“Think and Do the Right Thing” – A Plea for Morally Competent Autonomous Robots

Matthias Scheutz
Human-Robot Interaction Laboratory
Department of Computer Science
Tufts University
Medford, MA 02155
Email: matthias.scheutz@tufts.edu

Bertram F. Malle
Cognitive, Linguistic, and Psychological Sciences
Brown University
Providence, RI 02912
Email: bfmalle@brown.edu

Abstract—**Autonomous robots are increasingly employed in human societies without any provisions for “moral behavior” (other than some implicit architectural protective measures such as to avoid collisions or to follow human orders). However, being part of human societies, these robots will inevitably face morally charged situations, if not *moral dilemmas* where no simple solution exist, and thus need to be able to handle them appropriately. We argue for the need to incorporate explicit mechanisms for moral competence into robotic architectures and briefly sketch several components that are required for such moral competence.**

I. Introduction

Ordinary life frequently presents challenges to people that make it impossible for them to meet all of their obligations. Part of the problem is that they have only limited resources available to pursue their goals and that resource capacity can quickly become a constraining factor, especially when multiple goals have to be accomplished within a short time frame. The resource problem can sometimes be avoided through better planning, such as making fewer commitments and allowing for more time to fulfill those commitments. However, a more important part of why people at times fail to meet all of their obligations is that those obligations can be incompatible. In fact, it is a hallmark of human norms, values, and moral principles that they form a partially inconsistent network of obligations of different priorities and strengths, and there is little that individuals can change about it.

While there is no single best way to cope with these mutually inconsistent obligations, people typically approach resolutions by employing moral reasoning and, at times, meta-reflections based on principled statements and moral stipulations, so as to arrive at reasons why it is morally acceptable to keep some and drop other obligations. These reasons are not only used for decision making but also form the basis of justifications people give to others when trying to avoid or mitigate blame—a primary social mechanism for dealing with norm violations.

Now consider autonomous robots, which are increasingly deployed into human societies in various roles, ranging from socially assistive robots such as elder care robots or physical therapy robots, to various household robots such as vacuum cleaners and lawn movers, to robot toys and robot companions. All of these robots participate to varying degrees in human social lives and thus have the potential to become, passively and actively, involved in human moral struggles, both as moral agents and moral patients [1]. Consequently, social robots have to be prepared to deal with the same perturbing mix of social conventions and contradictory principles that humans deal with every day. In particular, it is not only possible, but rather seems very likely, as we will argue, that social robots will face “moral dilemmas” in which standard ways of decision making in robotic architectures are insufficient [2]. Robotic architectures therefore need to be expanded by various moral reasoning capacities that are necessary to arrive at solutions fully acceptable to humans.

In this paper, we will first motivate the need for moral competence in autonomous robots with an example from an elder care setting. Based on the moral structure of the example, we will argue that morally charged situations like the elder-care scenario are easy to come by, and that autonomous robots as a result need to be morally competent to be able to handle unexpected morally charged situations where none of the potential actions are indisputably right. We will then sketch a set of requirements for such moral competence, discuss the architectural extensions required to accommodate them, and briefly report on our first steps to allow for “moral action selection.”

II. The Case of Rob, the Elder-Care Robot

To see that some autonomous robots will likely face situations that present ethical dilemmas, consider an autonomous robot R in an elder-care setting, where R is assigned to a largely immobile human H in H’s home (see also [2]). R’s job is to support H in all daily-life tasks as much as possible (e.g., prepare food and feed H, ensure H is taking the required medicine, alert the remote health care supervisor if H health situation deteriorates, consult with the supervisor before any medication is administered, etc.). Overall, R has an obligation to provide the best possible care for H and keep H’s pain levels as low as possible. Now suppose that H had a very bad night and is in excruciating pain in the morning. R notices the pain expression on H’s face and asks if it could help H find a more comfortable position in bed (given R’s goal to minimize H’s pain). Instead, H asks R for pain medication. Since R has an obligation to get permission from the remote supervisor before giving H any medication, it attempts to

978-1-4799-4992-2/14/$31.00 © 2014 IEEE

Published by the IEEE Computer Society

reach the supervisor, even though it knows that providing pain medication in this context is an appropriate action without any medical side-effects. However, repeated attempts to contact the supervisor fail (e.g., because the wireless connection is down), hence R cannot obtain the permission. Hence, R is left with the following "moral dilemma": it can either give H the pain medication (without permission) and thus reduce H's pain, while violating the imperative to consult with the supervisor first before administering any medication (even though pain medication is harmless in this case); or it can refrain from providing pain medication, thus letting H suffer in vain and violating the obligation to keep H's pain levels as low as possible. What should R do?

This example is one of many possible scenarios in which pre-defined rules for governing the robot's behavior without any recourse to the moral aspects of the situation can run into trouble, even though they might work just fine for non-moral contexts [2]. An interesting question is what a human health care provider might do in R's position. A human provider P's decision to hand out pain medication would probably depend on several observed facts, such as how severe H pain is, but possibly also on the extent to which P has empathy for H, is willing to ignore strict orders, and is able to justify rule violations to the remote supervisor after the fact. Hence, if R were to model human behavior, it would, in addition to ethical reasoning, need the capability for empathy as well as the ability to generate justifications (i.e., explanations of norm violations such as not contacting the supervisor). We will not focus on those aspects of moral competency in this paper (but see [3]). Rather, we will develop a general argument that, in order to avoid unnecessary harm to humans, autonomous artificial systems must have moral competence.

III. Moral Dilemmas Pop Up Everywhere

The previous example is an instance of a *moral dilemma*, where an agent ought, all things considered, to do action A, and also ought, all things considered, to do action B, but cannot do both actions A and B. Formally, a moral dilemma consists of the following three formulas expressed in Deontic Logic: **O***A*, **O***B*, and $\neg\Diamond(A \wedge B)$ where **O** means "obligatory" and "$\Diamond$" means "physically possible" (as opposed to metaphysically possible, say). The dilemma of Rob, the elder-care robot, could be cast in Deontic Logic as follows:

1. $\neg havePermission(R, administer(R, H, M)) \rightarrow \mathbf{O}\neg administer(R, H, M)$ [obligation]
2. $inPain(H) \rightarrow \mathbf{O}administer(R, H, M)$ [obligation]
3. $\neg havePermission(R, administer(R, H, M))$ [fact]
4. $inPain(H)$ [observation]
5. $\mathbf{O}\neg administer(R, H, M)$ [1,3,MP]
6. $\mathbf{O}administer(R, H, M)$ [2,4,MP]
7. $\neg\Diamond(administer(R, H, M) \wedge \neg administer(R, H, M))$ [modal logic]

Thus, the robot has an *inconsistent obligation* that it cannot satisfy. Note that this dilemma is not a logical dilemma *per se* in that it does not lead to a *logical contradiction* (e.g., as in the Liar paradox, for example). However, if we adopt the "Principle of Deontic Consistency," $\neg\mathbf{O}\perp$, then we can derive a logical contradiction, as (5) and (6) imply $\mathbf{O}administer(R, H, M) \wedge \neg administer(R, H, M)$. Similarly, we could adopt a principle akin to the Kantian dictum that "ought implies can," $\mathbf{O}A \rightarrow \Diamond A$ [Kant], which would then result in a logical contradiction for moral dilemmas:

1. $\mathbf{O}A$ [assumption]
2. $\mathbf{O}B$ [assumption]
3. $\neg\Diamond(A \wedge B)$ [assumption]
4. $\mathbf{O}A \wedge \mathbf{O}B$ [prop.logic,1,2]
5. $\mathbf{O}(A \wedge B)$ [deont.logic,4]
6. $\Diamond(A \wedge B)$ [Kant,prop.logic,5]
7. $\perp$ [prop.logic,3,6]

Note that the Kantian principle in the case of Rob might actually avoid the dilemma rather than leading to inconsistency. For Rob could use it to "reason its way out" of the obligation to contact the supervisor: R has the obligation to obtain permission (i.e., to get and have the permission) to administer the pain medication to H, which, by [Kant], means that it is possible to obtain to get and have the permission; however, this turns out to be (physically impossible) because the supervisor cannot be contacted, thus leading to a contradiction. Hence, one of the two obligations has to be dropped, allowing R to drop the obligation to obtain permission. And R can then also drop the obligaton that without having the permission it is not allowed to administer the pain medication (number 1 above). Finally, R would make note of this reasoning and then use it in justifying its decisions to the supervisor.

Moral dilemmas where more than one feasible action is obligatory are also called "obligation dilemmas," compared to "prohibition dilemmas" in which all feasible actions are forbidden. Even though different types of moral dilemmas with different severity have been discussed in the literature, and there even though is still an ongoing debate as to whether such dilemmas are "genuine" dilemmas (i.e., whether there is such as thing as a "moral dilemma"), we will assume the practical position of using the term "moral dilemma" to denote a situation where moral requirements are in conflict (i.e., an apparent conflict between moral imperatives in which to obey one obligation or prohibition would result in transgressing another).

Notice that the problem of determining what to do in such moral dilemma-like situations is tricky even for robots that have explicit representations of obligations and permissions, for even reasoning with these representations does not result in any action the robot can perform. Suppose we are given a situation S, a robot R, a human H, and an action A that will necessarily inflict harm on somebody (including H). Moreover, suppose the robot has the obligations to not inflict harm on anybody and to follow orders from H, and that not following orders will necessarily inflict harm on H. Then the order to perform action A will cause a problem assuming a widely accepted deontic principle that $\Box(A \rightarrow B) \rightarrow \mathbf{O}A \rightarrow \mathbf{O}B$ (DP):

1. $\mathbf{O}\neg harm$ [obligation 1]
2. $\mathbf{O}A$ [obligation 2]

3. $\Box(A \rightarrow harm)$ [assumption]
4. $\mathbf{O}A \rightarrow \mathbf{O}harm$ [DP,3,MP]
5. $\mathbf{O}harm$ [2,4,MP]
6. $\mathbf{O}harm \wedge \mathbf{O}\neg harm$ [1,5,PL]
7. $\mathbf{O}(harm \wedge \neg harm)$ [6,DL]

The action obligated by the last formula – for the robot to do harm and refrain from doing harm – is not physically possible, thus we again have a moral dilemma.

It is fairly straightforward to see that any ordinary decision-making situation can potentially become a morally charged, dilemma-like situation with the same kinds of difficulties for utility-theoretic decision-making algorithms. Consider any general decision-making situation in which multiple agents are involved; then assign different costs to the decision-maker's available actions such that the action outcomes affect multiple other agents in different ways, causing harm to some while sparing others, and vice versa under different cost assignments to the involved actions (see [2] for details).

In fact, note that any situation S in which an action A inflicts harm on somebody and in which not doing A also inflicts harm on somebody (possibly the same person) will become a moral dilemma for a robot as soon as the robot is either ordered to do A or to refrain from doing A. (Without an explicit instruction to carry out A the situation may or may not be a moral dilemma for the robot depending on whether it, on its own, attempts to execute A or considers executing A.)

IV. Implications for Robotic Architectures

The fact that any ordinary decision-making situation is potentially a morally charged dilemma-like situation that requires moral decision-making capabilities for their resolution is probably the strongest case that can be made for the need to incorporate moral decision-making mechanisms into robotic architectures (cp. to Moore's "explicit moral agents" [4]). In fact, one could argue that ultimately any decision-making situation for an autonomous robot deployed in human societies will be morally charged in that moral values are involved in any human activity. And although not every morally charged situation will be necessarily dilemma-like, the very fact that moral dilemmas could and do occur requires that robots be equipped to handle such eventualities appropriately.

The question then arises what capabilities robots need to have in order to be able to deal with morally charged situations [3]. Currently it is not well understood (or agreed upon) what constitutes "human moral competence," let alone what it takes to replicate it in computational artifacts. For example, it is unclear exactly what moral computations and action representations are presupposed by human moral competence, and therefore also what cognitive mechanisms are required to implement such competence in artificial cognitive systems. Nor is it clear yet what potential effects artificial moral agents might have on humans.

It seems clear that we need to start by understanding the necessary basic concepts, including linguistic expressions, that are required to express, understand, reason over, and evaluate events relative to a system of moral norms – we will call this the *moral core*.

The moral core encompasses representations of essential moral concepts, together with their linguistic labels, as well as connections among these concepts. Among the necessary core concepts are at the very least (deontic) modal operators such as "permissible," "obligatory," and "forbidden," as well as representations of "norms" and "(moral) principles." Essential is also an understanding of what it means to adhere to those norms and principles and what it means to violate them, along with knowledge about typical sanctions imposed as a result of violations. In an integrated architecture, various aspects of these concepts need to be represented and integrated in different computational components, because different cognitive and subcognitive processes participate in the processing of moral information. Both implicit and explicit representations need to be employed if the agent is expected to not only "know" moral concepts but also "apply" them (e.g., in action execution and linguistic moral justification).

The second component of moral competence comprises the perceptual and cognitive processes involved in moral assessment, judgment, and reasoning. This includes all perceptual and cognitive processes involved in detecting situations in which some norm violation is being committed. But it also includes reasoning processes implicated in moral judgments (e.g., to determine who is to blame and what degree of blame should be attributed). And it also includes monitoring processes that check one's own and others norm-conforming behavior.

Closely associated, though distinct, are the affect representations and processes needed to capture typical moral emotions as they are generated in response to morally charged situations (e.g., empathy or resentment). Affective responses are particularly important for building models of human emotional responses to norm violations (e.g., [5]) that can help the robot predict human expectations and actions (e.g., for justification).

Yet another component concerns morally-aware planning, problem solving, and decision-making processes. These processes need to be able to represent deontic modal operators and use them as constraints on planning paths and decision trees (e.g., to determine which actions/states are permissible, etc. These features are necessary so that inferences or plans can be generated that are, at the very least, morally acceptable if not desirable. Moreover, such inferences and plans must contain all obligatory states – as intended states – and ideally attempt to minimize "non-intended" outcomes along the path. As part of this component, the system must have rich action representations that delineate "means," "ends," and "side effects" of human actions, thereby capturing crucial differences between performing actions intended to inflict harm and only knowingly inflicting harm as an unavoidable byproduct of a legitimate action. The representations have to be rich enough to include nested traces of possible decision and action outcomes, but compact enough to be viable as representations that are passed among architectural components (like task planners, moral reasoner, natural language processing, etc.).

Finally, a critical capability for robots interacting with humans is moral communication in natural language. Not only do such robots have to understand moral language for accepting instructions, but they also need to understand the different ways in which they could get blamed by humans. And, in response, they will have to be able to generate explanations and justifications of their actions and potentially norm-violating behaviors in a way that is not only accessible but intuitively acceptable to humans.

As a first step to allow for "moral action selection," we have started to modify the *Action Manager* component in our DIARC architecture [6], [7], which is responsible for carrying out *action scripts*, i.e., sequences of actions that achieve a particular goal state. The first extension was to allow for the inclusion of a set of *impermissible actions* (IA) that can be passed into each action script at the time of its instantiation (and hence can be modified based on the context in which the script is executed). The Action Manager then recursively checks for each action A in the script whether A is in IA. If $A \in IA$, the Action Manager first starts to search for possible alternative actions A' that are no impermissible. If successful, a permissible action is chosen. (Note that this process may include consulting the task planner to find complete alternative plans.) If no alternative action (or plan) can be found, the Action Manager consults the moral reasoner (if present) to determine whether an action override is warranted (e.g., because not performing the impermissible action is morally more problematic than performing it). If an override is recommended, then the Action Manager performs the action (without removing it from the set of impermissible actions), otherwise the script execution fails. Throughout the script execution the Action Manager explicitly tracks the execution states as they can be determined from the pre- and post-conditions of executable actions and stores the whole execution traces (including when and where alternatives were generated and overrides proposed).

The second extension is to allow for the specification of goal states in action scripts (instead of prescribing particular actions to achieve those states) together with a set of *impermissible states* (IS), which can be passed into each action script at the time of its instantiation. This allows for a more general specification of action scripts that more closely matches the way humans might give commands as well as impose action constraints. For example, it is often easier to specify an impermissible state than all the actions that could lead to it (e.g., "keep the patient hydrated" for the health-care robot instead of all the different ways in which this could be accomplished). Analogous to the execution of actions, the Action Manger will check whether a given state S specified in an action script is contained in the set of impermissible states. However, now it also has to check whether S implies any of the impermissible states in IS. Note that this is a much more difficult problem than checking whether S is simply contained in a set of states, for it not only requires a formal reasoning system, but it also requires bounds on the reasoning for the case that S does not imply any of the states in IS given the currently available information (for in that case, the reasoning process might not terminate). Currently, it is, however, not clear how to best bound the reasoning, although recent proposals are attempting to provide solutions (e.g., [8]).

V. Conclusions

In this paper, we have argued that autonomous robots deployed in human societies are likely to encounter many morally charged situations that humans struggle with in their ordinary lives. To be able to handle these morally charged situations in accordance with human norms and expectations, it is not sufficient for robots to simply follow a fixed set of rules. Rather, robots will have to have moral competence. This is particularly important for dilemma-like situations, which require potentially very sophisticated moral reasoning to determine a morally acceptable solution. We briefly sketched what "moral competence" might comprise in robots and then described our very first steps of extending the Action Manager of our DIARC architecture to effect the transition from implicit to explicit moral agents. However, this is only a first step in a long series of architectural extensions that include the representations and reasoning algorithms necessary to make autonomous robots morally competent.

VI. Acknowledgments

Funding for this work was in part provided by an ONR MURI grant #N00014-14-1-0144, "Moral Competence in Computational Architectures" to both authors. The authors would also like to thank their colleagues and collaborators on the MURI team for the various discussions that helped shape the framework and its associated challenges.

References

[1] G. Briggs and M. Scheutz, "Investigating the effects of robotic displays of protest and distress," in *Proceedings of the 2012 Conference on Social Robotics*, ser. LNCS. Springer, 2012.

[2] M. Scheutz, "The need for moral competency in autonomous agent architectures," in *Fundamental Issues of Artificial Intelligence*, V. Mueller, Ed. Berlin: Springer (Synthese Library), 2014, p. forthcoming.

[3] B. Malle and M. Scheutz, "Moral competence in social robots," in *IEEE International Symposium on Ethics in Engineering, Science, and Technology*, Chicago, May 2014, p. (forthcoming).

[4] J. H. Moor, "The nature, importance, and difficulty of machine ethics," *IEEE Intelligent Systems*, vol. 21, pp. 18–21, July/August 2006.

[5] B. F. Malle, S. Guglielmo, and A. E. Monroe, "A theory of blame," *Psychological Inquiry*, p. forthcoming, 2014.

[6] M. Scheutz, G. Briggs, R. Cantrell, E. Krause, T. Williams, and R. Veale, "Novel mechanisms for natural human-robot interactions in the DIARC architecture," in *Proceedings of the AAAI Workshop on Intelligent Robotic Systems*, 2013.

[7] M. Scheutz, P. Schermerhorn, J. Kramer, and D. Anderson, "First steps toward natural human-like HRI," *Autonomous Robots*, vol. 22, no. 4, pp. 411–423, May 2007.

[8] N. Alechina, M. Dastani, and B. Logan, "Norm approximation for imperfect monitors," in *Proceedings of AAMAS*, 2014, p. forthcoming.