Towards Morally Sensitive Action Selection for Autonomous Social Robots

Matthias Scheutz[1], Bertram Malle[2], and Gordon Briggs[1]

[1] Human-Robot Interaction Laboratory
Tufts University, Medford, MA 02155, USA
{matthias.scheutz,gordon.briggs}@tufts.edu
[2] Cognitive, Linguistic, and Psychological Science
Brown University, Providence, RI 02912, USA
bertram_malle@brown.edu

***Abstract*— Autonomous social robots embedded in human societies have to be sensitive to human social interactions and thus to moral norms and principles guiding these interactions. Actions that violate norms can lead to the violator being blamed. Robots thus need to be able to anticipate possible norm violations and attempt to prevent them while they execute actions. If norm violations cannot be prevented (e.g., in a moral dilemma situation in which every action leads to a norm violation), then the robot needs to be able to justify the action to address any potential blame. In this paper, we present a first attempt at an action execution system for social robots that can (a) detect (some) norm violations, (b) consult an ethical reasoner for guidance on what to do in moral dilemma situations, and (c) it can keep track of execution traces and any resulting states that might have violated norms in order to produce justifications.**

I. Introduction

Autonomous social robots are increasingly embedded in human societies. Different from other kinds of autonomous robots that might interact with humans for the purpose of specifying tasks or giving status updates, social autonomous robots do so *at a social level* (e.g., [1]). Specifically, they are connecting to people via multiple information channels (eye gaze, facial expressions, gestures, bodily postures, speech, etc.) and aim to engage in the kind of dynamic back-and-forth that humans engage in among each other (e.g., [2]). For social robots this kind of interaction with humans is an essential aspect of the robots' goals (e.g., for socially assistive robots such as wheelchairs or therapy robots [3]) or even the sole goal (e.g., for eldercare or companion robots like Paro [4]).

There are many challenges involved in ensuring that social robots are useful and effective interaction partners, from exhibiting joint attention [5], to respecting human timing in turn taking [6], [7], to being sensitive to human affect [8]. Critically, as autonomous social robots are becoming more complex in their behaviors and capabilities, they will ultimately have to respect the various levels of social norms and principles that govern human social interactions, from customary etiquette to legally binding ethics. For failing to respect human social and moral conventions and rules will ultimately result in interaction failures, where humans in the best case will become frustrated and interrupt the interaction but in the worst case will *blame* the robot, sliding into conflict or even avoid future contact. It is thus essential that autonomous social robots not only be aware of the ethical principles guiding human social interactions in the robots' deployment environment, but also that they continuously attempt to abide by those principles, even when cases arise where these principles are at odds with one another and no action that the robot can perform will ultimately be "blame-free".

In this paper, we present a first attempt at sketching an action selection and execution system for autonomous social robots that can detect (some) norm violations and consult an ethical reasoner for guidance on what to do in morally charged dilemma-like situations. We first start with three brief examples highlighting the urgency of development of such "morally aware" action selection mechanisms. Then we provide a more detailed architectural discussion of where in the architecture such mechanisms are needed and how they could interact with other components typically found in robotic control architectures. We conclude with a summary of our proposal together with suggestions for the next steps towards developing morally competent robots [9], [10], [11], [12].

II. Motivation

We have argued elsewhere [9] that even simple robots can cause physical or emotional harm to humans and animals. For example, vacuum cleaning robots or mobile robot toys might scare pets or children by inadvertently approaching or touching them; they might get into people's way and cause them to trip over them; they might accidentally knock over items causing spills and fires (e.g., by knocking over lit candles), and so forth. To illustrate the potential of autonomous social robots to inflict harm on humans, we will first describe two examples of a very simple existing robot, followed by an example of a much more complex robot envisioned for a more distant future. The goal of these examples is to reveal possible intervention points in the robot's action selection system where the integration of

This work was in part funded by 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 many helpful discussions.

specific capacities for moral competence [10] might be able to reduce or prevent harm.

Example 1: Imagine a simple vacuum cleaning robot that has no sensors other than bumper sensors around its circular base to determine when it hits an obstacle and a dust sensor to determine whether the area it drives over is still dirty. The robot has a fixed pre-programmed action pattern of driving in ever increasing spirals until it hits an obstacle, at which point it will attempt to follow the obstacles contour. The robot thus has no representation of what it is doing, what environment it is in, or even what type of objects it circumvents. Moreover, the robot does not have any memory of its actions, hence it does not know what control commands it had used in the past that got it into its present state. Nor can the robot plan ahead its future actions. Rather, the robot operates entirely in the *hic-et-nunc*, as its behavior is completely determined by its current sensor status and its *reactive* control system. In short, this robot is about as simple as it gets for a robot that can still do something useful in the world.[1] Now suppose that the robot is cleaning an upstairs room in a single-family home where a little boy is playing in the closet. The robot, by a sheer coincidence, bumps into the open closet door. And even though it immediately reverses direction upon contact of its bumper sensors with the door (as instructed by its reactive control to avoid obstacles), the small impact is sufficient for the door to shut and lock the boy in. The little boy is shocked and starts screaming immediately, while the robot continues it cleaning mission, oblivious to what has just happened. The boy's parents cannot hear him as the boy starts hyperventilating because the closet door is closed and eventually stops breathing. When the parents eventually go upstairs to check on their son, they find him suffocated in the closest. Meanwhile the cleaning robot has moved on to their bedroom where it eventually ran out of power. Once their shock is starting to subside, they are looking for explanations and are at a loss as to how their son managed to lock himself in.

This dramatic example is intended to illustrate that even the simplest of all robots are capable of exerting harm to humans, directly or indirectly, because of their impact on the environment but their lack of awareness of this impact (e.g., see [9]). While accidents like the one described in the above example are hopefully very unlikely, it is easy to imagine other cases where robots are used as tools to perform harmful acts to other humans.

Example 2: Imagine that the robot from Example 1 could also be given simple instructions through an app on a mobile phone. Specifically, the robot can be stopped and started again ("move forward"), and the moving direction of the robot can be specified ("turn left" and "turn right"). The interface is simply consisting of buttons and dials that graphically represent the possible actions and allow users to easily specify them. Also imagine the robot has a camera mounted on the front, which can stream the video feed to the app via the robot's built-in 4G network adapter. Hence, owners of the robot can connect to it from any location in the world with cell phone data connections, observe the robot's actions and give it new commands (e.g., steering it to a different location in the environment). Now consider the above Example 1 again, except that this time the robot is given the command to push the door closed as an act of revenge by the boy's brother downstairs who got a hold of the parents' cell phone with the installed app. The tragedy unfolds as before, but this time it was not the robot's fault that the door got shut; rather, it was the brother's purposefully timed command that made the robot drive into the door and shut it as a result.

While both examples had to be fairly contrived given that the robot's behavioral repertoire is so limited, the opportunities for robots to inflict harm on humans, physical and emotional, will overall increase with increasing behavioral sophistication, in particular, with social robots.

Example 3: Consider a future eldercare robot (not unlike the robot in the recent movie "Robot and Frank") that takes care of chores around the house while also keeping the elderly person company. The robot has natural language interaction capabilities to be able to take verbal instructions. And while the robot can also hold short limited conversations about the tasks around the house, it cannot go much beyond such conversations (e.g., engaging in small talk about the most recent football game). Now imagine the robot's lonely elderly owner who experiences growing gratitude[2] for all the robot is doing around the house and who then wants to connect to the robot on an emotional level. This is by no means far-fetched, as long-term interactions with socially assistive robots are likely to cause people to develop unidirectional emotional bonds with the machine [14]. Unfortunately, the robot in this case is not able to do so because it does not have the capability to understand and reciprocate human emotions. Consequently, the perceived cold-heartedness and unemotional responses of the robot may be interpreted as emotional distance, even rejection, thus causing emotional harm to the human owner. At the same time, the elderly person does not want to give up the robot as it is the person's only social connection, hence the emotional pain is recurring and ongoing, with no improvement in sight.

Example 4: Imagine a robot similar to the one in Example 3. In addition to the ability to understand and carry out verbal directives in order to assist with a variety of household tasks, it also has the ability to issue directives. This capability provides the robot with a richer ability to

[1]The reader might have guessed that such robots are commercially available.

[2]People even experience gratitude for the simple vacuum cleaner in the first example after several weeks of watching it do its job, so much so that they opt to clean for it to make its job easier [13].

engage in collaborative activities, where the robot is also able to ask for help when needed [15], as well as enable the ability to give reminders, which may be an important feature for certain user populations (e.g. patients with Alzheimer's disease). However, like the capabilities from the previous examples, this increased sophistication comes with a variety of risks and potentially deleterious consequences. Much like the emotionally insensitive robot from Example 3, a robot with the ability to make requests could cause similar distress by formulating these requests in ways that do not conform to human norms of politeness, perhaps by being either too direct or being perceived as overly critical [16]. Not only is the manner in which directives are formulated an important consideration, the content and the situational and social context of the requests influence whether or not particular directives should be issued by the robot at all. For instance, a request for the human to help plug the robot into the wall to recharge is perhaps an appropriately modest request, as opposed to a request for the human to drive immediately to the store to by a replacement battery (or pay a large fee for express shipping of a new battery). Inappropriate requests that are insensitive to social norms based on consideration of roles, imposition, and respect to the interlocutor's self-worth and image, may cause distress and/or annoyance in the robot's human interaction partners, which is undesirable per se, but also may lead ultimately to the rejection of the robot and all the other positive benefits it may confer.

However, there is also a danger that any norm-based modulation of directives and other communicative action may be too conservative. Consider the case where carbon monoxide begins to fill the elderly person's apartment. Unfortunately, the dedicated carbon monoxide detector does not trigger an alarm, as its maintenance has been neglected and its battery is out of charge. The assistive robot, fortunately, also has a carbon monoxide detector, which is operational. Despite this, the robot does not sound the alarm because it knows that its owner is sleeping and it is inappropriate to interrupt this activity. The elderly owner then passes away in his or her sleep due to carbon monoxide poisoning.

All of the above examples are just a few instances of a very large set of possible problematic cases where autonomous social robots can cause harm to humans because they are unaware of their environment, the consequences of their actions or inactions, and the mental states of their interlocutors. In Example 1, the robot did not notice that it accidentally shut the door and that a boy was in the closet because it had *no way to notice*. If it had been able to notice the situation, it could have called for help (e.g., by sounding a loud alarm). Similarly, the robot in Example 2 could have refused to drive into the door had it been able to anticipate what was about to happen. It could have justified the refusal to its human operator through the app interface or through a natural language explanation ("There is human in the room. Moving forward would shut the door and lock the person in, which is not allowed. Hence, I cannot move forward.")[3] The robot in Example 3 could attempt to do its best at reminding its owner that it is just a machine, that the absence of affect in its voice is meaningless, and that its lack of emotional reciprocation is due to its cognitive and affective limitations, not due to lack of respect.[4]. Finally, the robot in Example 4 could avoid request forms that are considered impolite if it were able to reason about sociolinguistic norms, as well as known when it is appropriate to ignore these norms if the situation presents a more morally undesirable norm-violation ("It is ok to be direct and/or rude if there is a risk of physical harm"). Hence, it is clear that to minimize or altogether avoid such harmful encounters with humans, robots will need special mechanisms in their control system that allow them to detect morally charged situations and prevent the execution of actions that will cause harm. We will, in the next section, briefly describe the different types of mechanisms needed for robots to reduce harm.

III. Requirements for Moral Action Selection

To get a better understanding of what is required for moral action selection, consider two different cases of the same robot performing the same norm-violating action in the same situation, albeit with a difference in the robot's knowledge about the moral status of the action or action outcome. In the first case, the robot *knows* that executing the action will (likely) cause a norm violation in the given situation, while in the second case, the robot *does not know* that. Even though the robot's actions in both cases have the same outcome, we would not blame the ignorant robot but would be inclined to blame the robot that knows about the potential harm (unless there are mitigating circumstances such as that any other action would have caused even more harm). Hence, regardless of how the robot comes to have the relevant knowledge (through perceptions, explicit instruction, or inference), once it has the knowledge in some clearly specified sense (e.g., having instantiated a data structure that represents the action outcomes together with their moral evaluation), it ought to take this knowledge into account when selecting an action.

A. Permitted and forbidden actions and states

The requirement for taking knowledge about the moral status of action outcomes into account thus prescribes certain design features that robot control architectures need to exhibit in their action execution component (AEC) to be able to perform moral action selection. Here we describe three core aspects:

- *Represent permitted and forbidden actions.* Specifically, by explicitly marking actions together with sets of

[3]The question of whether or not people will heed the refusals of robotic agents is an area of current research. There is some evidence that they do [17].

[4]It is an open question whether such frequent reminders will succeed, but for now we are hopeful that there might be effective ways to counter the strong and automatic human tendency to anthropomorphize autonomous robots; see also [14]

their arguments as *permitted* or *forbidden*, possibly augmented by an explicit context argument (which further specifies the contexts in which an action is or is not allowed), the AEC can check the permission status of every action before executing it. And when an action is not permitted, the AEC can notify other components in the architecture that will then be able to react by choosing any combination of the following options: (1) replanning using only permitted actions; (2) consulting an "ethical reasoner" that might determine justified exceptions (see below); (3) notifying the robot's supervisor. Note that some actions may be forbidden, not because of their effects, but because they directly violate social norms (e.g., "don't move" in a situation where the robot is supposed to sit still).

- *Represent permitted and forbidden states.* In addition to representing permitted and forbidden *actions*, the AEC also needs to represent permitted or forbidden *states*. For it is possible that a permitted action nevertheless results in a norm-violating outcome (e.g., consider the robot's "turn the cup upside down" action, which might result in the "spilled coffee" state if the cup was filled with coffee, a state that may not be permitted). While one could, in principle, annotate every action with all possible outcome states, this is practically infeasible because of the potentially large number of context-specific "side effects"—i.e., states that are uniquely brought about by the action in the local context in which it is executed. Representing only more general sets of permitted and forbidden states is more realistic but requires the AEC, in each particular context, to determine what local states may result from executing the planned action. This, in turn, might require additional planning and reasoning capabilities, as well as possible simulation capabilities in order to "envision" how action execution would impact the world up to some point in the future. Note that making permitted and forbidden states explicit is another way of representing *action contexts*—i.e., specifying when an action is permitted or forbidden, because an action is not permitted when it leads to a forbidden state. Also note that in many cases, it will be easier to specify forbidden states (e.g., to not hurt a human) than to specify all the possible actions and argument combinations that could lead into those states.

- *Specify exceptions.* Since the *network of moral norms* [18] is not perfect and sometimes norms overspecify restrictions (e.g., "do not push others") or, worse yet, are mutually inconsistent, it is necessary to represent exceptions for forbidden actions and states (e.g., "pushing people is allowed if it prevents them from getting hurt"). Such exceptions may be included directly into the norm representation by way of a default condition (e.g., a "normal situation" antecedent in a conditional), under which the norm applies, as well as exception conditions, when alternative norms or rules are applicable. In the case of inconsistent norms, a preference ordering might be necessary to determine which norm has more weight and if no such norm can be determined, additional conditions might have to be used to determine the best action. For example, if a set of actions are in direct conflict because each action violates a different norm and there is no precedence relation among the norms, the AEC could compare the action outcomes and pick the action with the least number of norm-violating states, paying attention to the precedence relation among the violated norms (e.g., preferring violations of lower-ranked norms). In addition, the AEC might need to consult an ethical reasoner to determine how to properly handle those norms [19]. If such a reasoner is available, the AEC would hand the current state, the principles in conflict, and the possible action paths to the reasoner and expect back a sequence of statements together with a recommended action that can be used as a logical justification for executing the recommended action or refraining from executing it. The statement sequence can also be used to generate justifications in natural language (as in the case of Example 2 with the robot controlled by the app).

- *Modulation of norms throughout the architecture.* Not only is it important to explicitly represent and reason about exceptions to norms in the AEC, but it is important for the AEC to be able to influence or override norm-based behavior in other components in the robotic architecture. While other components may not have explicit deontic-style representations of norms, they nonetheless have implicit representations that facilitate specific aspects of the robot's interactive capabilities. For instance, in a natural language enabled robot, the dialogue component will implement a turn-taking algorithm that informs the architecture when it is appropriate for the robot to speak. However, it is not difficult to imagine a scenario in which turn-taking conventions ought to be violated for the sake of giving an urgent warning (e.g "Look out!").

In addition to the above core requirements for morally sensitive action selection, it is important that the robot have at least some rudimentary perceptions of objects in its environment. For example, if the robot in the first and second example could have perceived the door, then it could have used a very simple exception principle to the unrestricted "move forward" action: "move forward unless door in front". In both cases, this restriction would have prevented it from pushing the door shut, albeit without any understanding for "why" this was not allowed. In fact, with this principle alone, the robot would never shut any door, which may or may not be desired.

B. Norm-based reasoning during action execution

Additional perceptual capabilities such as being able to detect the human in the closet and also detecting that there

is not other door to the closet, would enable additional inferences and increase safety. Let us assume that the robot has an explicit norm stating that "It is not permitted to close doors to rooms that contain at least one human":

$$\forall r, d.room(r) \wedge door(d) \wedge has(r,d) \wedge \\ \exists h.human(h) \wedge in(r,h). \rightarrow \neg \mathbf{P} do(close, d).$$

where **P** means "permitted". Moreover, assume that the robot's perceptual system generates the percepts:

$$room(r3),\ door(d1),\ has(r3, d1),\ human(h7),\ in(r3, h7)$$

where $r3$, $d1$, and $h7$ are new constants denoting the perceived objects in the environment. Then robot can infer in two steps (by substituting the constants from the perceptual system in the norm principle and applying modus ponens based on the perceptual facts) that it is not permitted to close the door:

$$\neg \mathbf{P} do(close, d1)$$

Note that the normative principle is again very broad and that additional refinements are possible, e.g., by making explicit that the closing action is only forbidden when the human inside has no way of getting out when the door is closed, which would make it a forbidden action if the robot could determine that the person inside the closet is a child that cannot reach the lock, but would make it permitted for an adult who could open the door from inside. Such refinements, however, would also require additional perceptual capabilities as well as *common sense knowledge* about the capabilities of people, about how locks in closets work, where they are usually located, whether they are reachable for a human of a particular height, and so forth.

C. Obligations and goals

Up to now we have considered only permitted and forbidden states, but we have not explicitly mentioned *obligations*, i.e., actions and states that the robot *must do* and *make or keep true*, respectively. Generally, we distinguish between obligations to make a particular state true and obligations to maintain a state (e.g., "close the door" versus "keep the floor clean"). The former can be realized as an "accomplishment goal" while the latter can be realized as an "achievement goal" in the robot control architecture. Assuming that a robot will always attempt to achieve all its goals, we can then, conversely, take any of the robot's goals to be obligations the robot has that can be explicitly represented by the "obligatory" operator **O**. E.g., in Example 1 the robot has the obligation to keep moving and vacuuming, i.e., not to stop ($\mathbf{O}\neg stop$). Since obligations and permissions are closely related in standard deontic logic in that if an action A is obligatory, then it is not permitted to not do A, we can then derive the forbidden actions and states logically implied by the robot's obligations. In the case of the robot in Example 1 the implied forbidden actions is stopping, i.e., it is not permitted to stop: $\neg \mathbf{P} stop$.

Note that in Example 2 an interesting norm conflict might arise between the user input ("go forward") and the inferred prohibition to go forward because going forward would lock the user in. Specifically, there is a conflict between $\mathbf{O} move(forward)$ which implies $\neg \mathbf{P} \neg move(forward)$ and $\neg \mathbf{P} move(forward)$ which is obtained from the reasoning about locking the person it. This is an instance of a classical moral dilemma where an agent is not permitted to do an action and at the same time not permitted to not do the action [12]. As mentioned before, these kinds of instances can either be resolved by assigning precedence relations among obligations (e.g., the robot's obligation to avoid harm trumps the robot's obligations to execute user commands) or to pass the dilemma to an ethical reasoner for resolution.

D. Keeping track of states and actions

Regardless of how the robot ends up resolving such a dilemma, it is important to note that the robot will likely get blamed in either case, exactly because there is a clash of obligations (and, likely, moral principles). Hence, it is particularly important in cases of norm conflicts for the robot to be able to justify its decision making, and a prerequisite for this justification is that the robot be able to access its past decisions. It is straightforward for the AEC to keep track of each action it executes and each state before and after the executed action. Depending on the robot's perceptual capabilities and additional common sense knowledge, the AEC can also store any generated proofs for or against executing an action at any point in time to allow for later inspection of what led to a certain robot action. This record of the robot's inferences, decisions, and actions will then allow the robot's supervisor to determine what the robot knew at any point in time, what the robot did, and why it did what it did. Such a record will also allow the robot to generate increasingly rich justifications that can elucidate why it did or did not perform a given action.

E. Modulation of lower-level norms

As we discussed in Example 4, it is possible that "lower-level" norms such as those pertaining to politeness or turn-taking may interfere with morally desirable natural language interventions. This raises the issue of how the components responsible for implementing these behaviors can be properly modulated by the "higher-level" reasoning of the AEC. This could be accomplished by encoding exceptions that the lower-level components are designed to query for. For instance, the turn-taking algorithm could be querying the ethical reasoner for whether or not $\mathbf{P} interrupt(self, L)$ is supported (where $self$ corresponds to the robot and L corresponds to the addressee). If the interrupt is permitted, turn-taking can be adjusted to allow for immediate speech generation, whereas if the interrupt is not permitted, then turn-taking can operate in the usual fashion.

IV. Discussion and Conclusion

In this paper, we argued that social robots embedded in human societies have to be sensitive to moral norms and principles guiding human behavior. Even simple robots' actions

can violate norms and therefore lead to the violator being blamed [20], [21], [22] and disrupt human-robot relations. Indeed, we expect that, with expanding interaction capabilities and levels of autonomy, social robots will get increasingly blamed for violating human norms and principles. We thus suggested that social robots, even relatively limited ones, need to be able to detect possible norm violations and attempt to prevent them while they execute actions—not simply to avoid blame but, more importantly, to avoid harming people. We presented four examples of robots that could harm humans in different ways and used them to motivate modifications to a robot's action selection system that would allow the robot to detect norm violations and prevent the execution of forbidden actions, or actions that would result in bringing about forbidden states. Specifically, we argued for explicitly representing permitted and forbidden actions, states, and exceptions, in addition to normative rules that can be used to perform inference together with perceptions and common sense principles in order to detect potential norm violations. While some reasoning capability will have to be built into the action execution component in the robot's architecture, additional ethical reasoning capabilities might be necessary to deal with moral dilemma-like situations where the robot would otherwise be at an impasse. Future work will thus have to focus on ways to integrate such higher-level ethical reasoning capabilities in ways that are practically applicable (i.e., the relevant normative information is available and can be formally captured) and feasible (i.e., the processes can unfold within reasonable time and resource limits).

V. 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] C. Breazeal, *Designing Sociable Robots*. MIT Press, 2002.

[2] C. Yu, P. Schermerhorn, and M. Scheutz, "Adaptive eye gaze patterns in interactions with human and artificial agents," *ACM Transactions on Interactive Intelligent Systems*, vol. 1, no. 2, p. 13, 2012.

[3] P. Briggs, M. Scheutz, and L. Tickle-Degnen, "Are robots ready for administering health status surveys: First results from an hri study with subjects with parkinsons disease," in *Proceedings of 10th ACM/IEEE International Conference on Human-Robot Interaction*, 2015.

[4] K. Wada and T. Shibata, "Robot therapy in a care house - change of relationship among the residents and seal robot during a 2-month long study," in *Proceedings of the 16th IEEE International Symposium on Robot and Human interactive Communication*, 2007, pp. 107–112.

[5] C. Yu, M. Scheutz, and P. Schermerhorn, "Investigating multimodal real-time patterns of joint attention in an hri word learning task," in *Proceedings of the 2010 Human-Robot Interaction Conference*, March 2010.

[6] H. H. Clark, *Using language*. Cambridge University Press, 1996.

[7] J. L. Rotondo and S. Boker, "Behavioral synchronization in human conversational interaction," in *Mirror Neurons and the Evolution of Brain and Language*, M. Stamenov and V. Gallese, Eds. Amsterdam: John Benjamins, 2003, pp. 151–162.

[8] M. Scheutz, P. Schermerhorn, J. Kramer, and C. Middendorff, "The utility of affect expression in natural language interactions in joint human-robot tasks," in *Proceedings of the 1st ACM International Conference on Human-Robot Interaction*, 2006, pp. 226–233.

[9] M. Scheutz, "The need for moral competency in autonomous agent architectures," in *Fundamental Issues of Artificial Intelligence*, V. C. Müller, Ed. Berlin: Springer, 2014.

[10] B. F. Malle and M. Scheutz, "Moral competence in social robots," in *Proceedings of IEEE International Symposium on Ethics in Engineering, Science, and Technology (ETHICS)*, 2014.

[11] B. F. Malle, "Moral competence in robots?" in *Sociable robots and the future of social relations: Proceedings of Robo-Philosophy 2014*, J. Seibt, R. Hakli, and M. Nrskov, Eds. Amsterdam, Netherlands: IOS Press, 2014, pp. 189–198.

[12] M. Scheutz and B. F. Malle, ""think and do the right thing" - a plea for morally competent autonomous robots," in *Proceedings of IEEE International Symposium on Ethics in Engineering, Science, and Technology (ETHICS)*, 2014.

[13] J.-Y. Sung, L. Guo, R. E. Grinter, and H. I. Christensen, "'my roomba is rambo': Intimate home appliances," *UbiComp 2007: Ubiquitous Computing, Lecture Notes in Computer Science*, vol. 4717, pp. 145–162, 2007.

[14] M. Scheutz, "The inherent dangers of unidirectional emotional bonds between humans and social robots," in *Anthology on Robo-Ethics*, P. Lin, G. Bekey, and K. Abney, Eds. MIT Press, 2012.

[15] S. Rosenthal and M. Veloso, "Using symbiotic relationships with humans to help robots overcome limitations," in *Workshop for Collaborative Human/AI Control for Interactive Experiences*, 2010.

[16] G. Briggs and M. Scheutz, "Modeling blame to avoid positive face threats in natural language generation," *INLG and SIGDIAL 2014*, p. 1, 2014.

[17] ——, "How robots can affect human behavior: Investigating the effects of robotic displays of protest and distress," *International Journal of Social Robotics*, vol. 6, no. 3, pp. 343–355, 2014.

[18] B. F. Malle and M. Scheutz, "When will people regard robots as morally competent social partners?" under review.

[19] S. Bringsjord, J. Licato, N. S. Govindarajulu, R. Ghosh, and A. Sen, "Real robots that pass human tests of self-consciousness," under review.

[20] P. H. Kahn Jr, T. Kanda, H. Ishiguro, B. T. Gill, J. H. Ruckert, S. Shen, H. E. Gary, A. L. Reichert, N. G. Freier, and R. L. Severson, "Do people hold a humanoid robot morally accountable for the harm it causes?" in *Proceedings of the seventh annual ACM/IEEE international conference on Human-Robot Interaction*. ACM, 2012, pp. 33–40.

[21] B. F. Malle, M. Scheutz, T. Arnold, J. Voiklis, and C. Cusimano, "Sacrifice one for the good of many? People apply different moral norms to human and robot agents," in *HRI '15: Proceedings of the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction, Portland, OR, USA*. New York, NY: ACM, 2015, pp. 117–124.

[22] A. E. Monroe, K. D. Dillon, and B. F. Malle, "Bringing free will down to Earth: Peoples psychological concept of free will and its role in moral judgment," *Consciousness and Cognition*, vol. 27, pp. 100–108, July 2014. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1053810014000671