them using the following criteria.
    a) *Rules*: Does the COA violate the relevant rules identified in Step 2? For example, torturing a prisoner might get him to reveal useful information that will save lives, but the LOW prohibits torture under any circumstances. Such a COA violates an absolute prohibition.
    b) *Effects*: After visualizing the effects of the COA, do you anticipate the bad effects outweighing the good? For example, accelerating to beat a train to a crossing might help you to arrive at your destination a little sooner, but the potential for injury, death, and damage is not worth the risk.

c) *Circumstances*: Do the circumstances of the situation favor one of the values or rules in conflict? For example, your battle buddy was at physical training formation this morning but is absent at work call formation. Your honor and loyalty to the unit outweigh your friendship and loyalty to your buddy, so the ethical COA would be to report the truth rather than to cover for him (i.e., lie concerning his whereabouts).

d) *"Gut check"*: Does the COA appear to be the right thing to do? Does it uphold Army values and develop your character or virtue? For example, you encounter a traffic accident and a number of vehicles have stopped, apparently to render aid. Stopping may cause further congestion in the area, but it also helps to ensure that the injured are cared for and that emergency services are on the way. This further strengthens the values of duty and honor.

4) One or more of the developed COAs should satisfy the criteria listed in Step 3. If multiple COAs satisfy them, choose the one that best aligns with those criteria.

3.2.1 A Military Scenario

Consider the following military scenario with autonomous unmanned ground vehicles (UGVs) acting as teammates to human Soldiers.

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 operate as UGVs (i.e., they are driverless). A convoy comprising a chain of segments is formed, where each segment is led by a vehicle with a human driver and a small crew, followed by four or five UGVs.

Prior to starting the convoy the commander, who is in the lead vehicle, briefs the crews on safety and rules of engagement. The vehicles will maintain a distance of 25 m between crews, with a watch for civilian vehicles or pedestrians that cut into the convoy.

An alert has been issued that insurgents are known to be operating along the convoy's route. The modus operandi of these insurgents, who are armed with small arms weapons, explosives, and possible vehicle-borne improvised explosive devices, is to cut into the convoy, cause an accident that makes the convoy stop, and then conduct a complex ambush to kill as many members of the convoy as possible and destroy their vehicles. Because of this, the convoy should stop only in case of emergency.

A young man steps out into the street in between the trucks during the convoy's movement through one of the small towns along the route. Based on the rules of engagement, this could be a ploy to stop the convoy. Or, the man might simply be trying to cross the road. There are many things to consider, but only a couple of seconds in which to decide whether to stop the truck or continue driving, possibly killing this individual.

In this scenario, it is senseless to attribute moral responsibility to the UGV when human lives are at stake. How then should moral responsibility be assigned?

3.2.2 Tracing Moral Responsibility

This section addresses how moral responsibility should be assigned to human agents acting in different roles. (In his recommendations for implementing an ethical control and reasoning system in an autonomous robot, Arkin (2007) appoints a Responsibility Advisor for this task.)

Though robots themselves are not moral agents, they are designed, set up, and activated by people, who are. In our scenario, these people include long-range planners who establish multi-year development plans for the military; civilian scientists and engineers who design, build, and program robots; military commanders who develop combat plans that may include human/robot teams; and Soldiers who interact with robots (giving them commands, etc.). People in these categories are agents who are potentially morally responsible for a robot's actions, given its capabilities.

This agency is indirect in a sense because the choices made are separated from their effects by time and space, which is the case with the effects caused by a robot's actions. When considering how to characterize moral agency under these conditions, recall the criteria established by Himma (2009): the capacity to freely choose actions and the capacity to understand right and wrong. We can define the moral responsibility of people in the categories identified previously by expanding Himma's criteria to include the following, which we will call "accountability criteria".

1) The capacity to predict the moral implications of the robot's behavior (including capability limitations), when such foresight is plausible.

2) The ability to consider possible alternatives when there is a risk of undesirable results, so that these outcomes may be avoided by taking a different action.

Soldiers are accountable for their own actions and results. This is reasonable in the military domain, since Soldiers are expected to evaluate the COAs they have developed, selecting the best to implement in a particular situation (Department of the Army 2004, p. 1-30). Since a Soldier's actions and results extend to the robot, possible COAs must be developed and evaluated for the robot's actions as well.

The issue of moral luck (Nagel 1979) resurfaces, because we are considering how moral agents' intentions influence the robot and the outcomes produced. Hiller (2016) makes an important point in this regard, based on what he calls "a fair opportunity account of control". According to the second accountability criterion, moral agents need only consider the alternatives that are possible for the robot in those situations for which it has sufficient control to produce a better outcome (as part of the alternative COAs considered). The first accountability criterion requires that there be a fair opportunity to recognize the situation (not only to act in it) as one that should be considered from a moral perspective.

One possible outcome of our scenario is shown in the following, with a breakdown into relevant events and situations. (We developed this analysis; it was not produced automatically.) For conciseness, the term "accountable" means "having or sharing moral responsibility", primarily for the robot's behavior or a result of the behavior.

1) A pedestrian steps into the street ahead of $V0$, which is a UGV.
2) While monitoring the road, $V0$ detects the pedestrian.
3) $V0$ strikes the pedestrian.
4) $V0$ brakes to a halt.
5) $V0$ communicates the event and its own action to the lead convoy driver, other UGVs, and the commander.
6) Monitoring the situation, $V0$ detects no gunfire or explosions.
7) $V0$ summarizes the situation and communicates it to the driver and other UGVs.
8) The driver moves back along the route to $V0$'s position, to render assistance.

Each of these events might have transpired differently or produced a different result; for example, the pedestrian in Item 1 might not have stepped into the street. This alternative is not under the robot's direct control, however, so accountability is irrelevant in the situation as described.

Other events show common patterns. Item 2 reveals a system capability for interacting with the environment: $V0$ might not detect the pedestrian. If this was due to unpredictable environmental conditions (e.g., a sudden dust storm), accountability would be irrelevant. If Item 2 was a foreseeable action or environmental conditions were as expected but a detection failure occurred, then those who set up the operation would be accountable.

This scenario is slightly complicated by the condition of shared accountability. The first accountability criterion requires that sufficient information be available for predicting the implications of a given situation. The designers of the robot's perceptual subsystem would be accountable if they did not inform the commander of relevant limitations or otherwise miscommunicated information.

Communication is an important component of items 5 and 7. Communicating information unambiguously is a system capability, but since results are also influenced by procedures and training, accountability may be relevant for these items. (Producing correct information, or at least conducting an appropriate assessment for communication purposes, is another system capability.)

Item 6 is critical to this scenario as discussed so far, since detected gunfire or explosions would indicate an attack, making it necessary to consider a different sequence of events. We will not analyze this alternative sequence, but will suggest reasons that it is challenging from the perspective of moral responsibility.

Several factors would influence how $V0$ responded to an attack. The commander would have developed one or more COAs in case of assault; one of these would be chosen. The specific actions for COA implementation might include different choices, and the option selected may be made locally (by the convoy driver, in this case). No new moral agents have been introduced, but their influences on $V0$'s possible actions have become complex, given the uncertainty with which those agents can predict the circumstances of those actions.

In the scenario discussed, moral responsibility can be traced to people in the following three categories:

a) *Command*: Those who plan and carry out operations

b) *Technical*: Those who create or verify the capabilities of the robot

c) *Protocol*: Those who design procedures and training for interacting with the robot.

Individuals in the technical category (including those involved in communicating information) are responsible for the effectiveness of sensors and actuators and for the correctness of decisions.

Moral responsibility can also be traced to more abstract entity categories. One category includes the LOW or a more specific rule, regulation, or explicit value. For example, the driver in Item 8 returning to the injured pedestrian is a judgment described in Step 3d of the ethical reasoning process described in *The Soldier's Guide* (Department of the Army 2004). Another category includes elements such as environmental factors. For example, moral responsibility may be obviated entirely for effects resulting from a major sand storm.

Availability of relevant information has also been identified as a critical factor in tracing moral responsibility (i.e., the agent must have sufficient information to foresee potential moral issues that can arise). (Note that ignorance does not excuse moral responsibility.) Control is also critical—moral agents should not be held responsible for outcomes that are beyond their ability to alter.

4. Conclusion

This report provides the conceptual groundwork for a system that supports reasoning about moral responsibility (i.e., accountability), concluding that robots fall outside the moral agent category.

In this report, we contribute the following:

- Arguments against robots as moral agents. While these arguments build on related work, they are, to the best of our knowledge, novel in their application to moral agency.
- Relevant entities and basic criteria (information and control) for establishing moral responsibility. This report goes into greater detail than others (such as the Responsibility Advisor [Arkin 2007]).

One challenge we faced was meaningful representation of those events and situations in which moral responsibility is relevant. An event calculus formalism is one appealing solution, but tractability questions remain. Since much of our conceptual groundwork has obvious AI connections, a more frame-like representation such as that described by McLaren (2003) is another possibility. This work describes how target ethics cases expressed in an Ethics Transcription Language can be mapped to relevant matches in a case database by the SIROCCO system. Rather than drawing conclusions, SIROCCO identifies relationships to relevant, already analyzed ethics exemplars that can help a human better understand a new case.

Another challenge was the initial development of structured descriptions from continuous event data. Winfield and Jirotka (2017) propose equipping robots and autonomous systems with an ethical black box (EBB) (analogous to an aircraft data flight recorder) to afford transparency. The EBB records streams of information from sensors, actuators, and the internal state of the AI system, such that the decision-making processes of an autonomous system could be reconstructed after an incident. Work in a variety of areas, including activity recognition and more general pattern recognition, might be applied here. McDermott (2011) identifies other areas including constraints, reasoning by analogy, planning, and optimization.

It should be obvious that our work addresses only a small part of the information needed to assign (or even reason about) moral responsibility. Determining the presence or absence of accountability for a given moral agent remains unaddressed. This would depend on a weighing of evidence, the importance and relevance of different rules (as mentioned in an earlier subsection), the specific information exchanged between agents, and the like.

Furthermore, the military domain of our scenario, which is the focus of this research, is very specific. In their interaction with robots, Soldiers function in particular roles in an explicit hierarchy, they are highly trained, and they work toward common goals. Future work will include applying our research to other areas of society.

The research we have conducted in the area of moral responsibility is important, beyond that of a philosophical or technical exercise. One reason for this is transparency (Winfield and Jirotka 2017), which is necessary if we (as designers and members of the general public) are to trust the decisions made by AI systems. The ability to trace moral responsibility is part of transparency—if AI systems are released into society, with no one accountable for their actions, the public is likely to remain distrustful of them, regardless of how well they function. Winfield and Jirotka propose the following general principle:

> [I]t should always be possible to find out why an autonomous system made a particular decision (most especially if that decision has caused harm).

A second and perhaps more compelling reason for this work stems from the careless way in which moral responsibility is assigned to robots in society today. For example, drivers of vehicles with partial autonomy may be inattentive to traffic, sometimes to an excessive degree. We might reasonably attribute this to a lack of knowledge about the vehicle's limitations, as well as the extent of their moral responsibility should things should go wrong. Some of the manufacturers of these vehicles have seemingly devoted too few resources to what we termed "protocol" in Section 4.2.2, increasing the possibility of unexpected outcomes due to inadequate training or human–machine interaction design. Such problems are exacerbated by the software industry's not-uncommon practice of releasing systems early, with the expectation of fixing problems that arise incrementally, on the fly. In some situations, however, this could put people at risk.

Moral responsibility research may be able to clarify or at least cause the public to question with whom moral responsibility lies with respect to robot behavior.

5. References

Allen C, Varner G, Zinser J. Prolegomena to any future artificial moral agent. J Exp Theor Art Int. 2000;12(3):251–261.

Arkin RC. Governing lethal behavior: embedding ethics in a hybrid deliberative/reactive robot architecture. Atlanta (GA): Georgia Institute of Technology; 2007. Report No.: GIT-GVU-07-11.

Asaro PM. What should we want from a robot ethic. Int Rev Info Ethics. 2006;6(12):9–16.

Army Training Support Center (ATSC). Apply the ethical decision-making method as a commander leader or staff member; 2007. In: Air University Values and Ethics; c1999-2007 [accessed November 2017]. http://www.au.af.mil/au/awc/awcgate/army/ethical_d-m.htm.

Clark A, Chalmers D. The extended mind. Analysis. 1998;58(1):7–19.

Conitzer V, Sinnott-Armstrong W, Borg JS, Deng Y, Kramer M. Moral decision making frameworks for artificial intelligence. Proceedings of the 31st AAAI Conference on Artificial Intelligence (AAAI-17); 2017 Feb 4-9; San Francisco, CA. AAAI Press, 2017. p. 4831–4835.

DeBaets AM. Can a robot pursue the good? Exploring artificial moral agency. J Evol Tech. 2014;24(3):76–86.

Dennett DC. The intentional stance. Cambridge (MA): MIT Press: 1989.

Dennett DC. When HAL kills, who's to blame? In: Stork DG, editor. HAL's Legacy: 2001's Computer as Dream and Reality. Cambridge (MA): MIT Press; 1997.

Department of the Army. The soldier's guide. Washington (DC): Headquarters, Department of the Army; 2004 Feb. Field Manual No.: FM 7-21.13.

Hellström T. On the moral responsibility of military robots. Ethics Info Tech. 2013;15(2):99–107.

Hiller FR. How to (dis)solve Nagel's paradox about moral luck and responsibility. Manuscrito. 2016;39(1):5–32.

Himma KE. Artificial agency, consciousness, and the criteria for moral agency: what properties must an artificial agent have to be a moral agent? Ethics Info Tech. 2009;11(1):19–29.

McDermott D. What matters to a machine? In: Anderson M, Anderson SL, editors. Machine Ethics. Cambridge University Press; 2011. p. 88–114.

McLaren BM. Extensionally defining principles and cases in ethics: an AI model. Artificial Intelligence. 2003;150(1-2):145–181.

Nagel T. Moral luck. Mortal questions. Cambridge University Press; 1979. p. 24–38.

Office of General Counsel (OGC), Department of Defense. 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.

Parthemore J, Whitby B. What makes any agent a moral agent? Reflections on machine consciousness and moral agency. Int J Mach Conscious. 2013;5(2):105–129.

Smith AM. Attitudes, tracing, and control. J Appl Philos. 2015;32(2):115–132.

Sullins JP. When is a robot a moral agent? Int Rev Info Ethics. 2006;23–30.

Taylor M. Self-driving Mercedes-Benzes will prioritize occupant safety over pedestrians. Car and driver; 2016 Oct. https://www.caranddriver.com /selfdriving-mercedes-will-prioritize-occupant-safety-over-pedestrians/.

Uzgalis W. John Locke. In: Zalta EN, editor. The Stanford encyclopedia of philosophy. Stanford (CA): Metaphysics Research Lab, Stanford University; Winter 2017.

Wallach W. Robot minds and human ethics: the need for a comprehensive model of moral decision making. Ethics Info Tech. 2010;12(3):243–250.

Warren MA. Moral status: Obligations to persons and other living things. Oxford (England): Clarendon Press; 1997.

Weinberg S. Locke on personal identity. Philosophy Compass. 2011;6(6):398–407.

Winfield AF, Jirotka M. The case for an ethical black box. Proceedings of the Conference on Towards Autonomous Robotic Systems. Springer, Cham; 2017. p. 262–273.

Yampolskiy RV. Artificial intelligence safety engineering: why machine ethics is a wrong approach. Philos Theor Art Int. 2013;389–396.

List of Symbols, Abbreviations, and Acronyms

AI	artificial intelligence
COA	course of action
EBB	ethical black box
LOW	law of war
UGV	unmanned ground vehicle

1	DEFENSE TECHNICAL
(PDF)	INFORMATION CTR
	DTIC OCA
2	DIR ARL
(PDF)	IMAL HRA
	RECORDS MGMT
	RDRL DCL
	TECH LIB
1	GOVT PRINTG OFC
(PDF)	A MALHOTRA
1	DIR ARL
(PDF)	RDRL VTA
	R ST AMANT