Against the Moral Turing Test: Accountable Design and the Moral Reasoning of Autonomous Systems

Thomas Arnold

Matthias Scheutz [1]

Department of Computer Science

Human–Robot Interaction Laboratory

Tufts University

Abstract

This paper argues against the moral Turing test (MTT) as a framework for evaluating the moral performance of autonomous systems. Though the term has been carefully introduced, considered, and cautioned about in previous discussions (Allen et al. in J Exp Theor Artif Intell 12(3):251–261, 2000; Allen and Wallach 2009), it has lingered on as a touchstone for developing computational approaches to moral reasoning (Gerdes and Øhrstrøm in J Inf Commun Ethics Soc 13(2):98–109, 2015). While these efforts have not led to the detailed development of an MTT, they nonetheless retain the idea to discuss what kinds of action and reasoning should be demanded of autonomous systems. We explore the flawed basis of an MTT in imitation, even one based on scenarios of morally accountable actions. MTT-based evaluations are vulnerable to deception, inadequate reasoning, and inferior moral performance

[1] thomas.arnold@tufts.edu, matthias.scheutz@tufts.edu

vis a vis a system's capabilities. We propose verification—which demands the design of transparent, accountable processes of reasoning that reliably prefigure the performance of autonomous systems—serves as a superior framework for both designer and system alike. As autonomous social robots in particular take on an increasing range of critical roles within society, we conclude that verification offers an essential, albeit challenging, moral measure of their design and performance.

Keywords Robot ethics**,** Artificial moral agents**,** Moral Turing test**,** Verification**,** Human–robot interaction

Introduction

The increased range and depth of artificial intelligence in human life has turned the hypothetical means of moral competence for autonomous systems into an urgent computational linchpin. Given its iconic place in the intellectual history of computation and artificial intelligence, Turing's legendary test for machine intelligence has not surprisingly surfaced as an organizing concept for evaluating this effort (Turing 1952). While the term can serve as shorthand for a wide range of ways to test how humans regard machines (as seen cinematically most recently in 2014's "Ex Machina"), the idea of a "moral" Turing test still leans on the premise and method of the original version (Allen et al.2000). Appealing to the term in the context of evaluating machine moral competence suggests that answering the question of whether machines think (the original Turing test) is analogous to, if not strongly correlated with, how to determine whether a machine is moral. Allen and Wallach have broached this concept (coining the acronym MTT for moral

Turing test), rightly noting some difficult problems an MTT would have to resolve in order to be viable (Allen et al. 2006; Wallach and Allen 2008). But while their discussion acknowledges that the MTT is "inadequate", the idea is still not wholly abandoned. It is left open as a possible means of tackling the larger, well-trodden question of what sources of ethical theory computational approaches to moral judgment should employ (e.g. deontology, consequentialist, virtue theory). More recent work by Gerdes and Øhrstrøm (2015) has continued this pattern, looking to an MTT as a general goal to "pass" on the way toward specifying the logic of ethical reasoning that an autonomous system would need to do so.

As the evocative phrase continues to pop up in public discussion, it is worth lingering on the very idea of a "moral Turing test" to determine what value it could have for moral evaluation and the design of autonomous systems (Henig2015). We argue in this paper that such a test, if it carries enough similarity to the original Turing test to deserve that name, ultimately and unavoidably rests on *imitation* as a criterion for moral performance. In turn, the kind of deceptive responses consistent with imitation, both as a representation of the agent and as a substitute for moral action writ large, undermines a more accountable, systematic design approach to autonomous systems. Secondly, we argue that even addressing a "behaviorist" bias in an MTT (for example through dialogue or reason-giving) will still be insufficient for evaluating the moral competence of an autonomous system. In view of the various high-stakes situations in which autonomous robots will likely be called upon and designed to act, and the moral responsibility inherent in moral evaluation itself (including that which still attaches to a system's designers),

we advocate a perspective of "verification": designing controlled, accountable, and accessible processes of moral reasoning on the part of autonomous systems. Verification seeks predictable, transparent, and justifiable decision-making and action, without black-box processes that keep the springs of a selected action out of human purview. And if the idea of an MTT is ultimately just a symbolic wish list of what design and verification will achieve, we conclude, it is too misleading and counter-productive to employ any longer.

The imitation game and the turn toward morality

Turing (1950) famously and inventively approaches the question "Can machines think?" through a novel and empirically inflected means—the test he introduces as an "imitation game". The game is introduced as one of an interrogator trying to tell if a respondent is a man or woman, with Turing then substituting a machine for the person A (a man). Though there has been debate about whether the machine is meant to imitate gender rather then just a "human", Turing's subsequent discussions suggest this is indeed about distinguishing human from machine (Moor2001). The game addresses a seeming lack of essential criteria for the predicate "think" by means of an argument via simulation—if a test subject's questions and comments cannot expose a machine's responses as different from a human's responses (after both of them respond to that test subject from closed rooms), then what more does the machine need to perform in order to be considered capable of thought? If the machine's responses cannot be found lacking—never mind in what—then either (1) some other criterion of thinking must be identified, or (2) it has succeeded in meeting a behavioral test for thinking.

The Turing test has seen ongoing debate about when it could be passed, if it has not been passed already, including Ray Kurzweil's (2005) prediction of premature announcements that it has been. The trumpeted "success" of chatbot Eugene Goostman has met considerable skepticism and dismissal, given the transcript of an admittedly bizarre and awkward conversation between his program and the human tester (Sample and Hern 2014). If some kind of Turing test is to establish the moral competence of a system's behavior, one might well imagine that it could be even longer before a human being is ready to attribute "goodness" to the observed conduct (never mind for the moment how disguised) of an artificial agent. To begin with, what kind of conduct, in what circumstances, could the tester employ to be adequately satisfied that a moral agent was being observed?

Allen et al. (2000) introduce the first substantive treatment of a moral Turing test for autonomous systems. In the course of surveying the tradeoffs between deontological and utilitarian views as guides for imparting morality to artificial moral agents (AMA's), they invert their perspective toward that as testing output:

> A moral Turing test (MTT) might similarly be proposed to bypass disagreements about ethical standards by restricting the standard Turing test to conversations about morality. If human `interrogators' cannot identify the machine at above chance accuracy, then the machine is, on this criterion, a moral agent.

Their ensuing discussion rightly points out that an MTT has some obstacles to clear as a plausible hurdle for an AMA. For one, the response format

privileges the "articulation of reasons" rather than the good of the system's behavior, a point that Allen et al. claim Mill might insist is crucial. They suggest that MTTs might have, instead of a verbal response, a comparison of actions between human and AMA. The test subject would be given certain actions and asked whether the agent was a human or machine—if the subject could not distinguish between the two than the MTT is passed.

Of course, Allen et al. recognize that the standards for AMA's might be higher than they are for the average person. They introduce the cMTT as one in which the machine has to be seen as no less moral than the human agent, behind the veil of non-specified agents for actions. That is in turn not foolproof, they admit, since the standard for an AMA might be *never* to be less moral than a person. It is at that point that they have largely let the idea rest, recognizing limitations in the cMTT but wondering if it might ultimately be the "only workable yardstick" (Wallach and Allen 2008) there remains to develop. Even in briefer, more recent appearances of this idea of an MTT, there is an ambivalence about whether to retain it as a horizon. In (2013) Wallach and Allen say "we accept that full-blown moral agency (which depends on strong A.I.) or even "weak" A.I. that is nevertheless powerful enough to pass the Turing test...may be beyond current or even future technology," which at least juxtaposes the test and criteria for morality. In the public discourse Wallach has mentioned the moral Turing test yet again in considering how robots will responsibly act in increasingly fraught social roles (Henig 2015).

Throughout these discussions Allen et al. have performed admirable work for laying out the dense network of questions and considerations that should go into robust moral standards for autonomous systems. They

insightfully show the tension between top-down, rule-bound approaches to morality and developmental, bottom-up models. More recently Gerdes and Øhrstrøm have picked up where that question is left off, exploring what kind of hybrid approach—top-down in terms of ethical theory, bottom-up in terms of neural networks and machine learning—could best approach the horizon evoked by an MTT (2013). Ultimately, these discussions have declined to grant MTT's sufficiency for moral attribution *per se*. Nonetheless, these discussions leave a number of considerations on the table unresolved, and it is time to begin sorting what aspects of a moral Turing test, if any, are essential, which ones disposable, and whether the success of an MTT is a matter of feasibility or just principled incompatibility.

Though Allen et al. may peer, however ambivalently, into a vista of increased feasibility, the argument here will confine itself to an ambit of principle. This is particularly important for a systematic analysis of the concept of the Turing test, whose open-ended nature toward criteria is suggestive of an ongoing struggle between adequate questions or comparisons and ever-improving sophistication on the part of the responding machine. The uncertain and compelling prospect of that struggle, along with the spontaneous developments it might promise to stage, risks occluding the constant features of the test itself, the structure that would apply no matter how advanced the test got as a moral arbiter. The limitations of those principles are crucial to investigate if the moral Turing test is to have any standing as a guide for AI and design.

It should be noted that this investigation diverges from how others working on computational rendering of moral reasonings have launched from the concept of MTT. As mentioned earlier, Gerdes and Øhrstrøm use the idea

to enter the larger issues of what inferences autonomous systems must execute to embody moral principle in important contexts (2015); in many ways, this reroutes the issue of MTT back into how a moral agent is analyzed in terms of capabilities, not necessarily tested as such (Floridi and Sanders 2004). Another take has been to argue a robot cannot, lacking the robustness of human ethical judgment, ever pass a true MTT (Stahl 2004). In either case, there is some measure of credence lent to the MTT, either as a signpost for future design or a threshold that blocks full ascription of moral agency. The problem with the MTT on principle, and how it misleads those and other efforts, is left unaddressed.

The argument that follows is that any MTT, if it is to retain the structural identity of the original Turing test, will face moral compromises because of (1) the role of imitation in achieving successful comparison, (2) the inaccessibility of moral reasoning, (3) the gap between reason-giving and action, and (4) the heightened moral demands for an autonomous, artificial system. Given the rich and varied directions in which computational architecture is drawing upon top-down logics and bottom-up machine learning, the objection to any MTT may seem overly stark, if not laboriously cautious. But for the overarching project of designing autonomous systems that perform in the best way possible for larger goods, a project for which Allen, Wallach and others have so ably laid out the moral landscape, it is critical to uproot where methods are lacking in concept, not just present-day form.

The moral dissociations of imitation

It is straightforward but crucial for considering Turing's testing proposal to note it relies on imitation behind some mode or mechanism of obscured identity. Though its meaning has spread beyond the behind-the-door response Turing proposed, one must track how its format applies to moral performance. Is the behavior that could lead to attributing thought to a respondent analogous with that of granting moral competence? Are the terms and conditions of imitation as applicable, appropriate, and advisable in the case of morality or ethics as they are to thinking? In terms of Allen et al., the general condition of imitation must be thought of both as (1) *verbal reason-giving or dialogue*, closer to the original Turing test, and (2) *narrated action*, by which the test subject is told of the system's action along with that of the human control.

One can naturally begin by pointing out that imitation alone, strictly speaking, is not sufficient to establish thought or morality themselves. The utterance that a test subject receives may come from any type of mechanism or random process, just as a real person could recite a fact instead of truly understanding what is expressed. Turing explicitly acknowledges this dimension to understanding, pointing out that ordinary examinations of a student's ability will seek "viva voce" answers to test whether there is genuine understanding of the topic at hand. One may learn a few talking points about astrophysics, but follow-up questions and requests for elaboration may expose the imitation as shallow and the supposed understanding as fake. In that

sense, the quality of the Turing test will vary with the depth and intricacy with which the test subject follows up with responses.

The cumulative aspect of the Turing test with respect to thinking suggests how performance can progressively strengthen the case for attribution. A Cartesian take on the Turing test, in fact, might make a point similar to one suggested about the performative character of the *cogito* (Hintikka 1962). At a certain point being able to pretend one is thinking, similar to the rehearsal of doubts that one is thinking, seems to reproduce, not just imitate, thought. If one's responses pick up concepts correctly and apply them in a non-repetitive and adaptive manner, with fidelity to current circumstances and without clear prior scripting, there seems less and less ground to deny that thought is occurring. This is not just because thought and language are so hard to pick apart in terms of performance (though if one wanted to test an agent's physical movements through some identity-obscuring visual filter, that might be an interesting case of motor intelligence and interpretation). It is also because the deception of presenting oneself as a person is perfectly consistent with the attribute of thinking, just as pretending to be a male is conceptually consistent with being a female (the original condition Turing uses to set up the imitation game for machines). It is possible, in other words, that a test subject end up surmising that an agent is a machine while being convinced that the preceding performance has reached the level of thought. For example, if the machine perfectly gauges a "natural" time for performing multiplication or division over and against its ability to perform it instantly, it might be more impressive than being a "quicker" thinker through its circuitry. Likewise coming to an incorrect answer through a compelling set of inferences may project a more

cognitive ability than a simple spouting of a fact. Though Turing does not dwell on these possibilities at length, models of cumulative machine learning fit perfectly well within the parameters of the test design (Johnson-Laird 1988). To conclude, with attributions of thinking *tout court* one can say that the Turing test may succeed cumulatively without being undermined by a basic deception.

How about for morality? What would success in a moral Turing test look like? In the abstract, the test would lead a subject with no knowledge of the actor's human or machine identity to judge that the actor was performing morally. Still, we can consider a basic disanalogy to testing thinking: moral actions are often not cumulative in leading to attribution. No matter how many sophisticated rationalizations one gives for a cruel or kind act, no matter how many kind actions one performs, the designation of a subsequent action as cruel or kind may not change. In fact, the attempt to give more and more reasons for an action, or cite more and more previous actions, may evidence more sophistication and deviousness than morality. Being thoughtful about a decision has no inherent connection to being more moral about the decision, and choosing one moral-seeming action may be but a means to performing a much worse one. Pretending to think seems to approach the attribution of thought through depth of performance, but the imitation of morality seems wholly indeterminate in terms of what would count.

Perhaps one stark way to frame the problem of deception in MTT is this question: when does the test actually begin? In the case of thinking, a test connotes a challenge of intelligence with identified parameters, a problem that could be solved. But does not moral attribution apply across all contexts of actions in which the respondent would operate/act?

If the MTT is about real-time responses to an interrogator, this is a serious flaw. The decision to participate in an MTT itself can carry a moral weight, for better or worse, that could affect one's ultimate verdict. Even granting that a moral system could participate, a strict MTT would open itself to that system's deception from the very first question (e.g. "Are you a machine?" or "What can a machine lie about and still be moral?"). Effective deception as to human/machine identity seems to have a tenuous relationship to the moral judgments of the system. It seems as if the MTT could never get off the ground if it strictly stuck to identity questions and evasive, convincing responses. This is a problem not only because of negative effects on trust—at what point do deceptive answers about who/what one is contribute toward a less confident attribution of moral character?—but because it seems such questions cannot reach into the real world of moral challenges. An effective MTT should test the full range of moral response, across many more fields of conflict and struggle than what kind of system the respondent is.

Even granting that the MTT should properly begin outside of self-referential dodges, a second problem arises for a moral evaluation based on an imitation game. Allen et al. rightly note that such a test will privilege "justifications" that seem convincing to the subject receiving them, but not necessarily provide concreteness in terms of what follows from them. We will delve more deeply into that issue later, but for now there is a related yet distinct problem that comes with justification. For justifications from a respondent may not only be too abstract when compared to concrete action. The justification given is also wholly indeterminate as to its reference, given the priority of imitation. That is, success in an MTT is about giving the justification that would convince the subject of the agent's conformity to

morality. If there were any reason to justify an action, the morality of the action is always secondary to the success of the imitation.

Allen et al. (2000) rightly point out the general problem of "reason-giving" for an MTT—there may be an inherent tilt toward universal duties in how a system presents its reasons, rather than particular calculations of utility. But they consider that the MTT might well differ from the original Turing test in being a confined inquiry into moral test cases. The subject would then ask the respondent what they would do when facing a moral challenge of one sort or another, and why they would do it. Does such confinement help deflect the quandary of general deception or self-presentation onto real issues that the system could show its moral fitness? While concrete cases are certainly a step in the direction of testing action, this direction alone does not escape the criticism. On the contrary, it only reinforces the inherent gap between a hypothetical reason, given under the terms of successful imitation, and the action that the system would actually carry out in such a scenario. That gap, to be sure, is itself theoretical. Still, the ultimate and fundamental test for a moral attribution, after all, lies not in hypothetical judgment or imagined action. It lies in the relevant action itself, reflected upon, decided upon, and performed in real circumstances. While a system may be able to arrive at good reasons for this or that action when presented with a narrative, that is consistent with moral hypocrisy and cowardice (where one does not act according to one's stated reasons and values, due perhaps to weakness, delusion, or cynicism).

Notice, then, that the flaw of possible deception is not remedied by the move toward a cMTT, as described in Allen et al. (2000). By changing the terms of the MTT to *compared actions*, the problem merely takes on the

mirrored flaw of apparent morality without justification at all. That is, in the cMTT the interrogator is not interacting with the agent's reasoning, but merely comparing actions of both systems and human control as narrated to the interrogator. How the agent intends the action, what counterfactuals the agent has in mind, what good the action is aiming at—these all drop out. The cMTT imitation game still holds open a gap between justifications as outputs on the one hand and the whole apparatus of practical reasoning and execution on the other. Even if a system's responses are not disturbingly deceptive about the system's identity, and even if a system's reasons are commonsense and compelling, the game's essential barrier to full identification and observation precludes a fully tracked process of decision-making and action, whereby one can see the system in action.

Does a total Turing test help?

Harnad (1991) proposes the total Turing test (TTT) to get at some of the issues around veiling or a lack of transparency that seem to inhere in an imitation game's fundamental structure. Instead of putting a robot behind some partition (however that is designed), the TTT brings the robot into the room, so to speak, to see if all its actions and appearances can convince the subject that they are dealing with a thinking being. Applied to the case of the MTT, this is an important consideration of embodiment. It is not just a judgment on a hypothetical scenario, or a secondhand consideration, that should be given the designation of an evidently moral action, but, as it were, action in the real world. The total Turing test is an important reminder of the embodied character of social interaction, the source for many verbal and non-verbal performances with moral implications and expectations.

The cost of a total Turing test, however, still hinges on the difference between imitation and ordinary, or non-contrived, action. An MTT has only so much force as it can reproduce, with minimal discrepancy, the actions valued and judged as moral in the real world. Scenarios or settings that bear little resemblance to the situations where difficult moral decisions are made would be harder to justify as genuine tests of moral reasoning. The closer an autonomous system's action amongst humans was demanded to achieve such reasoning, the closer those actions would themselves be morally implicated outright. In other words, the imitative element of the test increasingly drops out, and what we are dealing with is no longer a test but a robot acting with moral consequences in the real world. Especially for moral evaluation, a "total" Turing test dissolves the imitative presence of the Turing test. It seems harder and harder to say why imitation is still a relevant framework.

Human performance versus moral performance

As the idea of the MTT leads into considerations of real action, a final problem with imitation of human actions emerges. This has to do not with moral failure but with what one might call moral feats. Turing himself recognized that machines might actually misrepresent how well they could perform relative to humans—rapid, complex calculation, for instance—in order not to be thought mere machine. As mentioned earlier, Allen et al. recognize a moral parallel—what if a machine arrived at a morally superior judgment that seemed at first blush bizarre or wholly idealistic? Could human standards of morality, in other words, hamper the full moral capabilities of a system? This is an important point from Allen et al., but in their ensuing discussion the question of whether a MTT can account for that problem is left behind.

But simply raising the issue is not sufficient for appraising what an MTT could accomplish, nor does it determine whether an MTT has a useful role to play. Considered closely, the prospect of moral superiority from an autonomous system only reinforces the criticisms we have leveled so far, calling into further question the moral claims of MTT, its reliance of imitation, and its distance from genuine action. If a responding machine could be more moral than human beings in a certain context, it indicts from the start the parameters of an MTT. What moral attribution would be earned by a system that failed to present the action it took to be the best? Why would passing the test take moral priority over actually arriving upon the utmost moral response to a challenge? On the issue of imitation, if a reason given by a robot for action were seen as insightful and a moral step forward, it is not clear why that should be attributed to moral reasoning, rather than seen as an attempt to awe and impress the test subject to whom the machine is responding. What would make the wise response a moral one rather than just a compelling one, inasmuch as it reflects an attribute of the respondent? Finally, even if the justification were attributed to the machine's reasoning, what would establish the action itself, in the real scenario, as being what the respondent would carry out? The machine could well have misrepresented its abilities in order to earn moral status, rather than commit to what it could accomplish.

Asymmetries of the answer: arithmetic versus moral judgment

There is a final feature to the fundamental insufficiency of MTT that extends the point of robotic performance vis a vis human performance. Again, in his original proposal Turing recognized that there might be certain

responses that would betray a machine's superior ability, and he cites the obvious example of arithmetic. A problem that would take a person a certain amount of time to calculate (even with a handheld calculator!) could yield an immediate, correct answer from a machine. Turing muses that mimicking a human being might well necessitate masking the machine's quicker processing. As we have discussed, an MTT raises the prospect of exceeding human moral judgment, not replicating it. However, there is a thoroughgoing asymmetry between the case of arithmetic and a difficult moral scenario. What moral scenario, at least one that has any challenge to it, has an answer as unambiguous and correct as a multiplication problem? What answer is demonstrably correct for human beings? What means would establish that—an empirical survey or another ethical theory? The force of moral dilemma, for instance, lies in being both unavoidable as a practical choice but unresolved as a universal decision across all contexts and circumstances. In the case of arithmetic the process is almost irrelevant—the test is pure outcome, the solution. But for a moral judgment the lines between justification, prior information, counterfactuals (what, given the background information one had, one could have done, and with risks/opportunities), final decision, and actual execution tell a story of both process and outcome. So not only is there deep empirical disagreement in terms of how people would identify the right answer to a moral situation, but the way the decision is made is inalienable from a final response.

This complexity only increases from the fact that the very difference between a robot and human may guide moral judgments on actions. To see why, let us for the moment grant as much as possible to the objections to imitation and genuine action we have raised thus far. Suppose a robot could

display its reasons and inferences as transparently as possible. Suppose, too, that the robot was also somehow acting in as real an environment as one could imagine. Let us even suppose, however implausibly, that a basic consensus exists about what to do in a particular moral challenge, say some feasible variation of a trolley-car scenario. Even then, the final evaluation of a robot acting—with the accompanying judgments of what it would not share with a human being in that scenario—could yield some irresolvable discrepancy in what "the answer" would be from the agent being tested by an MTT. Recent work has suggested a robot may be judged along more utilitarian lines than a human being facing the same moral dilemma (Malle et al. 2015). Though there may be computational architectures that mirror how human beings make sensitive moral distinction between, say, weighing risks to strangers versus loved ones, that does not resolve whether a robot should ideally do so to take the most appropriate action for it (Wilson and Scheutz 2015). The difference between human and autonomous machine not only challenges the distinction between outcome and process in identifying what any answer is. It also may change what the right answer should be.

The limits of moral imitation games

To take stock, what becomes ever clearer through explicating the conditions of an MTT is that its imitative premise sets up an unbridgeable gulf between its method and its goal. With moral attribution there can be no black box keeping reason opaque from observed action—they come together for evaluation to make any moral attribution about the action and its actor. That is why the possibility of deception or deviation is impossible to set aside for the "real" MTT—it continues to loom over any overall agent attribution

because the data can still shape a final attribution. The ability to imitate, or look moral compared to a human, but not sincerely follow the moral norms cited—the black box underscores this inherent possibility no matter how complex the scenario or justification. Being regarded as thinking may be cumulative, but for the attribution of morality, a final failure of moral judgment and action risk the abject collapse of any moral authority—a mistake can make every act before it a possible deception, ruse, or at least a disturbing lack of integrity. Only in verification, as we now turn to explore, does the whole action come under testing.

Verifying moral action: accountability, transparency, and prediction

There is a common distinction in software design between testing and verification. Testing receives outputs and judges them primarily from a user's vantage, whereas verification looks at the whole system—design and performance, inside and out, as it were—to determine with certainty what outputs the system will produce and why. Given the problems built into the Turing test as a framework or perhaps even metaphor for moral evaluation, we propose that a better concept for determining moral competence is *design verification*. While the elusiveness of criteria for the attribute "thinking" led Turing to confining one's analysis to responses from behind a veil, a moral attribution must rely on more as an accountable, practical, socially implicated act of trust. To be accountable for a system's moral performance means going to as full a length as possible to verify its means of decision-making, not just judging *ex post facto* from a stated response or narrated action. Verification aims for transparent, accountable, and predictable accounts of the system's final responses to a morally charged context.

Making the process of reasoning explicit

An initial, overarching argument for verification is that it brings together what moral evaluation should not put asunder, which is to say the whole process of moral reasoning from perception to assessment to decision to action (and, as seen with cMTT, back again). What defines an action is not just what it accomplishes but why and on the basis of what assumptions and experiences, and as we have seen imitation alone cannot give access to all those dimensions. Tracking how an autonomous system receives information, assesses the moral dimensions of its environment, decides on and justifies the best action given the situation in which it finds itself, and executes that action means examining the computational integrity of the system along the entire way.

As far as the first part of that process is concerned, verification means overseeing how a system can size up a scenario in terms of possible actions, rules and principles, and utility measures. The basis on which possible actions are narrowed toward the best one is similarly explicit in terms of the criteria the system shows itself to employ. Keeping track of how relevant information is organized makes the basis for the final action taken more explicit and accountable. If there are facets of the environment (say, other people involved) that can mistakenly be ignored, that fact can be evaluated and incorporated into initial design, based on the way information is represented. The link between what the system recognizes and what actions it can have in its repertoire should be accessible for criticism and improvement. In Allen et al.'s original MTT this will prove a fault in not bridging reason-giving with ultimate performance; inversely, for their cMTT the shortcoming will lie in

not bridging action back to the reasoning and adaptations (including alternative plans) that "full-blown moral agency" entails.

In this way the argument for regarding practical reasoning as an integrated process finds a supporting argument for making the process, especially its key transitions, explicit. The design and verification of systems actually depend on the same perspective, the former only differing in how responsible one is for directly making the system. Verification can isolate the components of moral decision-making as points of strength and weakness, and break down where in a chain of operations there are problematic assumptions, lacunae of information, or neglected considerations. What range of contexts, possible actions, and reasons does the system have available to it? How does the system connect and consider those three in light of one another? A standard of outputs alone, applied via the Turing test format, leaves those crucial connections in the black box, with one observed response being compatible with a myriad of strong and weak candidates for how and why that response was given. Even with follow-up questions, the output from that black box may not authenticate the structures of evidence and inference that provide the ultimate authority for its response. The more of that process that remains hidden, the less one's moral evaluation will be complete.

To drive home this advantage of verification over imitation, it is worth considering the practical contexts where the moral decisions of an autonomous system will come under scrutiny. In the case of individual actions as a healthcare attendant, or rescue worker, or tutor, for example, a robot may be asked to explain why it performed the way it did. Part of that questioning will involve what the robot perceived and understood about the situation it faced, what actions that understanding suggested, and why a particular action

emerged as the best option. As with people, however, an adequate answer to those questions often entails addressing counterfactuals. If something had not been the case, how would one have acted? Being able to acknowledge the dependence of the best action on particular background conditions, including how conditions would have led to different actions, is a key component of showing how one's decision was morally justifiable. Unlike with imitative outputs, systematic verification can better establish how counterfactual responses correlate with defined logical operations and context assessment, rather than just sounding reasonable. As with basic perceptions on the moral stakes of a situation, the ability to locate and improve upon counterfactuals in moral reasoning depends on knowing where it fits in the system's architecture. Verification ensures design and performance stay in view, so that the moral character of action might likewise be judged singly.

Before moving onto justification we can make a brief excursus on what counterfactuals show about moral reasoning for people, not just artificial intelligence. For in probing an actor's counterfactual considerations, one is arguably taking a verificationist, not output-testing, stance on the agent. The logic behind one's decision goes beyond achieving, in a straightforward cause and effect way, a goal or abstract principle. The counterfactual function is to expand on one's moral landscape, to show how one's concepts and norms work within certain conditions (but not others) that one faced. Recent advances in neurological scans notwithstanding, a human being cannot offer an authoritative schematic of its cognitive and affective workings—how the person was going to decide to act, given certain conditions. But that is what counterfactuals push toward—a demand that an agent not just act correctly, but that the agent represent a more systematic ability to do so going forward.

Verification of how an autonomous system will act in a range of conditions could be more than a means of certainty—it might reasonably signify a valid moral deduction from the kind of counterfactual expectation we have of each other.

Justification and execution of an action

Perhaps the most important way that verification should be both integrated and transparent lies in the relationship between deciding on the most justified action and performing it. The structure of imitation means that the justification of an action, in the form of a response from behind a veil, is as close to an authoritative action as one can get. Because access to the workings of the system are denied, the system cannot show what might prevent it from carrying out what its computation has yielded as being what it should do. Not only is this compatible with giving deceptive justifications, but it raises the systemic question of action management—what if, in the presence of the actual scenario, the system did not initiate and see through what was proposed? In pressing for a failsafe link between justification and action through systematic design, verification represents the dual character of ethical evaluation with respect to autonomous systems. In terms of testing the autonomous system itself, verification lays out the measures by which the system will decide upon and execute action. At the same time, verification also speaks to the conditions by which the *activity of evaluating* is itself subject to moral account. A faulty or flawed test does not just redound to the system that takes it, after all—the evaluators who employ it must to some degree be responsible for the outcomes a false "pass" or "fail" could risk in society. Verification answers to both forms of responsibility, as must any

approach to testing and designing autonomous systems with moral capabilities.

Alternative modes of testing

The inapplicability of the moral Turing test does not invalidate the general idea of testing morally relevant aspects of a system. The fact that when moral considerations are at stake the reasoning and action should ultimately be viewed from the design end, not just the output end, does not preclude supplementary tests nested within verification. The key criterion for integrating such tests would be (1) adequate protection from risks that the system might eventually face, and (2) a systematic overview that would map processing and execution at each step of the system's performance.

Virtual reality

The possibility of an autonomous system's action uncoupled or fully indeterminate with respect its decision-making may lead one to imagine that a more immersive, real-life field of action than an MTT or cMTT provides. Instead of justifications, or compared actions as narrated, a virtual reality test would remove any disguise from the system's appearance—in this sense it would approach Harnad's TTT idea. There would be no veil, but the format of the test would still be that of seeing the system in real (or what the system takes to be real) action. Virtual reality of one sort of another, where a system presumably finds itself in a real environment in which to make a morally implicated decision and act upon it, might seem like a better, but still output-oriented, test than any MTT discussed thus far. Is there validity in such a test for demonstrating a reliable relationship between justification and action, without being full-blown verification?

Putting aside feasibility, as mentioned earlier, a virtual reality test is not altogether implausible in terms of showing in real-time what, for example, a robot's performance will look like and how others might experience it. In terms of a moral evaluation, however, the VR test still cannot stand alone without a systemic overview. To categorize the action that was being observed would in some cases depend on the representation of the agent—what is the agent trying to do, and why? Filling out this test would entail monitoring and following the system's processing throughout the virtual reality interactions, as much to see what action it found most justified and why as to see what action it executes. The methodological question, in other words, is why any VR test should retain a veil between processing and action, and what useful attributions or criteria are met by having it. In terms of ensuring that the system's action meet agreed-upon moral standards, it seems more likely that the VR test would just be a supplementary demonstration of a prior verification.

Subsystems

If a VR test or similar exercise were to have a useful reason for output veils, it might more justifiably be to test subsystems. Testing more statistically-oriented mechanisms, like speech or object recognition, might be an ongoing facet of gauging how a system's social interaction will work in social situations where the system's processing is not wholly accessible (Ball 2015). These tests may yield interesting implications for how overall action management might better operate. Nonetheless, the governing evaluation for which those component abilities are tested would still demand a systematic, predictable, and accessible orchestration of moral reasoning. Reasons cannot disappear into an*anomalous monism*, where impenetrable data

generates—through means left mysterious—a decision on the other side. Likewise for moving from justification to action—the testing of systems would involve improving how dependably a system is executing what its decision-making process has arrived upon as the best action.

Human–robot interaction

What the preceding considerations of VR and subsystems have touched upon is the complexity and uncertainty that inheres in human–robot interaction. One kind of output-exclusive testing that undoubtedly has use can be seen in psychological studies into how humans interact with autonomous systems in various contexts. Veiling the nature of a robot's awareness, skills, and autonomy can yield wide-ranging insights into what dynamics in human–robot collaboration will have to anticipate and negotiate. Such testing may be instructive about how robots will be received by human beings, and how their design should accommodate various reactions in particular work contexts. For ethically charged scenario of high-risk work and personal interaction, these will no doubt be very important (Scheutz and Malle 2014).

It bears emphasizing, however, that whatever Turing-like qualities those tests possess, they involve testing *human beings*, not the moral competence of autonomous systems. Much as the imitation game can yield interesting patterns in what type of language a subject would choose to test an agent's identity, a variant on an MTT could be useful for probing what people themselves associate with essential or revealing moral judgments. Obviously, again, that is no longer an MTT properly speaking—it is an exercise in human self-discovery.

Test case: the self-driving car

If there could be no universal moral Turing test, both because of general problems with deception and the gap between justification and action, could there not be some small-scale tests of ethical reasoning given concrete scenarios? Before exploring that possibility one must point out how difficult such concrete scenarios can be in terms of achieving moral consensus. In the case of recent polling about self-driving cars facing a difficult decision, it seems no test by a human subject would be thorough enough to be called a test for "morality" (Millar 2014). As the Roboethics Initiative found when asking about whether a car should hit a child in the road or risk the life of the driver/passenger, respondents were split not just on what they would have the car do but on how easy a decision they thought it was to make (2013). What moral attribution could apply to the autonomous system's final response or action, given that in either case it would violate a large contingent's "easy" moral judgment (Lin 2013)? Would the sophistication of the justification really do anything but increase the suspicion that the agent responding was just a really good rationalizer, a conniving immoral person?

Im**itative role-playing and machine learning**

One possible objection to verification as a means of systematic evaluation would be that it too stringently ignores the role of imitation in learning how to act. While a Turing test tout court might rely too much on imitation, will formal logics of obligation and utility, say, be as effective as less abstract means of machine learning in learning how to serve in one role or another? If iterative imitation and ongoing learning are central to how robots learn from humans how to act in one capacity or another (especially those with social subtleties of etiquette, affect, and sympathy), then is not the best

test of that process what the robot in fact can do? If a robot healthcare worker is supposed to escort a patient in pain down the hall to their room, isn't the walking itself the true test of whether they have learned it? This objection could build off a sociological point to the effect that robots will have particular jobs before being general "actors" in society. So instead of seeking a universal sense of "morality" across all possible contexts, should moral competence not be confined to the particular role (and imitation) that the robots is designed to fill?

This is an important consideration, both in terms of design approach and in terms of social contexts for robotic action. Robots will not likely emerge from manufacture as free-ranging citizens of the world, with no particular vocation or role to define their actions and decision-making. While the range of social robots can be wide, one must ask what limited set of tasks or objectives they are at minimum designed to accomplish. At the same time, the point against imitation still stands. When social robots have multiple considerations, and in many cases face complicated decisions that do not conform to one they have seen modeled repetitively by a human instructor, their systematic perceptions, inferences and action management take on moral dimensions that a robotic arm welding the same part of a car over and over will never face. The social companion robot may find itself in the presence of dire need out in public (a child who is injured, for instance). Some level of awareness and decision options in such scenarios will seem necessary. In those cases it will be all the more crucial that output-tests alone will not be the criterion of their learning up to that point. Designers must have some verification that in those complex contexts the robots is not just acting in a comparatively moral way when set beside a human in the same scenario. The

system must be able to reveal how it makes its responsible decision and show how it will act accordingly.

Moral superiority

One of Allen et al.'s reservations about the MTT, as mentioned earlier, has to do with the possible moral superiority of an autonomous system. How would an MTT account for a robot intentionally feigning not quite-so moral a judgment so that it would not seem unnaturally superhuman? Verification not surprisingly offers a better, if not completely satisfying, approach toward such moral stances. Since imitation is not the standard being sought, a verified moral action from a robot that seems distinctively insightful, inspiring, or helpful can still be accounted for on the basis of the means by which the system arrived at it. In fact, under verification such actions should be conceivable before they occur, at least along broad parameters of the system being able to make certain inferences if other conditions held true.

Verification, then, is not threatened by extraordinary action, but on the contrary might provide the systematic tools for such moral decisions to occur. If science fiction has accustomed us to anything, however, it is that the spontaneous and unexpected behavior from a robot or machine forms the exciting and dramatic part of their story. Inasmuch as a verification does not sit back and receive an out-of-the-blue moral performance or response, as might the tester in an MTT, it does not reward or privilege its occurrence: it seeks the whole process that constituted the reasoning it exhibits. To the degree these putative breakthroughs are fantasized, however, verification once again shows its superiority to an MTT—it stresses transparency and accountability for the design that penetrates the sensationalized and oversold spectacle.

Here one comes to a juncture where verification as a means of evaluating moral competence intersects with its role as a principle for design. Being able to establish a reliable unity between justification and actual performance, like authenticating an explicit, accessible means of moral reasoning to that justification, suggests that systems ought to be designed to meet a verification standard. Before unpacking further, it is worth reflecting more on verification, justification, and action. Are there some moral actions that systematic overview does not capture? Are there residual roles for an MTT or some form or another, which could discern a system's moral action while not demonstrating an absolute relation between moral considerations and its performance?

Whose morality is tested? Verification and design responsibility

It was noted before that verification carries both an ideal of systematic evaluation and a principle for design. Yet, in fact, moral evaluation is inseparable from the design of the system tested in every case. A system that generates certain responses without an explicit, accountable reason reflects a lack of investment in such a reason in the design itself. It is compatible with imitation that proves deceptive, because the power of the imitation (whether in plausibility, or range, or adaptability) is the prime criterion behind the design. While fleshing out an experimenter's individual ethics may be bootless, an experimental ethos with respect to morality exists by default in the design of autonomous systems with the ability to make complex decisions. The means of holding those decisions and actions accountable must correspond with how they occur in the system.

What verification underscores is that an accountable, autonomous system will possess the grounding of its moral reasoning from its designers. While developments in robotics and artificial intelligence have heralded increased autonomy with respect to real-time instruction or teleoperation, in terms of responsibility they do nothing to cut the tether between design and moral competence. What a system is allowed to do, with what level of transparency and accountability, with what attendant risks and opportunities: these always entail a responsibility on the part of the designer, no matter how complex the ensuing responsibilities from institutions, users, and participants prove to be. There may be ways in which a user, colleague, or encountered agent could abuse the system, through deceit, manipulation, or physical attack. But the design of that autonomous system's understanding, even on a trajectory of growth and increased sophistication in the long-run, is still always on the hook. A social, morally-charged interaction will always, however distal, draw upon some principle or parameter in accord with which the system performs. It is no accident that legal theorists have stepped forward to chart some of these currents, re-examining notions of negligence, liability, and use of force and proposing reasoned, case-derived distinctions for how to understand design, reasonable expectations, and chance in the course of events (Calo 2015; Pagallo 2013).

Safety

The kind of predictability that verification aims at establishing is practically bound by the risks that an autonomous system's action will incur. The risks of human harm at the hands of an autonomous system take on a heightened public significance, given that autonomy suggests power that

might exceed human prediction and control. Industrial accidents occur every day, but a mishap that involves a robot causing harm (proximately or not) commands much more media attention and reflection. While that may stem in part from exploitative sensationalism, there is a moral intuition behind the idea that robotic action takes on particular burdens of safety and certainty due to the abilities it exercises.

The kind of safety that moral competence respects and to which it conforms will vary across the roles autonomous systems will play. Social robots promise to develop roles in health care, education, law enforcement, military, and domestic companionship—the range of emotional and physical harm they could cause has not received as fine-grained attention as it should in a time of "killer robot" concerns (though robots with weapons is unquestionably a critical area for public debate). Small decisions will have rippling impact on a system's social milieu, from feelings of abandonment to physical brutality to lethal panic in a crowd in the middle of a rescue effort. The ability to predict how the system will manage highly-charged social action comes urgently to the fore when one considers these impacts. So, too, one cannot underestimate the care and accountability that must attend designing means for moral decisions on the part of these systems. The stakes are simply too high, over a range of interactions, for output-testing to be the barometer of whether a system is field-ready. Verification may be as important for what it stops from hasty introduction as it is what it announces as reliably moral.

Competence or the awe of creation?

Looming over the practical demands for autonomous systems' actions, with their continually developing roles, risks, and opportunities, is a larger-

scale cultural question, one that also situates the best way to evaluate those actions. What purpose does the design of autonomous system serve? Does the designer of a system, not just the users, have an additional set of needs or desires being served through its performance? An unmistakable drama and aspect of mystery attach themselves to an MTT. Will the system convince the subject? Will it be moral enough for us as observers? What will be the clinching response, what the crucial insight? It may be an exaggeration, but if so only slight, to say that an MTT preserves the drama of creation, of something attaining personhood, in a way that verification does not.

Verification has no tipping point, no epiphany, no one mark that sets the system into moral competence—it will require continuous adjustment, refinement, and deliberate planning. The critical points of verification will be as incremental as the gradual confidence their design earns and gradual risks their performance assumes. As many films have vividly depicted, the dynamic of "fooling" a test subject can mean stark and topsy-turvy revelations through success and failure, either through malfunction or an opaque, inaccessible moral framework ("I can't do that, Dave" is in perfect accord with the mission HAL is designed to fulfill). Without systematic oversight, the staging of an MTT may be a thrilling guessing game and adventure. It may even invite the idea of a moral "savant", who arrives at an insight through imponderable means and amazes society with what it finds. But where will such insight make its way into society's deliberations of what is right and what is good? At some point any autonomous system's action will have to meet systematic measures. Given the societal stakes, verification is a better design and test horizon to pursue from the beginning, rather then being a hurried retrofit. By verification designers can at all points stay mindful of accountable reasoning,

identifiable improvements, and the responsible performance what needs to be done. This is not to say that verification does not have its own challenges to overcome, most importantly scale: as systems get more complex, the verification of their subsystems together with comprehensive system integrity tests can themselves seem insurmountable. Yet, these are the efforts that we simply will have to pursue for the safer operation of autonomous systems (very much in the way airframe manufacturers have to verify their planes' autopilots and formally prove their operational profiles).

Experimentation and the values of inquiry

While there may be a danger of grandiosity and wonder at the prospect of an artificially intelligent agent, it is also fair to ask whether verification as argued for here does not have ethical weakness as well. Is verification not a fantasy of full control, a vision of ethics as being an entirely predictable conclusion given a set of complex premises afforded by a particular context? Will deontic logic, say, ever lead an autonomous system to a simple act of kindness, and can it possibly preclude a rule-bound moral obtuseness from taking a formally correct, but morally inferior course of action? How can verification predict behavior that is morally instructive, actions that will enhance moral standards and strengthen them beyond their present form? Given the criticisms leveled here against the basic form of the MTT, specifically the moral implications that extend beyond the confines of any imitation game, it is indeed fair to ask to what moral standard verification itself might be subject. Moreover, since other forms of testing may have supporting roles to play, what kind of experimental or testing ethos does verification uphold?

At the most general level, the stance of verification is to keep that very question open and explicit throughout design. The principle of opening any potential black-boxes of decision-making may introduce feasibility challenges (Pasquale 2015). Representing in digestible form the sum total of a system's decision-making and action-taking will certainly be easier said than done, but there are certainly promising directions we can take. For one, as already mentioned, we can require that each component of an integrated architecture routinely perform checks on its own operation. This may include performing basic tests of all of its services to ensure that despite possible changes resulting from adaptation and learning it will still produce the expected results (e.g., that an image processing algorithm will still be able to properly categorize a set of test images). Whenever such a component test fails, this would be the first indication that the system's performance might have been compromised. In addition, integration tests can be performed to ensure to groups of components will continue to work together as intended, again in the light of adaptation and learning, but also noise and possible system deterioration. All of these results can be routinely recorded and made available to introspection for both the system itself and the human in charge of it. Any deviation from the expected outcomes can then be addressed immediately either by the system itself or its operator. The ability to record its states, however, needs to be accompanied by mechanisms that ensure that state will be recorded, that such recordings are not optional. In that way sequences of computations leading up to a decisions given a set of inputs and the system's overall knowledge state will make the system fully accountable and prevent the system from cheating itself (e.g., by applying inappropriate rules or classification it might have learned during its operation).

In any case, accountability itself, the asking for reasons and the analysis of how reasoning works, depends on accessible junctures of proof, inference, and aesthetic precision (e.g. judgment of pain) to isolate and enhance as needed. The cost of verification for moral competence will be a burden of proof laid upon process, not product—an action that invites the label "moral" will not be enough to pull a process of reasoning into moral respectability. The moral vulnerabilities of the approach will come from possible actions "in the field" not being unleashed before a thorough examination deems them ready. In this respect the development of autonomous systems carries some similar ethical challenges as the development of drugs—what tradeoff exists in keeping a drug off the market when no one knows why it works? While for drugs that tradeoff has been made for releasing drugs that work before knowing exactly why, the intimacy and detail of our engineering autonomous systems, and the way the designer is tethered to the work a system goes on to do, makes verification a better default from which to depart with care. Medical researchers would prefer to know how an effective herb is working the way that it is to not knowing—why should designers recreate that condition with algorithms?

Conclusion

To meet the distinct and uncompromising demands for moral competence from an autonomous system acting in society, verification should supplant the moral Turing test as an organizing concept. Verification admittedly does not obviate thorny and often critical questions as to the agential status of autonomous systems. Nor does it imply that a logical scaffolding of obligations and norms has been found to settle the morality of

an action to its fullest extent, nor that it eventually will. If anything, the turn to verification is a turn toward more trenchant, grounded work, toward more thorough explorations of ethical theory, the scope and function of moral norms, and the best computational means to fulfill them. The acknowledgement of limitations and weaknesses beyond controlled contexts may be an ongoing ethical exercise in honesty for designers and users alike. Nonetheless, design more ably approaches the demand for moral competence when it does not depend on veils to secure evaluations of a system's actions and justifications. Without the system's entire response being observed and understood, any moral attribution is left foundering on questions of imitation, deception, and executive incompetence. The tragic contours of deception, imitation, and rationalization to which flesh is heir may loom large, but there is no reason for those same conditions to migrate in contrived fashion into the evaluation of artificial autonomous agents. The virtue of autonomous systems should be to allow for the precise opposite—predictable, controlled, and transparent decisions that allow for explicit reworking and recasting. Thus with autonomous systems do verification and design implicate and call each other to better account, toward a horizon of needs that beckons both.

References

Allen, C., Varner, G., & Zinser, J. (2000). Prolegomena to any future artificial moral agent. *Journal of Experimental & Theoretical Artificial Intelligence, 12*(3), 251–261.

Allen, C., Wallach, W., & Smit, I. (2006). Why machine ethics? *Intelligent Systems, IEEE, 21*(4), 12–17.

Ball, P. (2015). The truth about the Turing Test. *BBC*. http://www.bbc.com/future/story/20150724-the-problem-with-the-turing-test/.

Bringsjord, S. (1992). *What robots can and can't be*. Dordrecht: Kluwer Academic.

Calo, R. (2015). Robotics and the lessons of cyberlaw. *California Law Review, 103*, 2014-08.

Floridi, L., & Sanders, J. W. (2004). On the morality of artificial agents. *Minds and Machines, 14*(3), 349–379.

Gerdes, A. (2015). The issue of moral consideration in robot ethics. *SIGCAS Computers & Society, 45*(3), 274.

Gerdes, A., & Øhrstrøm, P. (2013). Preliminary reflections on a moral Turing test. In *Proceedings of ETHICOMP*(pp. 167–174).

Gerdes, A., & Øhrstrøm, P. (2015). Issues in robot ethics seen through the lens of a moral Turing test. *Journal of Information, Communication and Ethics in Society, 13*(2), 98–109.

Harnad, S. (1991). Other bodies, other minds: A machine incarnation of an old philosophical problem. *Minds and Machines, 1*(1), 43–54.

Harnad, S. (1992). The Turing test is not a trick: Turing indistinguishability is a scientific criterion. *ACM SIGART Bulletin, 3*(4), 9–10.

Henig, R. (2015). Death by robot. *New York Times*. www.nytimes.com/2015/01/11/magazine/death-by-robot.html.

Hintikka, J. (1962). Cogito, ergo sum: Inference or performance? *The Philosophical Review, 71*, 3–32.

Johnson-Laird, P. N. (1988). *The computer and the mind: An introduction to cognitive science*. Cambridge: Harvard University Press.

Kurzweil, R. (2005). *The singularity is near: When humans transcend biology.* Harmondsworth: Penguin.

Lin, P. (2013) The ethics of autonomous cars. *The Atlantic.*www.theatlantic.com/technology/archive/2013/10/theethics-of-autonomous-cars/280360/.

Lin, P. (2015). We're building superhuman robots. Will they be heroes, or villains? *Washington Post.*www.washingtonpost.com/news/in-theory/wp/2015/11/02/were-building-superhuman-robots-will-they-be-heroes-or-villains/.

Malle, B. F., Scheutz, M., Arnold, T., Voiklis, J. T., & Cusimano, C. (2015). Sacrifice one for the good of many? People apply different. In *Proceedings of 10th ACM/IEEE International Conference on Human–Robot Interaction.*

Millar, J. (2014) An ethical dilemma: When robot cars must kill,who should pick the victim? *Robohub.*robohub.org/an-ethicaldilemma-when-robot-cars-must-kill-who-should-pick-thevictim/.

Moor, J. H. (2001). The status and future of the Turing test. *Minds and Machines, 11*(1), 77–93.

Moor, J. (2009). Four kinds of ethical robots. *Philosophy Now, 72*, 12–14.

Open Roboethics Initiative (2014). My (autonomous) car, my safety: Results from our reader poll. *Open Roboethics Initiative.* http://www.openroboethics.org/results-my-autonomous-car-my-safety/.

Pagallo, U. (2013). *The laws of robots: Crimes, contracts, and torts* (Vol. 10). Berlin: Springer Science & Business Media.

Pasquale, F. (2015). *The black box society: The secret algorithms that control money and information.* Cambridge: Harvard University Press.

Sample, I., & Hern, A. (2014). *The Guardian.* www.theguardian.com/technology/2014/jun/09/scientists-disagree-over-whether-turing-test-has-been-passed/.

Scheutz, M., & Malle, B. F. (2014). "Think and do the right thing"—A plea for morally competent autonomous robots. In *Ethics in Science, Technology and Engineering, 2014 IEEE International Symposium on* (pp. 1-4). IEEE.

Schweizer, P. (1998). The truly total Turing test. *Minds and Machines, 8*(2), 263–272.

Stahl, B. C. (2004). Information, ethics, and computers: The problem of autonomous moral agents. *Minds and Machines, 14*(1), 67–83.

Turing, A. M. (1950). Computing machinery and intelligence. *Mind, 59*, 433–460.

Wallach, W., & Allen, C. (2008). *Moral machines: Teaching robots right from wrong.* Oxford: Oxford University Press.

Wilson, P. (1979). Utility-theoretic indexing. *Journal of the American Society for Information Science, 30*(3), 169–170.

Wilson, J. R., & Scheutz, M. (2015). A model of empathy to shape trolley problem moral judgements. In *Affective Computing and Intelligent Interaction (ACII), 2015 International Conference on* (pp. 112–118). IEEE.