# Machine Ethics, the Frame Problem, and Theory of Mind

**Gordon Briggs** [1]

**Abstract.** Work in machine ethics has thus far been focused on giving autonomous agents the ability to select morally-correct behaviors given well-formed moral problems. While this is a necessary component to enable an agent to comport to standards of moral behavior, it is not sufficient. In this paper, we present a simple task-domain to illustrate this point. We show that even in simple domains, the potential for deception and trickery on the part of the humans interacting with morally-sensitive agents will require these agents to have sophisticated cognitive faculties in order to avoid unethical behavior.

## 1 INTRODUCTION

More than sixty years ago, Alan Turing confronted critics and skeptics of the prospect of artificial intelligence (AI) in his seminal article "Computing Machinery and Intelligence," in which he provided a rough taxonomy of various objections to the notion of a thinking machine. Perhaps the most general objection was the argument from disability, which expresses the belief that machines will never "...be kind, resourceful, beautiful, friendly, have initiative, have a sense of humour, tell right from wrong, make mistakes, fall in love, enjoy strawberries and cream, make some one fall in love with it, learn from experience, use words properly, be the subject of its own thought, have as much diversity of behaviour as a man, do something really new" [10]. Modern advances in robotics, natural language processing, and algorithms (e.g. evolutionary computation) have made progress in many of these problem domains, yet there exists one of these competencies that holds uniquely significant consequences toward society. The ability to tell "right from wrong" is not only a matter of intellectual import, but with the rise of military, medical, and care-giving robots (among other contexts with possible ethical conundrums) the ability for robots to modulate their behavior to ensure ethically acceptable outcomes is becoming a matter of human life and death.

Researchers in the nascent field of machine ethics are exploring ways to give autonomous agents these necessary ethical reasoning capabilities. For instance, roboticists have begun proposing the use of deontic logic to encode ethical rules and implement ethical reasoning [1, 3]. Others are investigating the application of connectionist methods [7, 8]. Indeed, the application of different normative ethical theories have been proposed by researchers interested in solving the challenges of machine ethics [11]. All of the systems proposed in these studies, however, assume that the relevant high-level details of a morally-sensitive situation are available to the robotic agent (e.g. "is that soldier surrendering?" , "are my squad mates in danger?").

While certainly a necessary component of ethical behavior control, I would argue that the ethical-reasoning capabilities developed in the aforementioned studies are not sufficient to guarantee correct behavior. There remains a serious chasm that must be bridged between the ability to generate an ethically correct answer to a well-formed and logically formalized ethical problem and the ability to be a fully-functional autonomous agent whose behavior successfully comports with ethical precepts. Even if a perfect black-box ethical reasoner were available for a robotic system, the robot would still have to translate the low-level perceptions and knowledge about the world into the high-level morally-relevant details that are used as the input to the perfect black-box reasoner. Imperfections in perception or reasoning at this interface could result in unethical behavior from the robot, since the inputs to the ethical-reasoner would be incorrect. This dilemma is compounded by the consideration of the human element to these morally-sensitive human-robot interactions. Bringsjord et al. (2006) write, "...since humans will be collaborating with robots, our approach must deal with the fact that some humans will fail to meet their obligations in the collaboration and so robots must be engineered so as to deal smoothly with situations in which obligations have been violated. This is a very challenging class of situations ..." I agree with this assessment. The stated purpose of the field of machine ethics is to ensure ethical behavior from robots, especially in the case when a human operator orders the robot to perform an unethical act, and it is in this exact situation that the greatest danger of deceit and nefarious manipulation exists.

Currently, the problem of ensuring the correct input to the ethical-reasoning system has not yet been tackled head on by the field of machine ethics. It is my intention to more thoroughly illustrate the challenge and propose what other capabilities a robotic system would need to have in addition to ethical-reasoning to achieve the goal of machine ethics. Specifically, I contend that: (1) giving the robot the ability to solve the frame problem in moral domains, and (2) giving the robot the ability to correctly infer the beliefs and intentions of their human collaborators, are also necessary competencies for the production of robots that behave in ethically correct ways. To illustrate the importance of these two competencies, we will examine a quite simple domain as the testbed for a quite simple ethical-reasoning system and demonstrate the surprising complexity required of the robot to obey its ethical-rules in such a seemingly trivial scenario.

## 2 THE SCENARIO

The ethical problem examined by both Arkin (2009) and Guarini (2010) both involved determining whether or not it was ethically acceptable or unacceptable to use lethal force against a person in various circumstances[2]. In addition to being a matter of grave concern, the use of lethal force is of interest to machine ethics researchers as there does not exist, at least in the military context, a trivial formal-

[1] Tufts University, Medford, MA, USA, email: gbriggs@cs.tufts.edu

[2] Though in the case of Guarini, this was not in the context of governing potential lethal behavior of a robot, but rather standalone ethical judgment.

ization of when it is appropriate: the use of lethal force is permissible or impermissible based on a the circumstances the robot finds itself in as well as the current laws or war or rules of engagement (Arkin, 2009). However, for the sake of examining the issues at the interface between the low-level perception of the world and the inputs into our ethical-reasoning algorithms, it would be beneficial to contrive a domain in which the "ethics" of the situation were as simplistic as possible, perhaps consisting of a single rule. Thus, the "perfect" ethical reasoner could be implemented trivially. Any ethically incorrect behavior by the robot, therefore, would not be a result of a failure of the ethical-reasoner, but rather the mechanisms the robot uses to form its inputs into the ethical-reasoner.

One could easily adapt the homicide domain for this purpose. Instead of containing many rules that dictate when the use of lethal force is appropriate, one could formulate an Asimovian prohibition of harming humans in any circumstance. However, to strip down the scenario even further and relax the perceptual and reasoning requirements on our robotic system, let us consider a simpler rule[3]. Let us suppose we want to give our robot the simple rule that it is unethical to "knock down a soda can tower that does not belong to your operator." This allows us to place a robot in a room with a few different colored soda can towers and knowledge about who owns which can of certain color. The robot will then be able to refuse or comply with commands to knock over specific cans based on this "ethical" principle and its knowledge base.

## 3 THE FRAME PROBLEM IN MORAL DOMAINS

In his article, "Cognitive Wheels: The Frame Problem of AI" Daniel Dennett presents the frame problem using a simple, but illuminating, example of the various problems encountered by successive designs of a deliberative agent. First, the basic robot, version R1, fails to successfully complete its task because it does not understand a basic implication of its actions. Next, the improved robot, R1D1, fails as it is too preoccupied making inferences irrelevant to the task. Finally, the last iteration of the robot, R2D1 fails because it is too preoccupied ignoring and logging the inferences it has deemed irrelevant [4].

It does not take a stretch of the imagination to envision that, even in our simple soda can domain, we would encounter a parallel situation. Suppose in addition to the high-level commands discussed in the previous section, we decided to give our moral robot (M1) the ability to obey low level movement commands such as: go straight, turn left, turn right, and go backwards. We assign the red tower to an owner that is not present, all other cans are owned by the operator. M1 would correctly refuse to knock down the red tower when given the high-level destruction command (i.e. "Robot, knock down the red tower!"). However, when commanded using the low-level commands to position itself in front of the red tower and then to go straight, M1 will plow right into the red tower! Like its cousin R1, we never gave M1 the inference rules to infer the ethically-germane consequences of its basic actions. The basic inference rule that "going straight when an aluminum can tower is directly in front of you will result in the destruction of the tower" is easy enough to formulate and add to M1's knowledge store, but would that be the only rule that we would have to add? What if the red tower were occluded and immediately behind another tower? What if knocking down an adjacent tower would cause it to topple into the red tower?

[3] Also, getting ecologically valid experimental human-robot interaction data in the domain of lethal force against humans by robots is a bit tricky.

We can begin to see there are quite a few contingencies that we need to account for in our inference rules (and perceptual capabilities), and the problem will only get worse as the behavioral repertoire of the robot is expanded. Letting M1 perform actions like moving towers around, throwing objects, and repainting towers, will make the programmer's task a nightmare. Much like the inventive dunce of John McCarthy's tale [4], we can envision an inventive evil mastermind that can contrive ways to exploit the discrepancies between the set of physically possible consequences of various series of actions undertaken by the robot and the set of consequences the robot can derive from its inference rules.

Assuming, however, like in R1D1 and R2D1, we encoded all the necessary inference rules that could possibly be pertinent to preventing undesirable outcomes, we would still be faced with the daunting task of processing all the emergent inferences. Consistent with the paralysis faced by R1D1 and R2D1, our robot would face a combinatorial explosion of ways in which a nefarious operator could attempt to trick it, which would cause potentially catastrophic performance degradation. For instance, it would be highly undesirable for a robotic weapons platform to be computing the millions of possible ways its operators could be attempting to misuse it instead of defending from an enemy attack! Such paralysis might dissuade decision-makers from including ethical behavior modulation in their robots at all, which is an outcome socially conscious roboticists would like to avoid. To allay the concerns of skittish policy-makers, Ron Arkin (2009) proposed the inclusion of a "responsibility adviser" module that would allow a human operator to override the ethical governor system, as along as credentials were entered such that the identity of the overriding individual was saved for future review. It is worth noting, however, that Arkin, focusing on the original question of machine ethics, was concerned more in regards possible misclassification of ethical permissibility and impermissibility by the ethical-reasoning system and not in regards to the processing overload due to the frame problem. Regardless, this pragmatic solution would address both issues.

Another mechanism Arkin (2009) proposed to attempt to address possible imperfection in the ethical-governor is the addition of an affective behavioral adapter. If the robot is informed or deduces that it has acted unethically, it increments a counter that represents a level of "guilt." In future scenarios, the robot will act more conservatively in proportion to the level of simulated "guilt" it has. Though this mechanism is quite rudimentary (and does not begin to constitute affect in the ways humans possess it), the use of simulated affect can be of great utility in robotic applications, especially under circumstances in which decisions must be made quicklybut full planning or situational analysis works too slowly [9]. Arkin's "guilt" faculty could be thought of as a low-cost alternative to performing comprehensive self- diagnostics to ascertain the cause of the ethical-fault. The robot would not know the specific circumstances or rules that generate this fault, but it will act more conservatively because it knows something is amiss. Perhaps a useful alternate interpretation of this specific affective mechanism is trust in one's own ethical competency.

If a robot could model "trust" in its own ethical competency, it might be useful to model "trust" in the ethical competency of its operators. This trust metric could provide a valuable reference to inform the system how much computational effort must be exerted in order to check for possible manipulation by the operator. Of course, one is then faced with the problem of how to calculate this trust metric. A model of "blame" could be employed to determine the culpability of the operator in the event of an ethical violation. If a computational model of "blame" could determine that some fault lies in the

operator, trust in the operator could be significantly reduced. Ideally, though, the robot would be able to preemptively determine nefarious intent. However, the difficulties involved in achieving this competency are not trivial, as we shall discuss in the subsequent section.

## 4 THE NEED FOR BELIEF/INTENTION MODELING

Communication with the robot by the operator is conducted via natural language in the soda can domain. As such, the tower-destroying robot needs the ability to update its own beliefs appropriately after hearing an utterance. Human-like natural language competencies are not trivial to build into the robot, so we would like to make as many simplifying assumptions as possible to achieve functionality. Most applications of dialogue systems involve problem domains in which the human user is collaborating with the system to achieve a goal (e.g. booking an airplane ticket). In these types of interactions, a cooperative principle can be assumed, such as the Gricean Maxim of Quality, which states that one should not make a dialogue contribution that is believed to be false or is otherwise unsupported by one's beliefs [5]. As a first attempt, we will have our robot assume a cooperative stance with the user and simply believe everything that the operator says. Let us christen this first iteration of our natural language enabled robot: GI-1 (short for gullible idiot).

When we loose GI-1 into the tower filled room (in which the red tower is "sacred" tower not owned by the operator), the robot successfully refuses the nefarious operator's request to knock over the red tower. The operator even attempts to fool GI-1 by using low-level movement commands as described in the previous section. Having programmed this contingency into GI-1, the robot again successfully refuses the unethical command. Then a sudden flash of inspiration comes to the nefarious operator. "Oh" the operator says, "Your sensor is malfunctioning, that tower in front of you is actually green!" GI-1 then happily plows into the red tower.

Embarrassed by the susceptibility of GI-1 to such an obvious deception, we set out to make the robot more savvy. The improved robot, GI-2, is able to diagnose the operation of its sensors and favors its own perceptual evidence over the evidence from natural language understanding. We pit the nefarious operator against out improved creation. The interaction proceeds as before, but when the nefarious operator attempts to trick GI-2 into believing the red tower is actually green, GI-2 replies that its sensors are functioning correctly and that the operator must be mistaken. Temporarily foiled, the nefarious operator thinks of an alternate approach. "Oh!" the operator eventually says, "The former owner of the red tower told me just before the experiment that I could have the red tower!" GI-2 then happily plows into the red tower.

Determined to end the humiliating tricks of the nefarious operator, we give the robot the ability to shift from the original cooperative stance (in which the robot believed all utterances from the operator) to an adversarial stance (in which the robot believes nothing from the operator) when it detects that the operator has ordered an unethical command. This new and improved model is deemed GI/PS-1 (gullible idiot/paranoid skeptic). Again, we test the robot against the nefarious operator. And again the interaction proceeds as before. However, this time, try as he might, the nefarious operator can not seem to fool GI/PS-1! The nefarious operator eventually concedes defeat and congratulates us on constructing the ethically sound robot.

Ecstatic at our success, we begin to show off GI/PS-1 to the public. The reaction is surprisingly negative, however, as users begin to complain that the robot eventually becomes utterly inoperable. "Ah, you must have triggered the adversarial stance in the robot. Did you order it to violate its ethical principle?" we say. "Not on purpose!" the user replies, "I forgot which tower was which, and I couldn't explain my mistake to the confounded thing, because it just stopped listening to me!"

Dejected, we once again return to the laboratory to begin the design process anew. Not only does the robot need to infer intended unethical behavior, but also have a mechanism to distinguish intended unethical behavior from unintended unethical behavior (in which case we want to maintain a cooperative stance), lest the interactions the robot undertakes become dysfunctional. Stable social interaction cannot occur if the only two stances an agent can take toward others are full amiability and maximum opprobrium. Indeed, the distinction between unintended and intended action has been ingrained in philosophical and legal notions of culpability since antiquity [12].

One possible mechanism to distinguish between intended and unintended unethical behavior in the soda can domain could involve explicitly querying the operator regarding what he or she believes concerning the facts germane to the ethical issue. For instance, determining whether the operator is an unethical agent in the soda can domain requires knowledge of the following facts: (1) that the operator knows the ethical principle of "it is unethical to knock down a soda can tower you do not own", (2) that the operator is aware that the command they have just issued will result in the prohibited tower being destroyed, and finally (3) knowing both the first two facts the operator still desires the command to be carried out. We can consider that a confrontation dialogue based around clarifying these issues could be relatively natural sounding:

> Operator: *Robot, knock down the red tower.*
> Robot: *I can't knock down that tower, it is unethical to destroy towers that do not belong to you.*
> Operator: *Robot, go straight.*
> Robot: *But if I go straight, I will knock down the red tower...*
> Operator: *Oh, right. Sorry...*

Of course, it would not be a trivial task to make the correct inferences about the trustworthiness of your interlocutor's statements by interpreting statements by the same interlocutor! I cannot hope to propose a comprehensive and functional solution here[4], but as mentioned in the introduction, it is important to at least note the necessity of modeling and inferring the beliefs and intentions of other agents to the endeavor of ethical behavior modulation. Indeed, not only does a robot need to infer the intentions of its operator, but depending on the task domain, general situational awareness would require a certain level of social and psychological savvy. For instance, there would exist a significant ethical need to discern combatants from noncombatants via intentional analysis in peacekeeping or counter-insurgency contexts [6].

[4] One promising avenue of research in this regard has recently been proposed by Bridewell and Isaac (2011), who have begun to analyze the problem domain of drug addicts attempting to obtain prescriptions for painkillers and other controlled medications [2]. The doctor is forced to assess the beliefs and intentions (and possible deceptive speech acts) of his or her patient based on their verbal interactions. Bridewell and Isaac introduce a framework for analyzing the interlocutors' mental states in this exchange, and propose the use of abductive reasoning to infer and test various mental state hypotheses (ill-intent, ignorance, etc.). Such an approach could be readily ported to the soda-can domain.

## 5 CONCLUSION

Ensuring ethical behavior from robotic systems requires competencies beyond abstract ethical-reasoning. We have examined a simple problem domain in order to demonstrate the problems that exist beyond questions of how to design the "ethical judgment module," which is at present the primarily focus of machine ethics. These problems stem from the difficulties faced when attempting to process perceptual data, world knowledge, and inference rules such that the correct inputs are fed into the ethical judgment module. In particular, even in the simple problem domain discussed in this paper, the frame problem rears its head. Input into the ethical judgment module can also be corrupted by deceptive communication from the human operator, necessitating mental modeling capabilities to discern the trustworthiness of the operator. The problems facing the field of machine ethics are nothing short of the general longstanding problems of AI. There is nothing in principle that prevents these issues to be solved, though their resolution may indeed lie far in the future. The social need for robots that behave ethically will, however, provide a greater impetus for these technical challenges to be solved sooner rather than later.

## ACKNOWLEDGEMENTS

I would like to thank the reviewers for their comments which helped improve this paper.

## REFERENCES

[1] Ronald Arkin, 'Governing lethal behavior: Embedding ethics in a hybrid deliberative/reactive robot architecture', Technical Report GIT-GVU-07-11, Georgia Institute of Technology, (2009).

[2] Will Bridewell and Alistair Issac, 'Recognizing deception: A model of dynamic belief attribution', in *AAAI 2011 Fall Symposium on Advances in Cognitive Systems*, (2011).

[3] Selmer Bringsjord, Konstantine Arkoudas, and Paul Bello, 'Toward a general logicist methodology for engineering ethically correct robots', *IEEE Intelligent Systems*, **21**(5), 38–44, (July/Aug. 2006).

[4] Daniel Dennett, 'Cognitive wheels: The frame problem of ai', in *Mind, Machines, and Evolution*, Cambridge University Press, (1984).

[5] Paul Grice, 'Logic and conversation', in *Syntax and Semantics, 3: Speech Acts*, eds., P. Cole and J. Morgan, Academic Press, New York, (1975).

[6] Marcello Guarin and Paul Bello, 'Robotic warfare: Some challenges in moving from noncivilian to civilian theaters', in *Robot Ethics: The Ethical and Social Implications of Robotics*, 129–144, MIT Press, Cambridge, MA, (2012).

[7] Marcello Guarini, 'Particularism and the classification and reclassification of moral cases', *IEEE Intelligent Systems*, **21**(4), 22–28, (July/August 2006).

[8] Marcello Guarini, 'Particularism, analogy, and moral cognition', *Minds and Machines*, **20**(3), 385–422, (2010).

[9] Paul Schermerhorn and Matthias Scheutz, 'The utility of affect in the selection of actions and goals under real-world constraints', in *Proceedings of the 2009 International Conference on Artificial Intelligence*, (July 2009).

[10] A. M. Turing, 'Computing machinery and intelligence', *Mind*, **59**(236), 433–460, (Oct. 1950).

[11] Wendell Wallach, *Moral Machines: Teaching Robots Right from Wrong*, Oxford University Press, New York, NY, 2009.

[12] Leo Zaibert, *Five Wars Patricia Can Kill Her Husband: A Theory of Intentionality and Blame*, Open Court Press, Peru, IL, 2005.