Value Alignment, Fair Play, and the Rights of Service Robots

Daniel Estrada
New Jersey Institute of Technology
djestrada@gmail.com

Abstract

Ethics and safety research in artificial intelligence is increasingly framed in terms of "alignment" with human values and interests. I argue that Turing's call for "fair play for machines" is an early and often overlooked contribution to the alignment literature. Turing's appeal to fair play suggests a need to correct human behavior to accommodate our machines, a surprising inversion of how value alignment is treated today. Reflections on "fair play" motivate a novel interpretation of Turing's notorious "imitation game" as a condition not of *intelligence* but instead of *value alignment*: a machine demonstrates a minimal degree of alignment (with the norms of conversation, for instance) when it can go undetected when interrogated by a human. I carefully distinguish this interpretation from the Moral Turing Test, which is not motivated by a principle of fair play, but instead depends on imitation of human moral behavior. Finally, I consider how the framework of fair play can be used to situate the debate over robot rights within the alignment literature. I argue that extending rights to service robots operating in public spaces is "fair in precisely the sense that it encourages an alignment of interests between humans and machines.

Value Alignment and Turing's Test

A substantial portion of contemporary research into ethics and artificial intelligence is devoted to the problem of "value alignment" (hereafter **VA**) (Allen, Smit, and Wallach 2005; Yudkowsky 2008; Yampolskiy 2013; Soares and Fallenstein 2014; Russell, Dewey, and Tegmark 2015; Arnold, Kasenberg, and Scheutz 2017). Rather than deriving ethically appropriate action from first principles or from a direct recognition of the good, VA takes as its goal the (presumably simpler) task of designing AI that conforms to human values. AI that reliably conforms to human values is said to be "aligned". A primary concern in this literature is to establish methods that guarantee alignment, potentially within tight parameters, since it is argued that even small and seemingly innocuous cases of misalignment can quickly develop into a serious threat to general human safety (Yudkowsky 2008; Bostrom 2012; Babcock, Kramár, and Yampolskiy 2016).

There are reasons to be optimistic about VA as an approach to AI ethics, perhaps most significantly that the framework of "alignment" seems to lend itself to contemporary machine learning techniques like supervised learning (Mohri, Rostamizadeh, and Talwalkar 2012), where machines systematically improve their performance relative to a specified training set. There are also reasons to be skeptical that today's machine learning techniques are adequate for generating the complex forms of alignment required for participating in human moral communities (Arnold, Kasenberg, and Scheutz 2017). However, rather than critiquing the VA literature directly, my goal in this paper is to reflect on connections between the discourse on value alignment and the historical discussion of Turing's notorious "imitation game", with the hopes that lessons from the latter might better inform our developing discussions of the former.

Turing's test, originally offered as an alternative to the question "can machines think?", has since become a standard benchmark for evaluating the intelligence of machines (Turing 1950; Saygin, Cicekli, and Akman 2000; Copeland 2004; Copeland et al. 2017). The test revolves around a comparison to human performance: if the machine cannot be correctly identified by a human interrogator after a few minutes of conversation, it is said to "pass" the test and can be called intelligent. The central criterion for passing the test is *indistinguishability from human behavior* (Dretske 1997; Saygin, Cicekli, and Akman 2000). We might describe the demand for indistinguishability in terms of "behavioral alignment": a machine is behaviorially aligned just in case it behaves indistinguishably from a human. (Allen, Varner, and Zinser 2000) already recognized that since the set of moral behaviors is a proper subset of the total behaviors, what today is called "value alignment" can be interpreted as a special case of behavioral alignment. From this insight they propose a Moral Turing Test (**MTT**) (Allen, Wallach, and Smit 2006; Wallach and Allen 2008; Arnold and Scheutz 2016). The MTT is passed by a machine that behaves indistinguishably from a human in a conversation about moral actions.[1] Just as passing the original Turing Test is supposed to suggest a degree of intelligence on the basis of behavioral alignment, so

[1](Arnold and Scheutz 2016) argue that even a Total MTT variation, where evaluation of behaviors is not restricted to conversation but encompasses the full range of moral behaviors, is not sufficient for saving the MTT as a viable ethical criterion. For this reason, I will not dwell on the restriction to conversation behaviors in this paper. See (Harnad 1989).

too does passing the MTT suggest a degree of moral agency on the basis of moral alignment.

Although the Turing Test is widely known and discussed, it is generally not accepted as a reliable test for intelligence. Criticisms of Turing's test abound in the literature, perhaps best summarized by Dretske: "despite indistinguishability, all is dark..." (Dretske 1997). Critics worry that the mere imitation of human behavior is not sufficient for either intelligent or moral agency, and so Turing's test doesn't tell us what we want to know (Searle 1980; Dennett 1981; Dreyfus 1992). Here the theoretical goals of the MTT come apart from those of VA. Researchers concerned about value alignment don't care whether the machine is a genuine moral agent; a pure automaton (like the Paperclip Maximizer (Bostrom 2003)) might still pose a threat to humanity if it is sufficiently misaligned. And conversely, a sufficiently aligned machine is no guarantee of moral agency, just as convincing automation is no guarantee of intelligent agency. For this reason, strong rejections of MTT sit awkwardly in the literature alongside expansive research programs into the constraints on alignment, even while the former is a clear example of the latter. For instance, (Arnold and Scheutz 2016) criticize the MTT as a standard for building moral machines, and yet they go on in (Arnold, Kasenberg, and Scheutz 2017) to develop some theoretical constraints on applying machine learning to value alignment, while drawing no strong connections between these discussions. The effect is to make it appear as if the MTT is either irrelevant or unhelpful to the to the discussion of value alignment.

Arnold et al. reject the MTT on several grounds, including that imitation cannot serve as the basis for intelligent moral agency. Echoing the traditional criticisms, they write, "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" (Arnold and Scheutz 2016) The framework of an "unbridgeable gap" is familiar from the philosophy of mind (Dennett 1991), and seems to render Turing's proposal inadequate for the task of developing genuine moral agents. However, if our task is not to develop intelligent moral agents *per se* but merely to align our machines with our values, than the MTT may continue to prove useful. In the next section, I argue that Turing's principle of "fair play for machines" (**FP**) (Turing 1947) provides a non-imitative ground for evaluating the alignment of machines. I argue that the FP avoids many of the classic criticisms of Turing's test, and provides a satisfying method for applying Turing's insights to the problem of value alignment distinct from the MTT.

Fair Play for Machines

(Proudfoot 2017) points to a variety of sources in developing a rich, comprehensive account of the Turing test ("from every angle"). Special emphasis is given to his discussion of a "little experiment" (Turing 1948) involving chess playing computers in an experimental design that is clearly the germ for his landmark 1950 paper. However, curiously missing from Proudfoot's analysis (and mentioned only in passing in (Leavitt 2017)), is a short passage from the end of Turing's 1947 Lecture on the Automatic Computing Engine to the London Mathematical Society (Turing 1947; Copeland 2004; Hodges 2012). Here Turing is also concerned with evaluating the performance and "I.Q." of chess-playing computers, which suggests this passage should be read alongside his 1948 and 1950 papers for a full appreciation of the developing proposal (Estrada 2014). Since it is so regularly overlooked, I quote Turing's argument below in full, with paragraphs and emphasis added:

> "It might be argued that there is a fundamental contradiction in the idea of a machine with intelligence. It is certainly true that acting like a machine has become synonymous with lack of adaptability. But the reason for this is obvious. Machines in the past have had very little storage, and there has been no question of the machine having any discretion. The argument might however be put into a more aggressive form. It has for instance been shown that with certain logical systems there can be no machine which will distinguish provable formulae of the system from unprovable, i.e. that there is no test that the machine can apply which will divide propositions with certainty into these two classes. Thus if a machine is made for this purpose it must in some cases fail to give an answer. On the other hand if a mathematician is confronted with such a problem he would search around a[nd] find new methods of proof, so that he ought eventually to be able to reach a decision about any given formula. This would be the argument.
>
> **Against it I would say that fair play must be given to the machine.** Instead of it sometimes giving no answer we could arrange that it gives occasional wrong answers. But the human mathematician would likewise make blunders when trying out new techniques. It is easy for us to regard these blunders as not counting and give him another chance, but the machine would probably be allowed no mercy. In other words then, if a machine is expected to be infallible, it cannot also be intelligent. There are several mathematical theorems which say almost exactly that. But these theorems say nothing about how much intelligence may be displayed if a machine makes no pretense at infallibility.
>
> To continue my plea for fair play for the machines when testing their I.Q. A human mathematician has always undergone an extensive training. This training may be regarded as not unlike putting instruction tables into a machine. One must therefore not expect a machine to do a very great deal of building up of instruction tables on its own. No man adds very much to the body of knowledge, why should we expect more of a machine? Putting the same point differently, **the machine must be allowed to have contact with human beings in order that it may adapt itself to their standards**. The game of chess may perhaps be rather suitable for this purpose, as the moves of the machines opponent will automatically provide this contact." (Turing 1947; Copeland 2004)

There are many striking things to note about this passage. First, Turing is responding to a critic of the very idea of machine intelligence, whose argument points to some necessary (and therefore unbridgeable) gap between the performance of humans and machines. In this case, the critic appeals to Gödel's incompleteness theorem (Gödel 1931; Smullyan 2001) as evidence of such a gap, an objection he returns to under the heading of "The Mathematical Objection" in (Turing 1950). Recall that Turing's major mathematical contribution (Turing 1937) is the formal description of a "universal computer", which can in theory perform the work of any other computer. On my interpretation (Estrada 2014), the universality of his machines is what ultimately convinces Turing that computers can be made to think. Without any assumption of behaviorism or appeal to a principle of imitation, the syllogism runs as follows: if the brain is a machine that thinks, and a digital computer can perform the work of any other machine, then a digital computer can think. This syllogism is both valid and sound. However, Turing recognizes that Gödel's theorem shows "that with certain logical systems there can be no machine which will distinguish provable formulae of the system from unprovable". This straightforwardly implies that there are some things that even Turing's universal machines cannot do. This result does not invalidate the syllogism above. Still, Turing's critics draw an inference from (1) there are some things machines cannot do, to (2) humans can do things that (mere) machines cannot do. Although this inference is clearly invalid,[2] arguments of this form persist even among respected scholars today (Penrose 1999; Floridi 2016). The passage from his 1947 Lecture shows Turing contending with this perennial challenge several years before his formal presentation of the imitation game. In other words, Turing was clearly aware of an "unbridgeable gap" objection, and both his "little experiment" and the principle of fair play serve as ingredients in his response. A full appreciation of Turing's position in this debate ought to take this evidence into account.

Second, the core of Turing's response is to offer a "plea" for what he calls "fair play for machines". This suggestion is proposed in the context of "testing their I.Q.", making explicit the connection between FP and the developing framework of Turing's test. Essentially, Turing is worried about a pernicious double standard: that we use one standard for evaluating human performance at some task, and a more rigorous, less forgiving standard for evaluating the machine's performances *at the same task*. Other things equal, a double standard is patently unfair, and thus warrants a plea for "fair play". Of course, one might worry that the mere fact that the performance comes from a machine implies that other things *aren't* equal. Since machines are different from humans, they ought to be held to different standards. But on my interpretation , Turing is primarily motivated by a conviction that universal computers can perform the work of any other machine, and so humans and computers are not essentially different. Turing's test isn't designed to prove that machines can behave like humans, since in principle this follows from the universality of the machines. Instead, the test is designed to strip the human evaluator of his prejudices against the machines, hence the call for fair play.

[2]Turing's original response to the Mathematical Objection remains satisfying: "The short answer to this argument is that although it is established that there are limitations to the powers of any particular machine, it has only been stated, without any sort of proof, that no such limitations apply to the human intellect." (Turing 1950)

Notice that calling a standard of judgment unfair does not imply that the machines treated unfairly can "think". Therefore, FP cannot serve as a basis for evaluating the intelligence of machines in the style of the Turing Test. And indeed, Turing's argument makes clear that his appeal to FP is concerned not with the intelligence of the machine, but instead with the standards used to evaluate the machine's performance. After all, Turing's plea is made in defense of machines that are expected to be infallible, and whose performance might be compromised (by "occasionally providing wrong answers") in order to more closely approximate the performance of a human. Turing's point is that we'd never demand a human mathematician to occasionally make mistakes in order to demonstrate their intelligence, so it's strange to demand such performance from the machine. If Turing's test is motivated by a call for "fair play for machines", this should inform our interpretation of the test itself. Since the principle of fair play does not depend on an imitative premise, the rejection of Turing's test on this basis seems too hasty.

Finally, the quoted passage closes by highlighting the phrase "fair play for machines" again[3], and arguing that "the machine must be allowed to have contact with human beings in order that it may adapt itself to their standards." Clearly, Turing is approaching the challenge of evaluating the performance of machines, even in purely intellectual domains like chess, as a problem of behavioral alignment. Moreover, Turing argues that for machines to achieve that alignment, *they must be allowed* certain privileges, in the interest of "fair play". Specifically, Turing argues that if we expect the machine to learn our standards, we must afford access to our behavior. In other words, he's arguing that constraints on *human behavior* are necessary to achieve alignment: in how we evaluate and interact with out machines. This perspective is rare even in the alignment literature today, where concerns are overwhelmingly focused on how to constrain the machine to stay within bounds of acceptable human behavior.

More importantly, Turing suggests that we must be willing to interact with machines, even those that aren't intelli-

[3]It may be interesting to consider why the phrase "fair play for machines" doesn't appear in the 1950 paper. Many of the arguments from the passage appear in the final section of his 1950, under the subsection "Learning Machines", where he proposes building "a mind like a child's" in response to Lovelace's Objection (Estrada 2014). In this section he proposes a number of games to play with machines, including chess and twenty questions. He also expresses a worry that "The idea of a learning machine may appear paradoxical to some readers." Turing's 1950's paper is self-consciously written to a popular audience; perhaps Turing worried that a plea for "fair play for machines", including those that aren't even intelligent, might also confuse his readers too much, and undermine the constructive argument he's given. Hopefully, readers 70 years later are not so easily confused.

gent, if we expect these machines to align to our standards. And this is precisely the kind of interaction Turing's proposed imitation game encourages. These reflections open a new route to defending the importance of Turing's test in today's alignment literature. Turing's test is usually understood as a benchmark for intelligence, and the MTT as a benchmark for moral agency. Commentary traditionally recognizes Turing's worries about standards of evaluation (Saygin, Cicekli, and Akman 2000; Arnold and Scheutz 2016), but they interpret Turing's imitation game as attempting to settle on some specific standard of evaluation: namely, indistinguishability from human performance, or perfect imitation, as judged by another human. If the machine meets this standard, the machine is considered intelligent. We might call this a "benchmark" interpretation of Turing's test, or **BTT**. The MTT is an instance of BTT that sets the benchmark to imitate human moral behavior, for example. Many machine learning applications today will present themselves as meeting or exceeding human performance (at discrimination tasks, image recognition, translation, etc.), a legacy of Turing's influence on the field. Criticisms of Turing's test focus on whether this benchmark is appropriate for evaluating the machine's performance, with most concluding it is not an adequate measure of general intelligence. But the principle of fair play suggests Turing is less interested in setting a particular benchmark for intelligence, and more concerned with establishing that the standards for evaluation that are fair. Call this interpretation the Fair Play Turing Test **FPTT**. A machine passes the FPTT when it meets the same standards of evaluation used to judge human performance at the same task. On this interpretation, Turing's imitation game is meant to describe a scenario of "fair play" where the human biases against machines can be filtered out, and the machine could be judged in their capacity to carry a conversation by the same standards as any other human. We typically think of someone who can hold a conversation as being intelligent, so if a machine can also hold a conversation without being detected as non-human, we should judge it intelligent too. Not because conversation is some definitive marker of intelligence, as the BTT interpretation suggests, but rather because conversation is a standard that is often used to evaluate the intelligence of humans, and the principle of fair play demands holding machines to the same standards. On this interpretation, the sort of hostile interrogation typically seen in demonstrations of Turing's test (Aaronson 2014) seems straightforwardly unfair, since we wouldn't expect an intelligent human to hold up well under hostile interrogation either.

Since the principle of FP does not depend on imitation, the FPTT works in a subtly different way than the BTT. Passing the FPTT doesn't merely imply a machine performs at human levels; it implies more strongly that the machine performs at these levels *when evaluated by the same standards* used to judge human performance. For instance, we usually aren't skeptical of mere imitation when talking to a human, so raising this concern in the context of evaluating machine could signal a change in the standards of evaluation, and thus a violation of FP. Perhaps a double standard is warranted in cases where machine performance is expected to diverge significantly from machines. We might, for instance, expect driverless vehicles to adhere to more rigorous safety standards than we typically hold human drivers. Recognizing these misaligned standards as a violation of fair play doesn't necessarily imply the situation is unethical. Instead, it draws attention to the multiplicity of standards at play, and the lack of a unifying, consistent framework for evaluating all agents. The framework of FPTT easily extends to evaluating performance at tasks other than "general intelligence", including the task of moral alignment in particular contexts. (Arnold and Scheutz 2016) reject the MTT as a standard for evaluating moral agency on the basis of its imitative premise. But FPTT doesn't depend on an imitative premise, and only checks for alignment within the standards used to judge humans at a task. In the next section, I argue that this framework of fair play has direct application for evaluating the alignment of robots operating in our world.

The Rights of Service Robots

The question of robot rights has historically turned on questions of legal personhood (Gunkel 2012; Bryson, Diamantis, and Grant 2017). Conditions on personhood typically involve both cognitive and moral attributes, such as "recognizing the difference between right and wrong" (Christman 2008). The consensus among scholars is that robots do not yet meet the conditions on minimal personhood, and will not in the near future. However, this consensus is inconclusive, and has been used to argue that robot rights might be necessary to protect machines that operate below the level of human performance (Darling 2012; 2015). For instance, in 2017 San Francisco lawmakers implemented restrictions on autonomous delivery services on sidewalks and public right-of-ways, citing safety and pedestrian priority of use as motivating concerns (Rodriguez 2017). The proposal raises a natural question of whether these robots have the right to use public spaces, and to what extent a ban on robots might infringe on those rights. These questions seem independent of more general concerns about moral agency and personhood that typically frame the rights debate. Furthermore, it is well known that service robots operating in public spaces are typically subject to bullying and abusive behavior from the crowd ((Salvini, Laschi, and Dario 2010; Salvini et al. 2010; Brscić et al. 2015). Protecting robots from such treatment seems necessary independent of whether they meet strict conditions for personhood.

Like many cases of moral alignment, the rights for service robots to operate on public sidewalks seems to demand a framework for evaluating the performance of machines that does not turn on any imitative premise. Delivery robots do not have nor require the intellectual and moral capacities typical of humans; to compare their operation with human performance seems at best mismatched, at worst insulting. Interpreted as a benchmark of performance, these machines operate well below the threshold where Turing's test is relevant and the vocabulary of rights and personhood applies. However, in contrast to the benchmark interpretation, the principle of fair play asks that standards of evaluation are consistent across humans and machines. In the case of service robots, the question turns on the nature of the task these robots are performing, and the standards already in use for evaluating such performances. There's an obvious comparison between service robots and service animals that is tempting, but I think is ultimately unhelpful. Importantly, animals feel pain and suffer, and service animals are used to support persons with disabilities who can't otherwise access public resources. Service robots, in contrast, are used by tech companies to better service their clients. Given the distinct nature of these tasks, holding service robots to the standards of service animals seems inappropriate.

A closer analogy to the work of service robots can be found (Chopra and White 2011), who propose an alternative approach to robot law centered not on legal personhood but instead on a framework of legal agency. A legal agent is empowered to act on behalf of a principal, to whom the agent holds a fiduciary duty that contractually binds the agent to act in the principal's interest. For instance, a lawyer or accountant operates as a legal agent in service of their clients. In the context of agency law, an agent's right to operate turns both on the capacities of the agent to faithfully represent the principle, and also on the nature and scope of the role . The framework of agency law offers a systematic defense of robot rights which focuses legal and policy attention to the roles we want robots to play in our social spaces, and the constraints which govern operation of any agent performing these roles (Chopra and Estrada in progress). A social role analysis of robot rights has clear application to the protection of service robots operating in public spaces, including delivery robots and self-driving cars. But it also has natural extensions for the regulation of robots in a wide variety of other social roles, including robots that provide services in the context of, for example, law and justice, finances, transportation, education, socio-emotional support, sex work, public relations, and security. For instance, agency law provides a straightforward path to the regulation of bots on social media that are used to influence voters and elections (Ferrara et al. 2016). From this perspective, social media bots are operating on behalf of their operators in the service of specific roles (campaign promotion, electioneering, etc), and therefore fall under the same legal frameworks that already exist to evaluate the ethics and legality of these agents.

Much more deserves discussion on the proposal for regulating robot rights under the framework of agency law (Chopra and Estrada in progress). I raise the proposal in this context to demonstrate how the principle of fair play provides a strong and satisfying justification for developing standards within agency law for consistent application across human and machine performances. Recall that the FPTT simply asks that we evaluate the machine according to the same standards of evaluation used to judge the performance of a human at the same task. Thus, FPTT focuses our discussion on the task-specific standards for evaluation, rather than on the details of the performance of any particular machine. In this way, FP also motivates an expansive research agenda for classifying and detailing the types of roles we might want robots to serve in, and the constraints on evaluating the performance of any agent filling that role. For instance, what should social media bots acting as campaign representatives be allowed to say or do? This is not an engineering question about the capabilities of any machine. It is a social policy question about what machines can and cannot do in the service of their role. And if we want machines to align to our standards of performance, then Turing demands that fair play must be given to the machines.

Of course, we don't want every machine to align them to our standards. For a simple example, a factory robot can produce more products than their human counterpart, and shouldn't be held to the same standards of performance as a human. If standards cannot be made consistent across humans and machines, it implies a kind of stratification of the social order that cleanly divide humans from machines on the basis of roles served. Indeed, one might worry that a social role defense of robot rights is ultimately not in the interests of robots, since it might be seen as imposing a necessary second-class status on the machines. In this way, a social role defense might appear to be a kind of "3/5th compromise for robots". This worry is reinforced by a review of the history of agency law, which itself develops out of a logic of slavery and indentured servitude (Johnson Jr 2016). However, the social role analysis of agency law provides a way around this worry by laying out a direct path to full legal agency for robots. To say that robots serve as agents for principals does not preclude the robot from being a full legal agent, since obviously lawyers and accountants retain their personhood and agency even while acting as an agent for their principal. But of course, serving as a principal is yet another social role to perform, one with its own standards of evaluation. Making the the standards for principals explicit allows for a robot to serve as principal, first for other robots, perhaps as a manager or oversight for other robots acting on its behalf, and eventually to serve as a principal for itself, thus bridging the gap to full legal agency.

Conclusions

In this article I have reviewed some primary concerns of the value alignment literature, and shown these interests were present in the development of Turing's test as early as (Turing 1947). Still, I argued that a widespread rejection of Turing's test as a standard of intelligence has led scholars to overlook Turing's test as a source of inspiration in developing machines that are value-aligned. I have proposed an alternate interpretation of Turing's test, FPTT, which is inspired by Turing's call for "fair play for machines". I carefully distinguish FPTT from the BTT interpretations like the MTT. Finally I have briefly discussed how FPTT might be used to justify a defense of robot rights and sketch out a path to full agency on the basis of a social role analysis of agency law.

Acknowledgements

Thanks to conversations with Jon Lawhead, Samir Chopra, Kyle Broom, Rebecca Spizzirri, Sophia Korb, David Guthrie, Eleizer Yudkowsky, Eric Schwitzgebel, Priti Ugghley, richROT, the participants in my AI and Autonomy seminars and the Humanities department at NJIT, all my HTEC students at CTY:Princeton, and everyone in the #botally and Robot Rights communities across social media, especially David Gunkel, Julie Carpenter, Roman V. Yampolskiy, Damien Patrick Williams, and Joanna Bryson.

References

Aaronson, S. 2014. My conversation with “eugene goostman,” the chatbot that’s all over the news for allegedly passing the turing test. *The Blog of Scott Aaronson*.

Allen, C.; Smit, I.; and Wallach, W. 2005. Artificial morality: Top-down, bottom-up, and hybrid approaches. *Ethics and Information Technology* 7(3):149–155.

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

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

Arnold, T., and Scheutz, M. 2016. Against the moral turing test: accountable design and the moral reasoning of autonomous systems. *Ethics and Information Technology* 18(2):103–115.

Arnold, T.; Kasenberg, D.; and Scheutz, M. 2017. Value alignment or misalignment–what will keep systems accountable? In *3rd International Workshop on AI, Ethics, and Society*. AAAI.

Babcock, J.; Kramár, J.; and Yampolskiy, R. 2016. The agi containment problem. In *International Conference on Artificial General Intelligence*, 53–63. Springer.

Bostrom, N. 2003. Ethical issues in advanced artificial intelligence. *Science Fiction and Philosophy: From Time Travel to Superintelligence* 277–284.

Bostrom, N. 2012. The superintelligent will: Motivation and instrumental rationality in advanced artificial agents. *Minds and Machines* 22(2):71–85.

Brscić, D.; Kidokoro, H.; Suehiro, Y.; and Kanda, T. 2015. Escaping from children’s abuse of social robots. In *Proceedings of the tenth annual acm/ieee international conference on human-robot interaction*, 59–66. ACM.

Bryson, J. J.; Diamantis, M. E.; and Grant, T. D. 2017. Of, for, and by the people: the legal lacuna of synthetic persons. *Artificial Intelligence and Law* 25(3):273–291.

Chopra, S., and Estrada, D. in progress. *Does a robot ban violate robot rights? A social-role analysis*. University of Michigan Press.

Chopra, S., and White, L. F. 2011. *A legal theory for autonomous artificial agents*. University of Michigan Press.

Christman, J. 2008. Autonomy in moral and political philosophy. *Stanford encyclopedia of philosophy*.

Copeland, J.; Bowen, J.; Sprevak, M.; and Wilson, R. 2017. *The Turing Guide*. Oxford University Press.

Copeland, B. J. 2004. *The essential turing*. Clarendon Press.

Darling, K. 2012. Extending legal protection to social robots: The effects of anthropomorphism, empathy, and violent behavior towards robotic objects. In Calo, Froomkin, K., ed., *We Robot Conference 2012*. University of Miami.

Darling, K. 2015. ’who’s johnny?’anthropomorphic framing in human-robot interaction, integration, and policy. *Browser Download This Paper*.

Dennett, D. C. 1981. *Brainstorms: Philosophical essays on mind and psychology*. MIT press.

Dennett, D. C. 1991. Real patterns. *The journal of Philosophy* 88(1):27–51.

Dretske, F. I. 1997. *Naturalizing the mind*. mit Press.

Dreyfus, H. L. 1992. *What computers still can’t do: a critique of artificial reason*. MIT press.

Estrada, D. J. 2014. *Rethinking machines: Artificial intelligence beyond the philosophy of mind*. University of Illinois at Urbana-Champaign.

Ferrara, E.; Varol, O.; Davis, C.; Menczer, F.; and Flammini, A. 2016. The rise of social bots. *Communications of the ACM* 59(7):96–104.

Floridi, L. 2016. Should we be afraid of ai? *Aeon Essays*.

Gödel, K. 1931. Über formal unentscheidbare sätze der principia mathematica und verwandter systeme i. *Monatshefte für mathematik und physik* 38(1):173–198.

Gunkel, D. J. 2012. *The machine question: critical perspectives on AI, robots, and ethics*. MIT Press.

Harnad, S. 1989. Minds, machines and searle. *Journal of Experimental & Theoretical Artificial Intelligence* 1(1):5–25.

Hodges, A. 2012. *Alan Turing: the enigma*. Random House.

Johnson Jr, H. N. 2016. From status to agency: Abolishing the very spirit of slavery. *Colum. J. Race & L.* 7:245.

Leavitt, D. 2017. Turing and the paranormal. In Copeland, J.; Bowen, J.; Sprevak, M.; and Wilson, R., eds., *The Turing Guide*. Oxford University Press. chapter 27.

Mohri, M.; Rostamizadeh, A.; and Talwalkar, A. 2012. *Foundations of machine learning*. MIT press.

Penrose, R. 1999. *The emperor’s new mind: Concerning computers, minds, and the laws of physics*. Oxford Paperbacks.

Proudfoot, D. 2017. The turing test from every angle. In Copeland, J.; Bowen, J.; Sprevak, M.; and Wilson, R., eds., *The Turing Guide*. Oxford University Press. chapter 27.

Rodriguez, J. F. 2017. Supervisor seeks robot delivery ban in sf.

Russell, S.; Dewey, D.; and Tegmark, M. 2015. Research priorities for robust and beneficial artificial intelligence. *Ai Magazine* 36(4):105–114.

Salvini, P.; Ciaravella, G.; Yu, W.; Ferri, G.; Manzi, A.; Mazzolai, B.; Laschi, C.; Oh, S.-R.; and Dario, P. 2010. How safe are service robots in urban environments? bullying a robot. In *RO-MAN, 2010 IEEE*, 1–7. IEEE.

Salvini, P.; Laschi, C.; and Dario, P. 2010. Design for acceptability: improving robots coexistence in human society. *International journal of social robotics* 2(4):451–460.

Saygin, A. P.; Cicekli, I.; and Akman, V. 2000. Turing test: 50 years later. *Minds and machines* 10(4):463–518.

Searle, J. R. 1980. Minds, brains, and programs. *Behavioral and brain sciences* 3(3):417–424.

Smullyan, R. 2001. Gödels incompleteness theorems. *The Blackwell Guide to Philosophical Logic. Blackwell* 72–89.

Soares, N., and Fallenstein, B. 2014. Aligning superintelligence with human interests: A technical research agenda. *Machine Intelligence Research Institute (MIRI) technical report* 8.

Turing, A. M. 1937. On computable numbers, with an application to the entscheidungsproblem. *Proceedings of the London mathematical society* 2(1):230–265.

Turing, A. 1947. The automatic computing engine. *Lecture to the London Mathematical Society*.

Turing, A. M. 1948. Intelligent machinery, a heretical theory. *The Turing Test: Verbal Behavior as the Hallmark of Intelligence* 105.

Turing, A. M. 1950. Computing machinery and intelligence. *Mind* 59(236):433–460.

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

Yampolskiy, R. V. 2013. Artificial intelligence safety engineering: Why machine ethics is a wrong approach. *Philosophy and theory of artificial intelligence* 389–396.

Yudkowsky, E. 2008. Artificial intelligence as a positive and negative factor in global risk. *Global catastrophic risks* 1(303):184.