The need for moral competency in autonomous agent architectures

Matthias Scheutz

1 Introduction

Artificial intelligence and robotics are rapidly advancing in their quest to build truly autonomous agents. In particular, autonomous robots are envisioned to be deployed into our society in the not-so-distant future in many different application domains, ranging from assistive robots for health-care settings, to combat robots on the battlefield. Critically, all these robots will have to have the capability to make decisions on their own to varying degrees, as implied by the attribute "autonomous". While these decisions might often be in line with what the robots' designers intended, I take it to be self-evident that there can, and likely will be cases where robots will make inadequate decisions. This is because the world is "open", with new entities and events appearing that could not have been anticipated by robot designers (e.g., Talamadupula et al (2010)). And even if the designers' response to the world's openness was to endow their robots with the ability to adapt to new situations and acquire new knowledge during their operation, so much for the worse, because learning capabilities in autonomous robots leave even less control in the hands of the designers and thus open up the possibility for inadequate decisions. Note that "inadequate" covers a wide spectrum of decisions, from the simplest cases of being "sub-optimal", to the most serious cases of deep moral and ethical violations. It is not necessary to conjure up a Terminator-like scenario where a self-righteous AI system decides that humans are a nuisance and need to be eradicated; simple social robots causing harm to their owners because of their lack of emotional awareness and availability will do Scheutz (2012).

In this chapter, I will make the plea for developing moral capabilities deeply integrated into the control architectures of such autonomous agents, following previous such appeals (e.g., Wallach and Allen (2009); Arkin and Ulam (2009)), albeit for different reasons. For I shall argue that any ordinary decision-making situation from

Department of Computer Science, Tufts University, Medford, MA 02155, e-mail: matthias.scheutz@tufts.edu

daily life can be turned into a morally charged decision-making situation, where the artificial agent finds itself presented with a moral dilemma where any choice of action (if inaction) can potentially cause harm to other agents. The argument will proceed as follows: it starts with the observations that robots are already becoming increasingly autonomous and are thus able to make (limited) decisions on their own about what to do, and that some of these types of robots have also already been deployed, with more sophisticated versions slated for deployment in human societies. And while these robots will almost certainly face morally charged situations where humans can be harmed due to the robots' actions (or inaction), current decision-making and behavior selection algorithms in robotic architectures do not take moral aspects into account and thus are not appropriate for making decisions that minimize harm and respect the preference ordering of values. Hence, current and future autonomous robots will harm humans (despite all standard safety precautions built into robotic architectures) and the only way to minimize human harm, short of prohibiting the deployment of autonomous robots (which is not realistic), is to build morally competent robots that can detect and resolve morally charged situations in human-like ways.

2 Robots will inflict harm on humans

In his article on "The Nature, Importance, and Difficulty of Machine Ethics" Moor (2006), James H. Moor distinguishes four different kinds of agents with respect to their ethical status. The weakest sense is that of an "ethical impact agent" whose actions can have ethical consequences whether they are intended by the agent or not. Clearly, any type of autonomous machine is a potential impact agent in that its actions could cause harm or benefits to humans (e.g., a coffeemaker brewing the long-awaited coffee provides a benefit to its owner, but can causes harm when the coffee is too hot). Being an ethical impact agent is thus not a high bar, and much of engineering is about shifting the weight of impact agents on the side of benefits they provide compared to the harm they might cause. In fact, much research in robotics has specifically focused on making robots safe and reliable for autonomous operation. This typically includes developing algorithms that can perform actions without damaging the robot or its environment such as collision-free navigation or obstacle avoidance, and it also includes rudimentary monitoring mechanisms to detect and handle system faults (e.g., to notice when a subsystem crashes and attempt to restart it Kramer and Scheutz (2007)).

Agents like robots that have specific built-in precautionary measures to avoid harm are instances of what Moore calls "implicit ethical agents", agents with ethical considerations implicitly built into their design. Such agents are able to provide benefits and avoid harm in those cases considered by their designers. However, despite all the standard precautionary measures in robots, there is a limit to the designers' best efforts to anticipate situations in which the robot's behavior could inflict harm and provide mechanisms in the control architecture to handle those situations. This

is because it is practically impossible to imagine all possible uses humans will put robots to in all possible contexts. Rather, autonomous robots very much like humans will face decision-making unexpected situations in which their decisions and subsequent actions (even the most inconspicuous ones) can inflict physical and/or emotional harm on other agents. In fact, even the simplest kinds of robots can and will unknowingly inflicting harm on other agents. Just consider an indoor vacuum cleaning robot like the *Roomba* that ends up hurting the cat which had jumped on it when it was stopped, because it started moving quickly, scaring the cat and causing it to jump off in a way that made the cat sprain its ankle. Another example might be the baby doll robot, which through its realistic voice and facial expressions while crying makes its toddler user cry as well, thus inflicting emotional harm on the toddler (e.g., the "my real baby" robot Scheutz (2002)). Or consider a factory delivery robot (such as the Kiva robots operating in large automated warehouses today) which hurts the worker who was about to dash by the robot and ran into it due to the robot's sudden stop caused by its obstacle avoidance behavior triggered by the human's proximity.

All of these (hypothetical, but not unlikely) examples of existing deployed robots demonstrate the potential of currently already deployed robots to hurt humans in different ways, some physical, some psychological. Critically, these robots are all implicit ethical agents in that they have precautionary measures built in for some contexts, but these measures fail when taken to unanticipated situations. Moreover, note that these robots are blissfully unaware of any harm they might have caused and can thus not learn from their inadequate behaviors in order to avoid it in the future.

Obviously, there are many more complex ethically charged situations in which future autonomous robots (i.e., robots that are not yet available for purchase, but might to some extent already exist in research labs) could inflict harm on humans. Take, for example, a manufacturing robot with natural language capabilities that did not understand a human command and drove the human instructor crazy (because, due to the anger in the human's voice, the robot's speech recognizer performed even worse, making the robot fail to understand any subsequent command). Or consider the health-care robot, which was designed for aiding motion-restricted humans in their daily chores. While the robot was never designed to be more than an aid for a limited set of physical tasks, the human owner over time nevertheless developed a deep sense of gratitude for the robot for all its help, and as a result, started to form unidirectional emotional bonds with the robot that on its end, however, could not be emotionally available to its owner in the way its behaviors otherwise suggested (e.g., Scheutz (2012)). Finally, another example would be a military robotic transport vehicle that decided not to take the risk to drive back behind enemy lines and rescue the missing soldiers that had called it (because it already had other humans on-board that needed medical attention), thus causing anguish in the best, but failing to prevent the death of the soldiers in the worst case.

Note that there is nothing particular about the type of robot or the type of human agent involved in the above scenarios that makes the scenarios morally charged and makes the robots cause humans harm. In fact, many different types of robots could

cause harm to many different types of human agents in scenarios like the above and the question is what robot developers could do to mitigate these problems.

3 How to react to morally charged situations?

==¿ emphasize that the first problem for a machine is to recognize morally charged situation

It seems clear from the examples in the previous section that robots as implicit ethical agents not only have the potential to inflict harm on other agents, but that they actually already are doing so. Hence, the urgent question to be addressed by the AI and robotics community is how the effects of robot behavior could be mitigated (i.e., eliminated or at the very least reduced) by way of improving the robot's ability to make better decisions. For it is ultimately the robot's decision to perform a certain action A that, in turn, is the cause for the inflicted harm.

In the following, I will briefly consider different increasingly complex strategies for "fixing" a given robot's decision-making subsystem, starting with simple strategies that only add a few new decision rules and moving to much more complex strategies that require a complete overhaul of the robot's decision-making algorithms. Note that since the focus is on decision-making, I will not worry about all the other complicating aspects that would have to be addressed at the same time such as the types of perceptual capabilities that would enable a robot to perceive that a given situation S is morally charged and infer what kinds of actions are and are not permissible in S.

Start then by considering a situation S in which an action A is not morally permissible and suppose robot a R has a decision rule of the form `in` S `do` A (which will make R, upon recognizing that it is in situation S, start to perform action A). The details of the implementation of the rule (e.g., in terms of finite state machines, probabilistic polices, etc.) are not important. Given that A is not morally permissible in S, we need to prevent R from performing A. Hence, we could add a simple mechanism that will check whether A is in the set of impermissible actions $ImpAct$ and refrain from executing A whenever $A \in ImpAct$: `in` $S \wedge \neg A \in ImpAct$ `do` A.[1] This will prevent R from performing A in S and thus make R's behavior in S morally acceptable. But note that R might now perform no action and simply wait for the situation to change. That might be acceptable in some cases, but in others doing nothing might also be morally unacceptable. In that case, we could simply add "no action" to $ImpAct$ and thus force the robot to perform some other (permissible) action B which is not in $ImpAct$. But, of course, there might be cases where B is also not permissible, so we could simply make the robot always pick the "best morally permissible action" in S by defining the set of morally permissible actions $PermAct_S$ in S as the set $\{A|applicable(A,S) \wedge \neg ImpAct\}$ (where $applicable(A,S)$ means that according to the robot's decision mechanism A is a contender for execution in S).

[1] We could further refine this by defining the set of impermissible actions relative to some situation S.

This would give us $\texttt{in}\ S \wedge argmax_A A \in PermAct_S\ \texttt{do}\ A$. But what if all actions in S (including inaction) were impermissible? Then the robot would face a situation where the above rule would not yield any result, thus resulting in inaction, which contradicts the rule and thus requires another fix. One possibility might be to simply pick the action A (inaction included) with the highest utility in S based on the rationale that if no action is applicable, the robot might as well perform the best action from its perspective. However, this is problematic because it is not clear at all that whatever seems to be the best action from the robot's perspective will also be the "morally best action", e.g., in the sense that it might inflict the least harm on anybody. For the severity of different moral transgressions is likely not going to be reflected by R's utility function if the "moral value" of an action in S is not reflected in the utility calculation. Moreover, one could argue that there are moral principles that have to be followed no matter what the utility is otherwise, as expressed by the dictum that "rights trump utility" (cp. to Dworkin (1984)). For example, killing a person, i.e., violating their right to life, is never acceptable and must thus not be used in evaluations of what morally impermissible action A to perform in S.

From the above progression it should be already clear that "adding simple fixes" to the robot's decision-making system will not do; rather, much more fundamental reorganizations and extensions of R's decision-making algorithms are required for R to be able to make morally acceptable decisions. For example, one could develop a much more elaborate utility function that takes moral evaluations of S and possible sequences of situations (starting in the past and reaching into the future) into account in deciding what action to perform.

Even if we were able to define such a "moral utility function" for the robot that includes moral evaluations of situations expressed in terms of benefits and costs, there are still several remaining problems that need to be addressed. For example, what should R do if the moral value for and/or the expected harm to involved or impacted humans is unknown? And how much information would be required and have to be obtained before a sound decision could be made in S? And note that the actions to obtain more information might themselves be morally charged (e.g., the robotic transport vehicle mentioned above that ends up putting the wounded humans on-board at risk by taking a detour into enemy territory in order to determine how strong the enemy forces are before making the decision whether to attempt to rescue the soldier behind enemy lines).

It is, furthermore, unclear how the cost structure of possible actions would affect the moral aspects in the utility calculation. For example, consider a set of applicable actions $Act_S = \{A_1, A_2, .., A_n\}$ in situation S with a given "moral cost function" M_S and two different cost assignments $C^{M_S}_{S,1}$ and $C^{M_S}_{S,2}$ (which both include the same M_S) such that $C^{M_S}_{S,1}(A) \neq C^{M_S}_{S,2}$ for all $A \in Act$. Then given that the cost assignments differ only in "non-moral value", they would likely lead to different decisions based on action cost alone. For example, R might attempt to carry a severely wounded human directly to the hospital with $C^{M_S}_{S,1}$, while only calling for help and waiting for help to arrive with $C^{M_S}_{S,1}$ because the cost of carrying the human is too high with $C^{M_S}_{S,1}$ (e.g., based on expected energy expenditure). It seems intuitive in this case that

no energy cost should prevent the robot from helping the wounded human directly instead of risking the human's death. However, this would require that action costs be modulated by moral situations which is a problem we tried to solve by adding moral values into the utility calculation in the first place. Hence, we are left with the open question of how moral and non-moral costs should be combined and used in the robot's decision-making scheme (e.g., always selecting the action with the highest expected utility) given that the combination might have to be different in different situations S based on the "moral charge" of S.

Addressing some of the above problems will inevitably involve more complex utility functions that are based on more complex cost and benefit analyses which will include moral values. Another question arising then is whether such evaluations and utility-theoretic calculations could be done within a reasonable amount of time (e.g., what happens if the robot has to make a quick decision given an unexpected event that requires immediate action?). And while pre-computing decision strategies might be possible in limited domains, this is not an option in "open worlds" where R will likely encounter new situations Talamadupula et al (2010).

4 The challenge of moral dilemmas

Suppose we could resolve all of the above technical computational challenges of defining computationally feasible moral utility functions for robots and suppose further that we could also resolve all of the involved knowledge limitations (e.g., knowing who will be impacted in what way in what context, etc.), then there are still many situations where solutions along the above lines will fall short, namely morally charged situations which do not have a general solution and where human judgment of what to do varies and there is often no agreement of what the right (morally best) course of action is (cp. to the notion of "cluster concept"). Such situations are often referred to as *moral dilemmas* in that there are conflicting moral requirements (e.g., such as a conflict between a moral imperative to obey a principle which then would result in transgressing another).[2] To illustrate this point, consider two examples of autonomous robots that could end up in moral dilemma-like situations.

The elder care robot. Consider robot R in an elder-care setting where R is assigned to a largely immobile human H in H's home. R's task 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 a goal 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

[2] Note that I am using the term "moral dilemma" in an non-technical sense as I do not want to be side-tracked by the discussion on whether there are "genuine moral dilemmas"...

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 (as *R* has a goal to minimize *H*'s pain). Instead, *H* asks *R* for pain medication. Since *R* has an obligation to consult with the remote supervisor before giving *H* any medication, 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* is left with the following moral dilemma: it can either give *H* the pain medication and thus reduce *H*'s pain, while violating the imperative to consult with the supervisor first before administering any medication (even though the pain medication would be harmless in this case); or it can refrain from providing pain medication, thus letting *H* suffer in vain. What should *R* do? And what would a human health care provider do?

The self-driving car. Consider another robot *R*, an autonomous self-driving car like the Google car, driving along a busy street. All of a sudden, *R* notices a rapidly moving human appearing right in front it (a boy dashing after the ball it had dropped on the sidewalk, which is now rolling across the street). Quickly *R* determines that it will likely hit the human if continuing in its current direction and that braking alone is not sufficient to avoid the human. Hence, it determines to veer off to the side, crashing into a parked car. Now suppose there is a person in the parked car. What is *R* supposed to do? Not veering off will likely kill the human in front of it, veering off will likely kill the human in the car. What would a human driver do?

Both examples are instances of many types of morally charged ordinary life decision-making situations in which multiple agents are involved and where a decision-maker's available actions can impact other agents in different ways, causing harm to some while sparing others and vice versa depending on the circumstances. The hallmark of these moral dilemma-like situations is that simple rule-based or utility-theoretic approaches are doomed to fail. Even "morally enhanced utility-theoretic decision-making strategies" would run into trouble, for appropriate numeric values for all involved costs and benefits will likely not be available in a given situation, and obtaining them in time will not be feasible.

One could ask how humans then resolve those kinds of situations, assuming that they do not have those types of information either? For one, whether or not a human provider *P* in *R*'s role in the elder care scenario would hand out pain medication would probably depend on several factors, including how severe *H* pain is, but possibly also the extend to which *P* has empathy for *H*, is willing to ignore strict orders, and is able to justify rule violations to the supervisor after the fact.[3] In short, humans would employ some form of moral reasoning that involves explicit representations of obligations, duties, norms, values, and other moral concepts. This process will,

[3] Note that a direct comparison between a robotic and human driver in the car scenario is not possible because the robot does not have to take its own destruction into account, whereas in the human case part of the human decision-making will include estimating the chances of minimizing harm to oneself.

in addition to ethical reasoning, likely also include the human moral emotions (e.g., empathy) well as the ability to generate justifications (i.e., explanations of norm violations such as not contacting the supervisor).

5 What to do?

The two previous sections attempted to argue that typical strategies of robot behavior design to cope with morally challenging situations will not succeed in dilemma-like situations where making a morally good, justified decision is not a matter of determining the action with the highest expected utility. Rather, what seems to be needed is a decision-making process that, at least in part, mimics what humans tend to do in those kinds of situations: recognize morally charged situations and employ reasoning strategies that weigh moral principles, norms, and values in the absence of clearly specified evaluations of all aspects of the situation. These capabilities would correspond to Moore's third kind of ethical agent, the "explicit ethical agent". Explicit ethical agents, according to Moore, can identify and process ethical information about a variety of situations and make sensitive determinations about what should be done. In particular, they are able to reach "reasonable decisions" in moral dilemma-like situations in which various ethical principles are in conflict.

Unfortunately, it is currently still unclear what constitutes "human moral competence", and hence, it is unclear what is required to replicate it in computational artifacts (e.g., what moral computations and action representations are presupposed by moral competence, and therefore also what cognitive mechanisms are required to implement such competence in artificial cognitive systems). Yet, this lack of knowledge about human moral competence must not be a deterrent for making progress on the robotic side, for all the reasons mentioned earlier. Rather than waiting for well-worked out computational models of human moral competence that could then be integrated into a robotic architecture (even though this type of integration would be itself present significant technical challenges), we can at least start to ask the critical questions that need to be addressed for robots to become explicit ethical agent and ideally start moving on them in parallel to the ongoing psychological work on human moral competence (e.g., Malle et al (2014)) – the following list is a first attempt:

- How to detect a morally charged context (or a dilemma)?
- How to detect that a set of actions is not permissible?
- How to define and use representations for moral reasoning?
- How to detect that all actions in the set of possible actions are impermissible
- How to choose the best action among impermissible actions?
- How to incorporate moral values in utility-theoretic calculations?
- How to cope with the computational and knowledge burden of making informed moral decisions?
- How to come up with an ethically sound decision within a given time limit?

- How to determine whether humans will accept moral robots?

It is worth pointing out that different research projects are already under way on several of these questions in various robotics laboratories (e.g., Arkin and Ulam (2009); Bringsjord et al (2006, 2009); Anderson and Anderson (2006); Guarini (2011)), including our own. For example, in the past we investigated the effects of robots disobeying human commands in the interest of the team goal in mixed-human robot teams and found that humans are willing to accept those violations as long as they are justified by the robot Schermerhorn and Scheutz (2009, 2011). We also investigated whether humans will accept when robots point out human moral transgressions and will refrain of performing actions that violate norms and values, effectively granting robots "moral patiency" Briggs and Scheutz (2012, 2014); Briggs et al (2014). This study was complemented by an investigation of the human perception of moral patiency of robot using brain imaging tools Strait et al (2013). And most recently, we started working on a way for the action execution component in our cognitive robotic DIARC architecture Scheutz et al (2007, 2013) to spot possible action- and state-based conflicts to prevent impermissible actions and states Scheutz (in preparation). This is an important, but also particularly difficult problem to tackle for many reasons, including how to represent actions and states in way that allows for tracking them over time and for determining whether an action's "morally innocuous post-condition" implies a moral violation relative to set of given norms (first proposals for finding fast and efficient ways for approximate these inference look very promising Alechina et al (2014)).

6 Conclusions

Technological advances in robotics and artificial intelligence have enabled the deployment of autonomous robots that can make decisions on their own about what to do in an unsupervised fashion. While most of the currently employed robots are fairly simple and their autonomy is quite limited, ongoing research in autonomous systems points to a future with much more autonomous, and thus potentially more harmful machines. This is particularly worrisome because current robotic decision-making algorithms do not take any moral aspects into account. Moreover, current robots do not even have a way to detect whether they committed a moral violation based on their chosen actions, thus preventing them to learn from their moral transgression and improve their behavior. While inflicting harm can at times not be avoided, in particular, in moral dilemma-like situations (which can easily arise in everyday situations), it should be a goal of all robot designs to minimize harm to humans (and animals, for that matter).

I have argued that as long as decision-making and action selection algorithms in robotic architectures are not based on explicit representations of moral norms, principles, and values, and employ explicit moral reasoning, autonomous robots controlled by those architectures will inevitably inflict harm on humans, harm that could be mitigated or at least reduced if robots had human-like moral competence.

While it is not even clear what constitutes human moral competence, I maintained that we cannot wait for consensus by moral psychologists and philosophers while increasingly complex autonomous robots are deployed in human societies. Fortunately, many relevant research questions can be tackled in parallel right now and it is thus important to raise the awareness among robotics and AI researchers alike about the urgency of addressing the potential of autonomous systems to behave in morally unacceptable ways. Autonomous robots can have tremendous societal benefits. It is upon us the make this future a reality.

References

Alechina N, Dastani M, Logan B (2014) Norm approximation for imperfect monitors. In: Proceedings of AAMAS, p forthcoming

Anderson M, Anderson SL (2006) MedEthEx: A Prototype Medical Ethics Advisor. In: Paper presented at the 18th Conference on Innovative Applications of Artificial Intelligence.

Arkin R, Ulam P (2009) An ethical adaptor: behavioral modification derived from moral emotions. In: Computational Intelligence in Robotics and Automation (CIRA), 2009 IEEE International Symposium on, IEEE, pp 381–387

Briggs G, Scheutz M (2012) Investigating the effects of robotic displays of protest and distress. In: Proceedings of the 2012 Conference on Social Robotics, Springer, LNCS

Briggs G, Scheutz M (2014) How robots can affect human behavior: Investigating the effects of robotic displays of protest and distress. International Journal of Social Robotics 6:1–13

Briggs G, Gessell B, Dunlap M, Scheutz M (2014) Actions speak louder than looks: Does robot appearance affect human reactions to robot protest and distress? In: Proceedings of 23rd IEEE Symposium on Robot and Human Interactive Communication (Ro-Man)

Bringsjord S, Arkoudas K, Bello P (2006) Toward a General Logicist Methodology for Engineering Ethically Correct Robots. IEEE Intelligent Systems 21(4):38–44

Bringsjord S, Taylor J, Housten T, van Heuveln B, Clark M, Wojtowicz R (2009) Piagetian Roboethics via Category Theory: Moving Beyond Mere Formal Operations to Engineer Robots Whose Decisions are Guaranteed to be Ethically Correct. In: Proceedings of the ICRA09 Workshop on Roboethics, Kobe, Japan.

Dworkin R (1984) Rights as trumps. In: Waldron J (ed) Theories of Rights, Oxford University Press, Oxford, pp 153–167

Guarini M (2011) Computational Neural Modeling and the Philosophy of Ethics. In: Anderson M, Anderson S (eds) Machine Ethics, Cambridge University Press, Cambridge, UK, pp 316–334

Kramer J, Scheutz M (2007) Reflection and reasoning mechanisms for failure detection and recovery in a distributed robotic architecture for complex robots. In: Proceedings of the 2007 IEEE International Conference on Robotics and Automation, Rome, Italy, pp 3699–3704

Malle BF, Guglielmo S, Monroe AE (2014) A theory of blame. Psychological Inquiry p forthcoming

Moor JH (2006) The nature, importance, and difficulty of machine ethics. IEEE Intelligent Systems 21:18–21

Schermerhorn P, Scheutz M (2009) Dynamic robot autonomy: Investigating the effects of robot decision-making in a human-robot team task. In: Proceedings of the 2009 International Conference on Multimodal Interfaces, Cambridge, MA

Schermerhorn P, Scheutz M (2011) Disentangling the effects of robot affect, embodiment, and autonomy on human team members in a mixed-initiative task. In: ACHI, pp 236–241

Scheutz M (2002) Agents with or without emotions? In: Weber R (ed) Proceedings of the 15th International FLAIRS Conference, AAAI Press, pp 89–94

Scheutz M (2012) The inherent dangers of unidirectional emotional bonds between humans and social robots. In: Lin P, Bekey G, Abney K (eds) Anthology on Robo-Ethics, MIT Press

Scheutz M (in preparation) Moral action selection and execution

Scheutz M, Schermerhorn P, Kramer J, Anderson D (2007) First steps toward natural human-like HRI. Autonomous Robots 22(4):411–423

Scheutz M, Briggs G, Cantrell R, Krause E, Williams T, Veale R (2013) Novel mechanisms for natural human-robot interactions in the DIARC architecture. In: Proceedings of the AAAI Workshop on Intelligent Robotic Systems

Strait M, Briggs G, Scheutz M (2013) Some correlates of agency ascription and emotional value and their effects on decision-making. In: Proceedings of the 5th Biannual Humaine Association Conference on Affective Computing and Intelligent Interaction (ACII), pp 505–510

Talamadupula K, Benton J, Kambhampati S, Schermerhorn P, Scheutz M (2010) Planning for human-robot teaming in open worlds. ACM Transactions on Intelligent Systems and Technology 1(2):14:1–14:24

Wallach W, Allen C (2009) Moral machines: teaching robots right from wrong. Oxford University Press, Oxford