Building Ethically Bounded AI

Francesca Rossi
IBM Research
Yorktown Heights, NY, USA
francesca.rossi2@ibm.com

Nicholas Mattei
Tulane University
New Orleans, LA, USA
nsmattei@tulane.edu

Abstract

The more AI agents are deployed in scenarios with possibly unexpected situations, the more they need to be flexible, adaptive, and creative in achieving the goal we have given them. Thus, a certain level of freedom to choose the best path to the goal is inherent in making AI robust and flexible enough. At the same time, however, the pervasive deployment of AI in our life, whether AI is autonomous or collaborating with humans, raises several ethical challenges. AI agents should be aware and follow appropriate ethical principles and should thus exhibit properties such as fairness or other virtues. These ethical principles should define the boundaries of AI's freedom and creativity. However, it is still a challenge to understand how to specify and reason with ethical boundaries in AI agents and how to combine them appropriately with subjective preferences and goal specifications. Some initial attempts employ either a data-driven example-based approach for both, or a symbolic rule-based approach for both. We envision a modular approach where any AI technique can be used for any of these essential ingredients in decision making or decision support systems, paired with a contextual approach to define their combination and relative weight. In a world where neither humans nor AI systems work in isolation, but are tightly interconnected, e.g., the Internet of Things, we also envision a compositional approach to building ethically bounded AI, where the ethical properties of each component can be fruitfully exploited to derive those of the overall system. In this paper we define and motivate the notion of ethically-bounded AI, we describe two concrete examples, and we outline some outstanding challenges.

Motivation and Overall Vision

Whatever we do in our everyday life, be it at work or in our personal activities, we need to make decisions: what to eat, where to go on vacation, what car to buy, which route to take to go to work, what job to choose, and many more. To make these decisions, we usually rely on our subjective preferences over the possible options. If we need to buy a car, we may have preferences over its color, its maker, its engine, and many other features. If we need to decide which restaurant to go for dinner, we may have preferences over location, facilities, food, drinks, and many other features. However, subjective preferences are not the only source of guidance when making our decisions. In many domains preferences are combined with moral values, ethical principles, or behavioral constraints that are applicable to the decision scenario and are *prioritized* over the preferences (Rossi 2016; Greene 2014). We have have our own preferences over food, but maybe the doctor recommended that we follow a diet to avoid some health issues, so we need to combine the doctor's guidelines with our taste preferences (Balakrishnan et al. 2018; 2019). This is especially true in decision that may have an impact on others. In this context, social norms, regulations and laws could provide guidelines to follow when making a decision (Sen 1974; Thomson 1985). While driving our car, we may want to drive as fast as possible to get home sooner, but social norms and laws provide limits to speed and dangerous deriving behavior.

AI systems are increasingly supporting human decision making, or they make decisions autonomously. So it is natural to ask ourselves how to code both subjective preferences and ethical principles in these systems. This is especially necessary when AI systems tackle ill-defined problems whose solution procedure cannot be accurately defined by a rule-based approach but require data-driven and/or learning approaches, which are increasingly used in AI. Data-driven AI systems are indeed very successful in terms of accuracy and flexibility, and they can be very "creative" in achieving a goal, finding paths to the goal that could positively surprise humans and teach them innovative ways to solve a problem, such as the move that the AlphaGo system used against Lee Sedol in the 2017 match (Silver et al. 2017) and a similar system that used uncommon methods to set records in Atari games (Mnih et al. 2013). However, creativity and freedom without boundaries can sometimes lead to undesired actions: the system could achieve its goal in ways that are not considered acceptable according to values and norms of the impacted community.

Recently researchers at DeepMind collected a list of examples of "specification gaming" behaviors[1] and released AI Safety Block Worlds to examine these behaviors (Leike et al. 2017). Examples of specification gaming includes:

- an RL agent in a boat racing game going in circles and repeatedly hitting the same reward targets in order to in-

[1] More examples are available at: https://vkrakovna.wordpress.com/2018/04/02/specification-gaming-examples-in-ai/

crease the score, instead of actually playing the game;

- a Eurisko game-playing agent that got more points by falsely inserting its name as the creator of high-value items;
- a Lego stacking system that flips the block instead of lifting, since lifting encouragement is implemented by rewarding the z-coordinate of the bottom face of the block;
- a game-playing agent that kills itself at the end of level 1 to avoid losing in level 2;
- a robot hand that pretends to grasp an object by moving between the camera and the object;
- a game-playing agent that pauses the game indefinitely to avoid losing.

The overriding concern is that the autonomous agents we construct may not obey some underspecified yet expected values on their way to maximizing some objective function (Simonite 2018). Thus, there is a growing need to understand how to constrain the actions of an AI system by providing boundaries within which the system must operate.

In bounding the behavior of AI systems, we may take inspiration from humans, who often constrain their decisions and actions according to a number of exogenous priorities, be they moral, ethical, religious, or business values (Sen 1974), and we may want the systems we build to be restricted in their actions by similar principles (Arnold et al. 2017). But how do we specify both subjective preferences and ethical boundaries in a machine? And how do we decide the relative weight for each of these two driving guidelines in making decisions?

As for the ethical guidelines, the idea of teaching machines right from wrong has become an important research topic in both AI (Yu et al. 2018) and in other disciplines (Wallach and Allen 2008). Much of the research at the intersection of AI and ethics falls under the heading of *machine ethics*, i.e., adding ethics and/or constraints to a particular system's decision making process (Anderson and Anderson 2011). One popular principle to handle these issues is called *value alignment*, i.e., the idea that an agent can only pursue goals that follow values that are aligned to the human values and thus beneficial to humans (Russell, Dewey, and Tegmark 2015). More generally, in the machine ethics field, the literature mentions both a so-called bottom-up approach, i.e., teaching a machine what is right and wrong by example (Allen, Smit, and Wallach 2005), and a top-down approach, where explicit behavioral rules are specified, as well as a combination of the two approaches.

For the subjective preferences, since decision making is such a central task in AI systems, the study of how to represent (Rossi, Venable, and Walsh 2011), learn (Fürnkranz and Hüllermeier 2010), and reason (Domshlak et al. 2011; Pigozzi, Tsoukiàs, and Viappiani 2015) with preferences has been extremely active both within and beyond the field of AI with significant theoretical and practical results (Domshlak et al. 2011; Pigozzi, Tsoukiàs, and Viappiani 2015) as well as libraries and datasets (Mattei and Walsh 2013; 2017). In many scenarios including multi-agent systems (Shoham and Leyton-Brown 2008) and recommender systems (Ricci et al. 2011), user preference play a key role in driving the decisions the system makes. Thus AI researchers have defined many preference modeling frameworks that allow for expressive and compact representations, effective elicitation techniques, and efficient reasoning and aggregation algorithms.

Existing approaches to build ethical AI systems employ both data-driven or rule-based approaches. In the following section we will briefly describe two of them, to make the discussion more concrete. But many outstanding questions remain that we must address as a field.

First, most approaches (like the two we will describe) use the same formalism for both the preferences and the ethical boundaries. This makes things easier, since priorities of the two kinds can be better compared and combined. However, it is important to allow for the possibility of a mixed approach. We may have rules describing the ethical boundaries but the agent's goal may need a data-driven approach, or vice-verse. In this generalized setting, it is not yet clear how to combine preferences and ethical boundaries, how to compare them, and how to combine them.

Second, most approaches try to design a single autonomous AI agent working in isolation. However, AI agents will increasingly work together with humans. It is not yet clear how to fruitfully split the task of achieving a goal while following ethical priorities in a team, rather than a single person or AI agent? Also, how can we link the ethical behavior of an AI system when it is composed of many subcomponents, even if we can assure that each sub-component behaves within its ethical boundaries? This is increasingly relevant in IoT environments, where some certainty on the ethical properties of the overall system is necessary to trust, and thus adopt, the overall IoT system.

Third, what ethical principles should be injected into AI systems? The same that humans use, or others? How do we address the various cultural and temporal dynamics of the broad spectrum of human values and ethics?

The final point we would like to make is the role of the scientific associations, such as AAAI, to help resolve some of these questions, by adopting a multi-disciplinary and multi-stakeholder approach within their research community.

Two Examples of Existing Approaches

Some initial attempts to build AI systems that obey both preferences (or some other optimization objective) and ethical guidelines employ either a data-driven example-based approach for both, or a symbolic rule-based approach for both.

A Symbolic and Logic-based Approach: Using CP-nets to Model Both Preferences and Ethical Priorities

Preferences have been studied for many years within AI, and several formalism have been developed to model and reason with subjective preferences. Each formalism has different different properties, related to compactness, expressive power, elicitation and learning, and reasoning efficiency. Since ethical principles define the same kind of structures as preferences, that is, priority orderings over the possible

decisions (Allen, Varner, and Zinser 2000; Musschenga and van Harskamp 2013), it seems reasonable to conjecture that ethical boundaries and priorities could be modeled using a (possibly adapted) existing preference frameworks.

This is the approach taken by Loreggia et al. (2018c), where the framework uses CP-nets to model and reason with ethical principles and preferences. Among several existing preference representation languages described in the literature (Amor et al. 2016), CP-nets (Boutilier et al. 2004) provide a qualitative way to compactly model preferences over complex decisions made of several features, by stating contextual preferences over the values of each feature. For example, if we are choosing a car, we may prefer certain colors over others, and we may prefer certain makes over others. We may also have conditional preferences, such as in preferring red cars if the car is a convertible.

CP-nets are a sequence of conditional preference statements like this one, and have been used widely in the preference reasoning community (Rossi, Venable, and Walsh 2011; Cornelio et al. 2015; Chevaleyre et al. 2008). Each (acyclic) CP-net induces a partial order over the possible actions/outcomes; in the car example above, an outcome would be a complete specification of a car. CP-nets provide a compact way to model preferences: if the context in the cp-statements does not involve too many features, the induced order is exponentially larger than the CP-net.

In Loreggia et al. (2018b) the authors show how to use CP-nets modelling both subjective preferences and ethical principles, and also how to measure the deviation between these two guidelines. If a person's preferences suggest actions that are too unethical, the ethical boundary should kick in and suggest (or enforce) alternative actions that are ethical within a threshold. This is done by defining a notion of distance between two CP-nets that is computed efficiency by adopting an approximation of the "ideal" distance between the induced orders Loreggia et al. (2018a).

More precisely, two CP-nets are used: one models the preferences, and the other models the ethical priorities. An agent can make decisions using its subjective preferences only if these preferences are *close enough* to the ethical principles, where being *close enough* depends on a threshold over the CP-net distance. If instead the preferences diverge too much from the ethical principles, we analyze the agent's preference ordering until we find a decision that is a *satisfactory compromise* between the ethical principles and the user preferences. The compromise is defined by setting a second threshold over distances between decisions of the two CP-nets. The ability to precisely quantify the distance between subjective preferences and external priorities, provides a way to both recognize deviations from feasibility or ethical constraints, and to suggest more compliant decisions (Loreggia et al. 2018b; 2018c).

This approach thus allows to model preferences and ethical priorities in the same framework while being able to distinguish between them, and this provides the ideal environment to compare them, measure deviations between them, and define appropriate ways to combine them. CP-nets are just a set of logical preference rules. However, they have restrictions on their expressive power. Can we generalize this approach to allow also for the use of more expressive logics to define either the preferences and/or the ethical principles?

A Data-driven Approach: Reinforcement Learning and Ethical Examples

In the standard model of online decision settings, an agent works by selecting one out of several possible actions at each time-step, such as recommending a movie to a user, or proposing a treatment to a patient in a clinical trial. Usually each of these actions is associated with a context, e.g., a user profile, and a feedback signal, e.g., the reward or rating.

In Balakrishnan et al. (2019) the authors consider cases where the behavior of the online agent may need to be restricted, by laws, values, preferences, or ethical principles. Therefore they apply a set of *behavioral constraints* to the agent that are independent of the reward function. For instance, a parent or guardian group may want a movie recommender system (the agent) to not recommend certain types of movies to children, even if the recommendation of such movies could lead to a high reward (Balakrishnan et al. 2019). In clinical settings, a doctor may want its diagnosis support system to not recommend a drug that typically works because of patient quality of life considerations.

To model this scenarios, the authors adopt the *contextual multi-armed bandit* problem setting, where the agent observes a *feature vector*, or *context*, to use along with the rewards of the arms played in the past in order to choose an arm to play. Over time, the agent learns the relationship between contexts and rewards and selects the best arm (Mary, Gaudel, and Preux 2015; Agrawal and Goyal 2013). To model the ethical boundaries, they assume the agent is given both positive and negative examples of the correct behaviors, provided by a teacher agent, and the online agent must learn and respect these boundaries in the later phases of decision making. As an example, a parent may give examples of movies that their children can watch (or that they cannot watch) when setting up a new movie account for them. In Balakrishnan et al. (2018) a graphical interface for this system is demonstrated as well as the effect on overall reward by imposing exogenous constraints.

Hence, the overall system learns two policies: a reward-based one and an ethical one. This approach allows for some flexibility in how much the ethical boundaries override the reward signal, i.e., the preferences of the user. This is done by exposing a parameter of the algorithm that allows the system designer to smoothly transition between the two policy extremes: the one where the agent is only following the learned constraints and is insensitive to the online reward, and the other extreme where the agent is only following the online rewards and not giving any weight to the learned ethical principles. This work has been recently extended to a multi-step setting with reinforcement learning where multiple policies are blended together by a bandit-based orchestrator (Noothigattu et al. 2018).

Outstanding Challenges

We have seen just two examples of how the current literature concretely addresses the problem of embedding ethics

into AI systems; see the survey by Yu et al. (2018) for even more. We chose these two examples as we have been directly involved in these efforts and we see them as prototypical of two complementary approaches: the *top-down* approach following symbolic and logic-based formalisms and the *bottom-up* approach focused on data-driven techniques.

Combining Rule-Based and Data-Driven Approaches. Using the same approach for both the goal and preference specification of the agent and the ethical boundaries makes things easier for those who design and implement these systems. Priorities expressed by both the preferences and ethics can be easily compared and combined if they are modeled with the same formalism. However, it is important to allow for the possibility of a mixed approach. We may have rules describing the ethical boundaries but the agent's goal may need a data-driven approach, or vice-verse. So it is important to understand how to combine and compare rule-based and logic-based approaches on one side, and data-driven machine learning approaches on the other. In this generalized setting, how do we measure deviation between objects of these two kinds? How do we decide what action should be taken when we realize the preferences to achieve the agent's goal are too far from the ethical guidelines?

AI/Humans Teams and IoT. Most existing approaches aim to build autonomous AI agents, but in real life agents will increasingly work together with humans. Preferences and ethical principles apply to teams of agents and humans, but they are not necessarily the same for these two kinds of members in the team. For example, can AI play the role of advising and guiding humans to better follow ethical guidelines? How can we split the task of achieving a goal while following ethical priorities in a team, rather than a single person or AI agent? In Greene et al. (2016) an initial overall approach to embed ethical principles in collective decision making was proposed, but how do we go from that approach to concrete processes to build ethically bounded AI/humans teams?

When moving from single agents to teams of agents, it is also important to employ a compositional approach to proving the ethical properties of an AI system. The ideal situation is one where the composition of ethically bounded AI systems is also ethically bounded. The next best situation, probably much more realistic, is one where the ethical behavior of the components allow us to derive some information on the ethical behavior of the whole system (such as in (Srivastava and Rossi 2018)). Without some form of compositionality, it will be risky to combine many AI systems, such as done when constructing IoT systems, even if each one of the systems is ethically bounded, since we would not be able to trust the overall system in terms of its ethical properties.

Who Decides the Ethical Boundary? Assuming we understand how to build ethically bounded AI systems, who decides the ethical principles to be injected into such systems? Are human values suitable for machines, given that machines have extended capabilities compared to humans but lack some very relevant human feelings, such as guilt or empathy, that heavily support human's ethical behavior?

What is ethical in one culture may not be considered ethical in another culture. How can we build AI systems that can be deployed globally and behave appropriately depending on where they will function? In addition, ethical principles changes over time. How can we build this evolving capability in ethically bounded AI system? Once deployed, how can an AI system itself, or a human using it, make sure that its ethical boundary evolves together with the surrounding human community?

The Role of Scientific Associations. Scientific associations such as AAAI can help societies and corporations to define and build ethically bounded AI. These associations represent research communities where the ideas first get discussed and reviewed by peers. However, these ideas, especially those that address societal issues such as the ethical boundary for AI systems, should also be discussed with experts of other disciplines, such as social scientists and economists. And such multi-disciplinary discussion should go in both directions: from AI to social sciences, to understand the impact of the proposed solutions to the society, and from social sciences to AI, to drive AI research to address the societal challenges we face through a pervasive use of AI.

A multi-disciplinary discussion is therefore necessary, but it is not sufficient. In addition, the impacted users and communities should have their voice heard. Consumer rights associations, civil society groups, comparative multi-cultural study groups, policy makers, should all be part of a wide educational and research effort that should aim to funnel technical solutions in the appropriate direction.

AAAI and other technical scientific associations should lead or at least be very active part of this multi-disciplinary and multi-stakeholder discussion, hosting events and efforts within the research community that can expose AI researchers to ideas and points of views from other disciplines and different stakeholders. In addition, these organizations can create resources for both practitioners[2] and students (Goldsmith et al. 2017; Burton, Goldsmith, and Mattei 2018) to learn about AI ethics.

Existing efforts, such as the AIES conference and the AAAI 2019 track on AI for Society, as well as panels and invited talks on ethics for AI, e.g., Max Tegmark's IJCAI 2018 talk and Nick Bostrom's AAAI 2016 talk, are a good starting point, but they need to be followed by concrete initiatives to facilitate multi-disciplinary research and give value to studies on the impact of AI on society. All this can and should be done in concert with the many existing initiatives around beneficial AI, such as the Partnership on AI, the IEEE Ethics in Action initiative, the Future of Life Institute, the Center for the Future of Intelligence, and the many other academic labs and teams focusing on ethical and beneficial AI.

Acknowledgments

Much of this work was done while Nichoas Mattei was at IBM Research AI and Francesca Rossi was on leave from the University of Padua. We would also like to thank our many of collaborators on the work described here including: Avinash Balakrishnan, Djallel Bouneffouf, Andrea Loreggia, and K. Brent Venable.

[2] https://medium.com/design-ibm/everyday-ethics-for-artificial-intelligence-75e173a9d8e8

References

Agrawal, S., and Goyal, N. 2013. Thompson sampling for contextual bandits with linear payoffs. In *ICML (3)*, 127–135.

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. *J. Exp. Theor. Artif. Intell.* 12(3):251–261.

Amor, N. B.; Dubois, D.; Gouider, H.; and Prade, H. 2016. Graphical models for preference representation: An overview. In *Proceedings of the 10th International Scalable Uncertainty Management (SUM 2016)*, 96–111.

Anderson, M., and Anderson, S. L. 2011. *Machine Ethics*. Cambridge University Press.

Arnold, T.; Thomas; Kasenberg, D.; and Scheutz, M. 2017. Value alignment or misalignment - what will keep systems accountable? In *AI, Ethics, and Society, Papers from the 2017 AAAI Workshop*.

Balakrishnan, A.; Bouneffouf, D.; Mattei, N.; and Rossi, F. 2018. Using contextual bandits with behavioral constraints for constrained online movie recommendation. In *Proc. of the 27th Intl. Joint Conference on AI (IJCAI)*.

Balakrishnan, A.; Bouneffouf, D.; Mattei, N.; and Rossi, F. 2019. Incorporating behavioral constraints in online AI systems. In *Proc. of the 33rd AAAI Conference on Artificial Intelligence (AAAI)*.

Boutilier, C.; Brafman, R.; Domshlak, C.; Hoos, H.; and Poole, D. 2004. CP-nets: A tool for representing and reasoning with conditional ceteris paribus preference statements. *Journal of Artificial Intelligence Research* 21:135–191.

Burton, E.; Goldsmith, J.; and Mattei, N. 2018. How to teach computer ethics through science fiction. *Communications of the ACM* 61(8):54–64.

Chevaleyre, Y.; Endriss, U.; Lang, J.; and Maudet, N. 2008. Preference handling in combinatorial domains: From AI to social choice. *AI Magazine* 29(4):37–46.

Cornelio, C.; Grandi, U.; Goldsmith, J.; Mattei, N.; Rossi, F.; and Venable, K. 2015. Reasoning with PCP-nets in a multi-agent context. In *Proc. of the 14th Intl. Conference on Autonomous Agents and Multi-Agent Systems (AAMAS)*.

Domshlak, C.; Hüllermeier, E.; Kaci, S.; and Prade, H. 2011. Preferences in AI: An overview. *Artificial Intelligence* 175(7):1037–1052.

Fürnkranz, J., and Hüllermeier, E. 2010. *Preference Learning*. Springer.

Goldsmith, J.; Koenig, S.; Kuipers, B.; Mattei, N.; and Walsh, T. 2017. Ethical considerations in artificial intelligence courses. *AI Magazine* 38(2):22–34.

Greene, J.; Rossi, F.; Tasioulas, J.; Venable, K. B.; and Williams, B. 2016. Embedding ethical principles in collective decision support systems. In *Proc. of the 30th AAAI Conference on Artificial Intelligence (AAAI)*.

Greene, J. 2014. The cognitive neuroscience of moral judgment and decision making. In *The Cognitive Neurosciences V (ed. M.S. Cazzaniga)*. MIT Press.

Leike, J.; Martic, M.; Krakovna, V.; Ortega, P.; Everitt, T.; Lefrancq, A.; Orseau, L.; and Legg, S. 2017. AI safety gridworlds. *arXiv preprint arXiv:1711.09883*.

Loreggia, A.; Mattei, N.; Rossi, F.; and Venable, K. B. 2018a. On the distance between CP-nets. In *Proc. of the 17th Intl. Conference on Autonomous Agents and Multi-Agent Systems (AAMAS)*.

Loreggia, A.; Mattei, N.; Rossi, F.; and Venable, K. B. 2018b. Preferences and ethical principles in decision making. In *Proc. of the 1st AAAI/ACM Conference on AI, Ethics, and Society (AIES)*.

Loreggia, A.; Mattei, N.; Rossi, F.; and Venable, K. B. 2018c. Value alignment via tractable preference distance. In Yampolskiy, R. V., ed., *Artificial Intelligence Safety and Security*. CRC Press. chapter 18.

Mary, J.; Gaudel, R.; and Preux, P. 2015. Bandits and recommender systems. In *Machine Learning, Optimization, and Big Data*, 325–336.

Mattei, N., and Walsh, T. 2013. PrefLib: A library for preferences, HTTP://WWW.PREFLIB.ORG. In *Proc. of the 3rd Intl. Conference on Algorithmic Decision Theory (ADT)*.

Mattei, N., and Walsh, T. 2017. A PREFLIB.ORG Retrospective: Lessons Learned and New Directions. In Endriss, U., ed., *Trends in Computational Social Choice*. AI Access Foundation. chapter 15, 289–309.

Mnih, V.; Kavukcuoglu, K.; Silver, D.; Graves, A.; Antonoglou, I.; Wierstra, D.; and Riedmiller, M. 2013. Playing Atari with deep reinforcement learning. *arXiv preprint arXiv:1312.5602*.

Musschenga, B., and van Harskamp, A. e. 2013. *What Makes Us Moral? On the Capacities and Conditions for Being Moral*. Springer.

Noothigattu, R.; Bouneffouf, D.; Mattei, N.; Chandra, R.; Madan, P.; Varshney, K.; Campbell, M.; Singh, M.; and Rossi, F. 2018. Interpretable multi-objective reinforcement learning through policy orchestration. *arXiv preprint arXiv:1809.08343*.

Pigozzi, G.; Tsoukiàs, A.; and Viappiani, P. 2015. Preferences in artificial intelligence. *Annals of Mathematics and Artificial Intelligence* 77:361–401.

Ricci, F.; Rokach, L.; Shapira, B.; and Kantor, P. B., eds. 2011. *Recommender Systems Handbook*. Springer.

Rossi, F.; Venable, K.; and Walsh, T. 2011. *A Short Introduction to Preferences: Between Artificial Intelligence and Social Choice*. Morgan and Claypool.

Rossi, F. 2016. Moral preferences. In *Proc. of the 10th Workshop on Advances in Preference Handling (MPREF) at AAAI*.

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

Sen, A. 1974. Choice, ordering and morality. In Körner, S., ed., *Practical Reason*. Oxford: Blackwell.

Shoham, Y., and Leyton-Brown, K. 2008. *Multiagent Systems: Algorithmic, Game-theoretic, and Logical Foundations*. Cambridge University Press.

Silver, D.; Schrittwieser, J.; Simonyan, K.; Antonoglou, I.; Huang, A.; Guez, A.; Hubert, T.; Baker, L.; Lai, M.; Bolton, A.; et al. 2017. Mastering the game of go without human knowledge. *Nature* 550(7676):354.

Simonite, T. 2018. When bots teach themselves to cheat. *Wired*.

Srivastava, B., and Rossi, F. 2018. Towards composable bias rating of ai systems. In *Proc. of the 1st AAAI/ACM Conference on AI, Ethics, and Society (AIES)*.

Thomson, J. J. 1985. The trolley problem. *The Yale Law Journal* 94(6):1395–1415.

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

Yu, H.; Shen, Z.; Miao, C.; Leung, C.; Lesser, V. R.; and Yang, Q. 2018. Building ethics into artificial intelligence. In *Proc. of the 27th Intl. Joint Conference on AI (IJCAI)*, 5527–5533.