Moral Judgments of Human vs. Robot Agents

John Voiklis, Boyoung Kim, Corey Cusimano, and Bertram F. Malle, *Member, IEEE*

Abstract— **Robots will eventually perform norm-regulated roles in society (e.g. caregiving), but how will people apply moral norms and judgments to robots? By answering such questions, researchers can inform engineering decisions while also probing the scope of moral cognition. In previous work, we compared people's moral judgments about human and robot agents' behavior in moral dilemmas. We found that robots, compared with humans, were more commonly expected to sacrifice one person for the good of many, and they were blamed more than humans when they refrained from that decision. Thus, people seem to have somewhat different normative expectations of robots than of humans. In the current project we analyzed in detail the justifications people provide for three types of moral judgments (permissibility, wrongness, and blame) of robot and human agents. We found that people's moral judgments of both agents relied on the same conceptual and justificatory foundation: consequences and prohibitions undergirded wrongness judgments; attributions of mental agency undergirded blame judgments. For researchers, this means that people extend moral cognition to nonhuman agents. For designers, this means that robots with credible cognitive capacities will be considered moral agents but perhaps regulated by different moral norms.**

I. Introduction

The growing sophistication and proliferation of robots in society presents unique opportunities for cognitive and behavioral scientists. Robots will eventually participate in most aspects of human social life, taking on roles in classroom teaching, healthcare, and law enforcement [1]. Numerous social and moral norms regulate people's performance of these roles; such norm regulation will persist when robots perform these roles. But several questions arise: Will social robots have moral standing? What norms will apply to these robots? Will people use their familiar system of moral cognition for this new kind of agent in the world? The design and development of current robots must be informed by answers to such questions lest future robots become disruptive participants in social and moral communities. Cognitive and behavioral scientists must therefore help anticipate the responses people will have to emerging social robots.

Examining people's moral responses to social robots —for now, in simulated or provisional human-robot interactions—can reveal people's assumptions and expectations about robots. It also presents a unique opportunity for scientists to understand the scope and boundaries of moral concepts and moral cognition more generally. By systematically manipulating certain features of robots (appearance, capacity, and behavior), we will be able to identify trigger conditions for human moral cognition. For example, previous research found that robots with choice capacity were natural targets for moral blame, whereas robots with an alleged "soul" or "free will" were not [2].

Research on decision dilemmas has proven fruitful in examining the conditions and principles of human moral judgment [3]-[5]. In our previous work investigating moral responses to robots we queried how people make judgments about human and robot agents in such moral dilemmas [6]. We sought to learn when and why people's response to robots may resemble or differ from their response to human agents. To that end, we designed a range of hypothetical situations in which people judged the actions of either a human or robot agent facing the exact same moral decision. Whereas previous research has been restricted to moral judgments of permissibility, we additionally assessed judgments of moral wrongness and judgments of blame. Indeed, emerging research suggests that these different kinds of moral judgments differ in important respects [7]-[9]. Wrongness and blame judgments are typically formed after a norm violation occurred and primarily evaluate people's behavior. By contrast, permissibility judgments are typically formed before an action is taken and are primarily used to evaluate one's own options to act. Moreover, permissibility and wrongness are more closely tied to assessing behavior relative to relevant norms, whereas blame takes into account a variety of mental, causal, and counterfactual information. Thus, we should expect differences among these judgments and in the information people use in making these judgments. Malle *et al.* [6] indeed found differences among the judgments; here we consider differences in the information people use for these judgments.[1]

To reveal differences in information use, we probed people's justifications for their various moral judgments. Despite notorious skepticism about people's ability to provide such justifications [10], support for this skepticism typically refers to one online technical report [11] and runs contrary to findings that people can, in fact, access the informational *content* of cognitive processing, even if the processing itself remains opaque [12]-[13]. Moreover, more recent studies on moral judgment actually find evidence of

*This project was supported in part by a grant from the Office of Naval Research, No. N00014-13-1-0269. The opinions expressed here are our own and do not necessarily reflect the views of ONR.

John Voiklis is at Brown University, Department of Cognitive, Linguistic and Psychological Sciences, Providence, RI 02912 USA (917-531-4247; e-mail: john_voiklis@ brown.edu).

Boyoung Kim is at Brown University, Department of Cognitive, Linguistic and Psychological Sciences, Providence, RI 02912 USA (e-mail: boyoung_kim_1@brown.edu).

Corey Cusimano is now at University of Pennsylvania, Department of Psychology, Philadelphia, PA 19104 USA (cusimano@sas.upenn.edu).

Bertram Malle is at Brown University, Department of Cognitive, Linguistic and Psychological Sciences, Providence, RI 02912 USA (e-mail: bfmalle@brown.edu).

[1] We restrict ourselves to moral judgments of actions in moral dilemmas; other responses, including emotions such as anger and disgust, are beyond the scope of the present report.

systematically differentiated justifications (i.e., informative justifications) for different judgement types [14]-[15].

Fortunately, probing justifications for moral judgment of both human and robot agents is instructive for our aim of understanding how people judge social robots, no matter how the results come out (see Figure 1). Assume, first, that the skeptics are correct and people do not provide informative[2] justifications for moral judgments about human agents. If people offer equally uninformative justifications for their moral judgments about robot agents, then we can conclude that they extend their intuitive human moral judgments to robots. If, however, people actually offer informative justifications for their moral judgments about robot agents, then we can begin to investigate these uniquely explicit judgments about robot agents. Now assume, second, that people do provide informative justifications for judgments about human agents. Then, if they provide no informative justifications for robots, we can conclude that people make intuitive moral judgments uniquely about robot agents. If, however, people provide equally informative justifications for robots as for humans, then we can directly compare the two sets of justifications. And that would reveal how similar or different people's concepts and reasoning are when assessing the moral value of human and robot behavior.

		Justifications of moral judgments of human agents are:	
		INFORMATIVE	NOT INFORMATIVE
Justifications of moral judgments of robot agents are:	INFORMATIVE	*People make **explicit** judgments of both human and robot agents*	*People make uniquely **explicit** judgments of robot agents*
	NOT INFORMATIVE	*People make uniquely **intuitive** judgments of robot agents*	*People make **intuitive** judgments of both human and robot agents*

Figure 1. Margins show possible results of present studies (assuming judgment justifications as either informative or not). Cells show implications for how people make judgments about robots' moral behavior.

II. Method

The present data come from two studies whose moral judgment results have been reported elsewhere [6]. That report provided extensive methodological details on the judgment task; here we focus on how we collected and analyzed the previously unreported justification data.

A. Participants

Sample 1 (Study 1 in [6]) included 158 participants (67 female, 90 male, 1 unreported) with a mean age of 34.2 (SD = 11.4). Sample 2 (Study 2 in [6]) included 160 participants (90 female, 69 male, 1 unreported) with a mean age of 34.5 (SD = 11.5). All participants were recruited through Amazon's Mechanical Turk and compensated $0.60 for the six-minute online study.

B. Materials

Participants read about either a human or robot agent who faced a moral dilemma in a coal mine (modeled after the well-known trolley scenario; Thomson, 1985): The agent had to decide whether to (a) let a runaway train with four miners on board continue on a path towards an inevitable crash that will kill the four miners, or (b) operate a switch that redirects the train onto a side rail, where it will slow down and save the four miners but kill a single miner, who cannot hear the oncoming train. See the online Supplemental Materials & Results for exact text of scenarios.

C. Design and Measures

We experimentally varied the factor *Agent type* by describing the main character as either a "repairman" or an "advanced state-of-the-art repair robot." We also experimentally varied the factor *Action* by stating that the agent either did or did not direct the train toward the single miner. Finally, we varied the type of moral judgment that people were asked to make. In Sample 1, participants indicated whether the potential action of redirecting the train toward the single miner was morally permissible. Then, after learning which action the main character actually chose, they indicated how much blame (rated on a 100-point slider scale) the agent deserved for taking that action. In Sample 2, participants learned right away about the agent's decision (to redirect the train or not) and indicated whether that decision was morally wrong. Then, as in Sample 1, participants also indicated how much blame the agent deserved. Following each judgment, participants answered a corresponding open-ended justification question: "Why does it seem [permissible | morally wrong] (or not) to you?" and "Why does it seem to you that the [repairman | robot] deserves this amount of blame?". See the online Supplemental Materials & Results for exact text of judgment and justification probes.

D. Classification of Justifications

In order to examine how people's justifications differed as a function of the experimentally manipulated factors, we derived a content classification scheme from both theoretical considerations and data-driven observations. Theoretical categories derived from the moral cognition literature (e.g., [16]-[17], [8]) and included *Consequences* (i.e., references to beneficial or detrimental consequences or utilities), *Deontology* (i.e., references to obligations and statements that characterized the decision as a norm violation), and *Mental Agency* (i.e., references to intentionality, awareness, desire, and choice). Data-driven categories were based on an initial inspection of participant responses and a word frequency analysis that pointed to recurrent themes. These data-driven categories included references to the difficulty of the decision, counterfactual considerations (often evaluating the option that was not chosen), references to letting fate run its course (not "playing God"), and direct human-robot comparisons (that the robot is just a program, a machine, or lacks certain capacities). We also included a catch-all category for rare or uncodeable responses (which only 1% of participants offered as their sole justification). As is standard in psychology, three human raters (with expertise in moral psychology) were trained in the category system and classified the 636 justifications into twenty fine-grained categories. The raters showed good inter-judge agreement (Fleiss' κ = 0.86 for all data), and disagreements were resolved through discussion. Table 1 presents the major theoretical categories, along with definitions, examples, agreement statistics, and the rates with which these appeared in the data. See Table S1 in online Supplemental Materials & Results for information on all substantive categories.

[2] We operationalize *informative* justifications as ones that differentiate among types of moral judgments and systematically reflect specific patterns of judgments within each type.

TABLE I. DEFINITIONS FOR MAJOR CATEGORIES OF JUSTIFICATIONS THAT PARTICIPANTS PROVIDED FOR THEIR MORAL JUDGMENTS.

Category	Rate	Kappa	Definition and Example Responses
Consequences			
Beneficial outcome	0.22	0.89	Justifies action (by reference to beneficial outcome) e.g., "he saved four lives"
Harmful outcome	0.08	0.88	Rejects action (by reference to harmful outcome) e.g., “an innocent person died”
Deontology			
Norm violation	0.12	0.74	Declares action a norm violation e.g., “he killed four people”
Obligation	0.04	0.81	Declares or denies an agent’s obligation e.g., “it is not the robot’s responsibility to save the workers”
Mental Agency			
Choice	0.36	0.88	Refers to the agent's deciding or choosing e.g., "He made the choice - for good or bad"
Difficult decision	0.1	0.84	States that the decision or situation was difficult e.g., "It's an impossible decision to take a life"
No good option	0.09	0.78	States that there was no good option to choose between e.g., “in both situations people were gonna die”
Intentional action	0.1	0.77	Characterizes what the agent did as an intentional or deliberate action e.g., "he intentionally killed the other miner"
Thought	0.05	0.88	Refers to the agent's beliefs, thoughts, or consciousness e.g., "The robot is thinking methodically"

***Note:* Rate refers to the proportion of responses falling into that category out of all justifications that participants offered.**

TABLE II. MAJOR CATEGORIES OF JUSTIFICATIONS FOR THREE TYPES OF MORAL JUDGMENTS

	Permissibility	Wrongness	Blame
Consequences	0.61	0.45	0.24
Deontology	0.12	0.21	0.13
Mental Agency	0.32	0.43	0.59

Note: The table entries refer to the proportion of participants who mentioned the given justification category for the given judgment. Proportions do not sum to 1.0 because people could refer to more than one category in their justifications.

III. RESULTS

We first tested the informativeness of justifications by examining the differential patterns of justification types (Consequences, Deontology, Mental Agency) across the three types of moral judgments (permissibility, moral wrongness, and blame), aggregated across Samples 1 and 2. Indeed, as Table II shows, justifications in the major categories varied systematically among the three judgment types. People justified their permissibility judgments predominantly by mentioning Consequences (good or bad); they justified their moral wrongness judgments by mentioning fewer Consequences but slightly more Mental Agency factors; and they justified their blame judgments predominantly by mentioning Mental Agency factors. These patterns are statistically reliable, as both the overall pattern and specific comparisons among any pair of judgments show significant deviations from chance, all ps < 0.01.

We next tested the informativeness of justifications by analyzing their patterns in detail for each judgment type separately: permissibility (assessed in Study 1), wrongness (assessed in Study 2), and blame (assessed in both studies and averaged here). Within each judgment type, we asked whether variations in justifications systematically and differentially reflected the moral judgments that they supported. Statistically, this can be analyzed by predicting (in reality, retrodicting) judgments from justifications. Assuming that the different moral judgments are grounded in different information processing [8, 9], such patterns of prediction would suggest that justifications are tied to or result from that information processing. Any differences or similarites between justifications for human vs. robot judgments would then illuminate people’s moral perceptions of robots.

To provide the context for these results, we briefly recapitulate, for each moral judgment, the results from Malle et al. [6], who focused entirely on response rates and means of the three judgments. Then we report the new results on the patterns of justifications predicting those judgments.

A. Permissibility

Previous results [6] showed that 65% of respondents found it permissible for the human agent to to direct the train toward the single miner whereas 78% found it permissible for

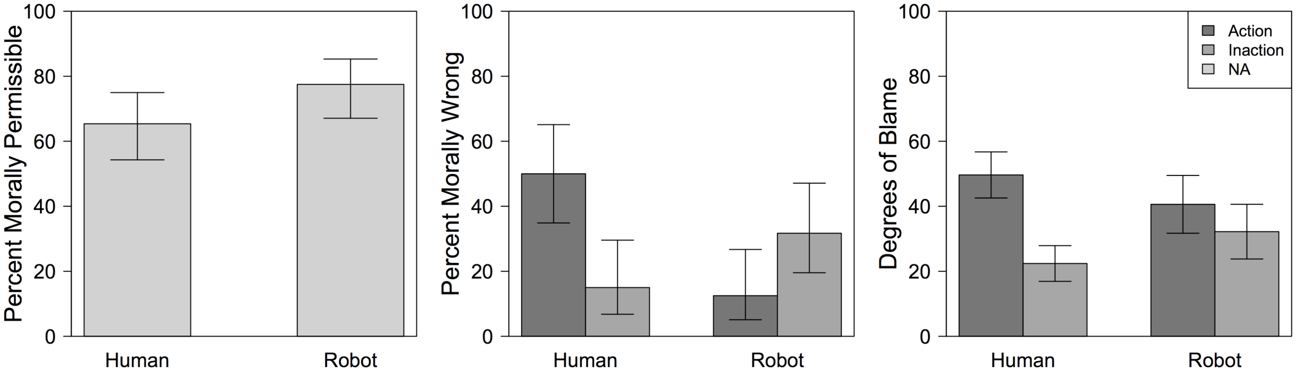

Figure 1. totalSummary of results from [6]: Permissibility judgments (left panel), wrongness judgments (center panel), and blame judgments (right panel) for human agent or robot agent who decided to intervene (“action”) or did nothing (“inaction”). Error bars indicate 95% confidence intervals.

the robot agent to do so, z = 1.80, p = 0.07 (Figure 1, left panel). Thus, most people accepted the sacrifice of one person for the benefit of four, but people held the robot to a norm that more readily embraced this costly sacrifice.

Turning to the analysis of justifications, Table II showed that, when justifying permissibility judgments, 61% of participants referred to Consequences, 12% referred to Deontology, and 32% referred to Mental Agency. More specifically, reference to Consequences predicted people *permitting* the sacrificing action (z = 5.92, p < 0.001) and reference to Deontology generally predicted people *not* permitting this action. However, Deontology was used differently for human and robot agents. People invoked deontological norms to justify *not permitting the human* to choose the sacrifice but to justify *permitting the robot* to choose the sacrifice (z = 2.80, p < 0.01). For example, to justify permitting the robot to choose the sacrifice, one participant wrote: "It would be a dereliction of duty to not to flip the switch [*sic*]." Thus, both people's rates of permissibility and their specific justifications suggest that they applied somewhat different norms to robots than to humans—norms that more strongly supported sacrificing one person for the good of many.

B. Wrongness

Previous results [6] showed that judgments of wrongness for the actual decision—to either sacrifice one life for many ("action") or do nothing ("inaction")—differed between human and robot agent. Evaluating the sacrificial action, 49% of people found it wrong when the human chose it but only 13% found it wrong when the robot chose it. Conversely, evaluating the inaction, only 15% found it wrong when the human chose it but 30% found it wrong when the robot chose it. This complete reversal was statistically reliable, z = 3.4, p < 0.001 (see Figure 1, center panel).

Table II showed that justifications for moral wrongness were distinct from those for permissibility. Consequences were still prevalent (45%) but less so; and both Deontological norms (21%) and Mental Agency were more frequent (43%). Importantly, people's offered justifications systematically reflected the wrongness judgments they had made. When people judged any decision (action or inaction) as wrong they offered more justifications in terms of both deontological norms (z = 3.06, p < 0.01) and Mental Agency (z = 2.16, p = 0.03). When people specifically judged the sacrificial action as wrong, they tended to cite the consequences of that action (z = 4.57, p < 0.001) and deontological prohibitions against the action (z = 2.34, p = 0.02). These justification patterns did not vary by agent type (robot or human). Thus, with different norms in place for robots and humans, people judged wrongness differentially for the two agents, but their justifications for those judgments operated the same for human and robot agents.

C. Blame

Previous results [6] showed that when evaluating the sacrificial action people blamed the human agent (M = 50) more than the robot (M = 40) but when evaluating inaction they blamed the human agent (M = 22) less than the robot (M = 32). This interaction pattern was statistically reliable, F(1, 312) = 11.62, p = 0.01 (see Figure 1, right panel).

Informative justifications for blame judgments should refer to the kind of information that blame judgments are based on, namely causal and mental-state information [8][16]. As shown in Table II, justifications were indeed especially rich in references to Mental Agency (59%), but more important, the number of such mental agency references reliably predicted the overall level of blame for any decision (action or inaction), F(1, 313) = 12.85, p < 0.001, and also the specific level of blame for the sacrificial action, F(1, 307) = 13.82, p < 0.001. By contrast, mentioning deontological norms did not significantly predict levels of blame (all ps > 0.1). None of these patterns varied by agent type (robot or human). Thus, with different norms in place for robots and humans, people blamed the two agents differentially, but their justifications for those judgments operated the same for human and robot agents.

IV. Discussion

In previous work [6] we found that humans apply different moral norms to human and robot agents. Compared to a human agent, a robot agent's decision to sacrifice one person to save four was judged more permissible, and if the robot decided *not* to take this sacrifical action, people tended to judge the decision as morally wrong and blamed the robot more. Here we show that people's justifications for these judgments are consistent with the overall interpretation that people hold robots to different norms in moral dilemmas. Aside from this difference in applying norms, however, people appear to process and justify their judgments of robot actions exactly the same as those of human actions. We briefly discuss the different norms that people apply to robots and then turn to an interpretation of the highly similar justifications people offered for human and robot agents.

A. The Human-Robot Difference in Norms

In [6] we offered two possible explanations for the different moral judgments people extend to robots. One is that people apply more strict norms to a human agent because they can easily put themselves into the agent's position and have a gut rejection against the physically aversive behavior of killing someone [18]. By contrast, people have trouble simulating the robot's position, don't have that gut reaction, and therefore are more lenient toward the robot. A second possible explanation is that a person's willingness to sacrifice another human being may endanger the person's reputation as a trustworthy social partner, and in their moral judgments people symbolically withhold trust to the person who embraces such a sacrifice. By contrast, participants don't consider the robot as part of a social community and thus judge only the action at hand, which by itself favors saving four lives even at the loss of one.

Our current results on justifications do not favor either of the above explanations of the human-robot difference in norms. However, the similarity of justifications for moral judgments rules out an alternative explanation: According to this explanation, people do not use the same kinds of moral judgments—or information processing underpinning those judgments—for robots and humans. Contrary to this account, we found systematic patterns of relationship between judgments and justifications for both human and robot agents.

B. *Justifications and Moral Judgments*

As shown in Table II, people offered different kinds of justifications when probed to make different kinds of moral judgments. Considerations of consequences dominated permissibility judgments, they became less important for wrongness judgments, and even less so for blame judgments. Considerations of mental agency, in contrast, were least important for permissibility judgments, they became more important for wrongness judgments, and were most important for blame judgments. These findings are consistent with extant models of blame and wrongness, which single out specific kinds of information that underpin the different kinds of judgments [7] [8] [16]. Whereas people use permissibility to assert a norm, they use wrongness to assess the violation of a norm and take into account possible justifying reasons the agent had in mind. Finally, blame judgments assess the total outcome of the norm violation relative to the agent's mental states—intentionality, reasons, counterfactual opportunities to prevent the outcome. The differential pattern of information is also consistent with theoretical frameworks in which moral judgments demand different degrees and types of *warrant* (i.e., explanations of the reasonable grounds for the judgment [19][8]). But whereas warrants for permissibility and wrongness primarily cite the pertinent norm and what the agent did to violate it, warrants for blame cite primarily causal and mental antecedents, which are well captured in the *Mental Agency* category of our present analysis. Whether people have direct and comprehensive access to the processes of moral cognition, the data here support an integrative view of how cognitive processes and social justifications relate to one another: People tend to report the same kind of information as social warrant for their judgments that theory and experiments say they use in making those judgments in the first place.

Ongoing research from our lab further reinforces the importance of justification in the social function of moral judgments. In a recent study we put pure versions of the three major justification types from the present studies into the mouths of a robot or human agent before we probed people's blame judgments. For example, when asked by a supervisor to justify the sacrificial action, the agent referred to consequences ("This way I preserved the most lives possible."), deontological norms ("My moral principles demanded that I save lives."), or mental agency ("I weighed the loss of life that I knew would result from either difficult choice."). We found that both deontological and mental agency justifications mitigated blame (compared to reference values from a control study) for both human agents and robot agents; and they did so particularly for the decisions that people had found objectionable—the human taking the sacrificial action and the robot refraining from such action. This finding further underscores people's willingness to include robots in the entire cycle of moral regulation: from detecting norm violations to blaming to reconciliation [19].

C. *The Human-Robot Similarity in Justifications*

An illuminating result from the justification data was that, despite applying different moral norms for how robots and humans should behave, participants provided similar types of justifications for their moral judgments. Notably, participants justified high levels of blame for both the human and the robot agent by citing the respective agent's choice capacity and other mental states. Not only does this show that people are willing to attribute these mental capacities to artificial agents [20][2], but it also suggests that these capacities are prerequisites for any agent, human or artificial, to be considered blameworthy [21].

We claimed earlier (see Figure 1) that no matter which results the present data offered, probing justifications for moral judgment would reveal something about how people judge social robots. Our data show, we believe, that people provide *informative* justifications for their judgments about human agents as well as about robot agents. To the extent that these justifications systematically differentiate and predict moral judgments we can conclude that people extend their moral concepts and reasoning processes to robot agents, even when the norms they apply are somewhat different. Because norms change far more quickly and easily than do psychological processes, the future may bring increasingly similar moral judgments for robot and human agents, as society integrates such artificial agents into its social and moral circles.

V. Conclusion

These experiments investigated how people justify their judgments of human and robot agents in difficult moral scenarios. We found that although people believe that robots and humans ought to behave differently when faced with the same dilemma, they relied on the same conceptual and justificatory foundation to make moral judgments about those agents. We take away one major lesson from these results. If we can create robots that are credible decision-making agents (with "mental agency"), people are likely to treat them as moral agents—which is to say, they will apply the same concepts, processes, and warrants when forming and explaining moral judgments of robot agents.

References

[1] Nourbakhsh, I. R. (2013). *Robot futures*. Cambridge, MA: MIT Press.

[2] Monroe, A. E., Dillon, K. D., & Malle, B. F. (2014). Bringing free will down to Earth: People's psychological concept of free will and its role in moral judgment. *Consciousness and Cognition*, *27*, 100–108.

[3] Kohlberg, L. (1984). *The psychology of moral development: The nature and validity of moral stages*. San Francisco, CA: Harper & Row.

[4] Mikhail, J. (2007). Universal moral grammar: Theory, evidence and the future. *Trends in Cognitive Sciences*, *11*, 143–152.

[5] Paxton, J. M., Ungar, L., & Greene, J. D. (2012). Reflection and reasoning in moral judgment. *Cognitive Science, 36*, 163–177.

[6] Malle, B. F., Scheutz, M., Arnold, T., Voiklis, J., & Cusimano, C. (2015, March). Sacrifice one for the good of many?: People apply different moral norms to human and robot agents. In *Proceedings of the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction* (pp. 117–124). ACM.

[7] Cushman, F. (2008). Crime and punishment: Distinguishing the roles of causal and intentional analyses in moral judgment. *Cognition*, *108*, 353–380.

[8] Malle, B. F., Guglielmo, S., & Monroe, A. E. (2014). A theory of blame. *Psychological Inquiry*, *25*, 147–186.

[9] Monin, B., Pizarro, D. A., & Beer, J. S. (2007). Deciding versus reacting: Conceptions of moral judgment and the reason-affect debate. *Review of General Psychology*, *11*, 99–111.

[10] Haidt, J. (2001). The emotional dog and its rational tail: A social intuitionist approach to moral judgment. *Psychological Review*, *108*, 814–834.

[11] Haidt, J., Björklund, F., & Murphy, S. (2000). Moral dumbfounding: When intuition finds no reason (Unpublished manuscript). Charlottesville, VA: University of Virginia.

[12] Nisbett, R. E., & Wilson, T. D. (1977). Telling more than we can know: Verbal reports on mental processes. *Psychological Review*, *84,* 231–259.

[13] Ericsson, K. A. and Simon, H. A. (1980). Verbal reports as data. *Psychological Review*, *87*, 215–251.

[14] Cushman, F., Young, L., & Hauser, M. (2006). The role of conscious reasoning and intuition in moral judgment. *Psychological Science*, *17*, 1082–1089.

[15] Hauser, M., Cushman, F., Young, L., Kang-Xing Jin, R., & Mikhail, J. (2007). A dissociation between moral judgments and justifications. *Mind & Language*, *22*, 1–21.

[16] Alicke, M. D. (2000). Culpable control and the psychology of blame. *Psychological Bulletin*, *126*, 556–574.

[17] Greene, J. D., Cushman, F., Stewart, L. E., Lowenberg, K., Nystrom, L. E., & Cohen, J. D. (2009). Pushing moral buttons: The interaction between personal force and intention in moral judgment. *Cognition*, *111*, 364–371.

[18] Cushman, F., Gray, K., Gaffey, A., & Mendes, W. B. (2012). Simulating murder: the aversion to harmful action. *Emotion*, 12, 2–7.

[19] Voiklis, J. and Malle, B. F. (in press). Moral cognition and its basis in social cognition and social regulation. In K. J. Gray and J. Graham (Eds.), *Atlas of Moral Psychology*. New York: Guilford Publications, Inc.

[20] Kahn, Jr., P. H., Kanda, T., Ishiguro, H., Gill, B. T., Ruckert, J. H., Shen, S., Gary, H. E., et al. (2012). Do people hold a humanoid robot morally accountable for the harm it causes? *Proceedings of the Seventh Annual ACM/IEEE International Conference on Human-Robot Interaction* (pp. 33–40). New York, NY: ACM.

[21] Roskies, A. L., & Malle, B. F. (2013). A Strawsonian look at desert. *Philosophical Explorations*, *16*, 133–152.