Sociotechnical Systems and Ethics in the Large

Amit K. Chopra
Lancaster University
Lancaster LA1 4WA, UK
a.chopra1@lancaster.ac.uk

Munindar P. Singh
North Carolina State University
Raleigh, NC 27695-8206, USA
singh@ncsu.edu

Abstract

Advances in AI techniques and computing platforms have triggered a lively and expanding discourse on ethical decision-making by autonomous agents. Much recent work in AI concentrates on the challenges of moral decision making from a decision-theoretic perspective, and especially the representation of various ethical dilemmas. Such approaches may be useful but in general are not productive because moral decision making is as context-driven as other forms of decision making, if not more. In contrast, we consider ethics not from the standpoint of an individual agent but of the wider sociotechnical systems (STS) in which the agent operates.

Our contribution in this paper is the conception of ethical STS founded on *governance* that takes into account stakeholder values, normative constraints on agents, and outcomes (states of the STS) that obtain due to actions taken by agents. An important element of our conception is *accountability*, which is necessary for adequate consideration of outcomes that *prima facie* appear ethical or unethical. Focusing on STSs avoids the difficult problems of ethics as the norms of the STS give an operational basis for agent decision making.

1 Introduction

Advances in computing platforms and artificial intelligence techniques have led to the situation where machines perform more and more of the tasks that humans have traditionally performed (Russell, Dewey, and Tegmark 2015). We refer to such a machine, potentially consisting of both hardware and software, as an *autonomous agent* or *agent* for short. We reserve the word *system* for social entities of two or more agents. In addition to constrained tasks such as flight control, which achieved capabilities of near autonomy decades ago, we see harbingers of autonomous capabilities in virtually any domain of human endeavor. Applications range from transportation and warehouse automation to healthcare and education.

A question that has fostered significant interdisciplinary discussion is whether we can ensure that agents make ethical decisions (Bostrom and Yudkowsky 2014; Conitzer et al. 2017). Partly this interest stems from science fiction media that depicts machines that are sentient and self-interested and may therefore come into conflict with humans. This question has inspired various versions of "do no harm to humans" maxims, from Asimov to Bostrom and Yudkowsky (2014). And, partly this interest stems from imagining that agents are deliberative entities who will make choices much in the same way humans do: faced with a situation that demands deliberation, an agent will line up its choices and make the best one that is also the most ethical. The *trolley problem*, a moral dilemma that has been the subject of extensive philosophical discussion, has been discussed extensively in the context of self-driving vehicles (Bonnefon, Shariff, and Rahwan 2016).

Concurrently, there has been an expanding body of work in the broad AI tradition that investigates designing and verifying, not individual agents, but *sociotechnical systems* (STSs), e.g., (Pitt, Schaumeier, and Artikis 2012; Singh 2013). STSs comprise social entities (principals, such as people and organizations, who may be represented computationally by agents) and technical entities (mechanisms, such as those comprising the infrastructure). In our conception, STS are computational systems that capture the normative context for the actions of the participating agents (Chopra and Singh 2016). There is a huge difference between (a) using the context merely to inform an agent's design and decision making; and (b) formalizing the context itself as a computational system (with attendant properties such as specification, composition, execution, and state) and using it to *potentially* inform an agent's design and decision making.

Our formulation of STSs enables us to pose a novel question: Is a given STS ethical? This question is crucial to ethics because whether an agent's actions are ethical depends upon whether the relevant STS is ethical. For example, in Victor Hugo's Les Misérables, the protagonist, Jean Valjean, steals a loaf of bread to feed his sister's family. Let us accept that stealing is unethical as is letting your sister starve when you can prevent her starvation. Valjean's actions are unethical but his alternative was unethical as well: he is truly in a moral quandary (Richardson 2014). Our approach would not pass judgment on Valjean's actions but instead find fault with the STS (the French society of the book) in which he functions. The STS is at fault for placing its participant in a quandary.

In general, an autonomous agent may be a participant in several STSs, so to be more precise, we ask if the autonomous agent's actions are ethical from the point of view of a given STS. But what does it mean for an STS to be ethical? This paper's main contribution is in providing an answer to this question.

2 Ethics in the Small

Ethics in the small concerns the decision making of individual agents.

2.1 Atomistic: Single-Agent View

Ethics intrinsically has a normative basis in the sense that the purpose is to distinguish an agent's correct actions from incorrect ones. Although ideas such as moral values and social norms, as standards of correctness, necessarily arise in any discussion of ethics and AI, their scope has largely been limited to the *design* and *impact* of autonomous agents.

The design concern is about *informing* the design of autonomous agents with ideas from ethics. Is it sensible at all to talk about agents being ethical? Assuming it is, when can we say an agent is ethical? Are there categories of ethical agents (Moor 2006)? And, how can we ensure that the agents we design are ethical? Assuming we had a suitable specification of ethics, then agent designs could be verified for ethics. For instance, Dennis et al. (2016) provide a language to express ethical requirements and an ordering over those requirements. They give techniques for verifying where an agent always selects the most ethical plan. In fact, there is a tradition in multiagent systems research to use deontic logic toward the verification of agent designs (Meyer and Wieringa 1993).

The impact concern broadly is about the relationship between agents and society, though the literature traditionally approaches them from the standpoint of a single agent. An important element of impact is *accountability* (Diakopoulos 2016), which we address at length later. Specifically, who would be accountable for the decisions made by an autonomous agent? For example, it would appear unsatisfactory for a manufacturer of an autonomous vehicle to design and sell it but not be accountable for mishaps because the vehicle was autonomous. Autonomous agents based on machine learning further complicate the picture because it is difficult to tell in advance what specific facts an agent will learn and therefore how it will act (Danks and London 2017). Specifically, we may encounter situations where the autonomous agent behaves correctly as designed but produces outcomes that are ethically suspect because the facts it has been trained on are biased or wrong.

Both the design and impact concerns are worthwhile, though they are inherently too narrowly conceived from the single-agent perspective.

2.2 Decision-Making, Principles, and Dilemmas

An agent's decision making involves ingredients such as these: beliefs about the world, goals, capabilities, and normative relationships with other agents. An *ethically conscious* agent's decision making would maximize achievement of its goals while minimizing violations of its ethics. A staple of the discussion in literature is how would an agent act when placed in an ethical dilemma. What makes an ethical dilemma interesting is that all the choices it exposes are unethical. In the trolley problem, the choice is between letting five people die or saving them but in the process killing one. The practical problem is to determine the best of the bad lot of choices. The hope is that solving the question will shed light on the nature and content of valid ethical principles.

Centuries of moral philosophy though have yielded no universal answer about ethical judgments. What philosophers have learned is that we find a number of ethical principles useful. For example, Sen (2011) discusses *utilitarianism*, *egalitarianism*, *libertarianism*, and so on, and shows that these principles are not mutually compatible. The list of ethical principles goes on: Kant's Categorical Imperative, Golden Rule, the Doctrine of Double Effect, and so on. As a way of overcoming the challenges of determining a valid ethics for autonomous agents, Bonnefon et al. (2016) suggest that their ethics could be informed by empirical studies of people's conception of ethics (Bonnefon, Shariff, and Rahwan 2016).

We claim that though ethical dilemmas may be interesting, they do not yield productive research questions. In essence, studies that seek people's opinions about the dilemmas are little more than parlor games. And, in addition to the problem of validity, *bounded rationality* means that most broad ethical conceptions and maxims are not operationalizable. How could one possibly compute all the consequences of one's action? Motivated by the problem of developing an agent that is capable of general intelligence (in contrast to being intelligent in a particular domain, say playing chess or driving), Bostrom and Yudkowsky propose "any action that causes no harm to humans is permitted" as the ethical principle by which the agent should act. It is not clear what would constitute *harm* in the first place, let alone one computing whether harm was caused.

3 STSs from the Standpoint of Ethics

The idea of a sociotechnical system builds upon work on institutions and organizations of autonomous agents (Artikis, Sergot, and Pitt 2009; Modgil et al. 2015; King et al. 2015). A key idea in this body of work is the distinction between *regulation* and *regimentation* (Jones and Sergot 1993). Regulation is the idea that the norms relevant to the setting are specified and satisfaction is left up to the agents. Regulation is thus autonomy-promoting but at the same time yields explicit standards of correctness. Regimentation is the idea that interactions between agents can be *implemented* in machines; its standards of correctness are implicit. Our motivation for emphasizing the *sociotechnical* view is that a real-life system must embody both a social architecture (regulatory) and a technical architecture (regimented) (Singh 2015). How to identify and work with the tradeoffs between them is a crucial challenge (Kafalı, Ajmeri, and Singh 2016; 2017).

We conceive of STSs in terms of three major kinds of elements: (1) stakeholders; (2) information, namely, stakeholder values, prescriptive norms, and outcomes, and (3) processes for governance, including for purposes of respecifying an STS.

3.1 Values, Norms, and Outcomes

Let's consider each of these in a hypothetical sociotechnical system for transportation by driverless vehicles. Here the stakeholders may include automobile owners, passengers, the public, road authorities, insurance providers, manufacturers, garages, and regulatory agencies. Suppose the values the stakeholders want to promote are comfort and well-being of owners and passengers, their privacy, economic sustainability, well-being of property, and speed of transportation. Upon deliberation, the stakeholders come up with the (prescriptive) norms for their STS (values of course don't *determine* norms).

Norms are a general and powerful construct. For concreteness, we adopt the conception of regulatory norms advocated by Von Wright, the father of deontic logic (1963; 1999); we adopt an enhancement of our previous formulation (Singh 2013). We give a few examples of the norms in relation to selected values—the small caps refer to roles abstracted from stakeholders.

- *Promoting well-being.* OWNER *authorizes* MANUFACTURER to apply software updates for enhanced safety and performance. MANUFACTURER *commits* to OWNER to apply all mandated updates. A REGULATORY AGENCY has been *empowered* over OWNER to mandate a software update.
- *Promoting speed of transportation and safety.* ROAD AUTHORITIES *commit* to OWNER to provide up to date information about directions and road conditions. Further, they *commit* to all other stakeholders to maintain roads and information systems up to the required standards.
- *Promoting safety of humans and property.* OWNER is *prohibited* by REGULATORY AGENCY from operating the vehicle unless it is certified roadworthy. MANUFACTURER *commits* to REGULATORY AGENCY for installing a device that safely disables driving the car if it operated for more than an hour without certification. OWNER is *empowered* by REGULATORY AGENCY to override the disablement by declaring an emergency.
- *Promoting privacy.* MANUFACTURER is *authorized* by OWNER to record operational information but *prohibited* from sharing it with third parties unless consent is obtained.
- *Promoting safety.* REGULATORY AGENCY commits to PUBLIC to monitor, investigate, and document the nature and causes of incidents.

We adopt the distinction between *prescriptive* (or injunctive) norms and *descriptive* norms (Cialdini, Reno, and Kallgren 1990). The distinction is that prescriptive norms state what an agent ought to do whereas the descriptive norms what agents *normally* do. For clarity, by *norms* of an STS, we always mean its prescriptive norms. And by *outcomes*, we mean the state of the STS as represented by the events that have occurred. The descriptive norms would be some time-limited aggregation of an STS's state; they are by definition *emergent*. Prescriptive norms have regulatory force and could arise merely from *constituting* emergent norms in a process of governance (as we explain below).

For instance, the regulatory agency may be unable to monitor incidents in bad weather conditions because of failures of sensors. Some outcomes may even fall outside the scope of the norms in the sense that they result from actions that do not result in norm violations but could be in conflict with the values. For example, the manufacturer may monitor what passengers are talking about in the car and use that information to target advertisements. Although not in violation of any specified norm, such monitoring would be in tension with preserving the privacy of passengers.

3.2 Governance

Governance is an interactive activity among stakeholders whereby they try to align the norms of an STS with their values, which may be themselves be informed by ethical considerations and information about the performance of the STS, including the behavior of agents and technical components. The following three activities are crucial to effective governance.

Design. Does the STS, conceived of as a specification of norms, satisfy all stakeholders' requirements? A related question is whether adequate consideration was given to identifying and engaging all potential stakeholders.

Enactment. Did the principals in an STS behave as expected? For example, assuming an STS for self-driving cars has been designed, we can ask if a manufacturer designed the car's control algorithms as expected by safety standards bodies; if an owner took it for inspection and maintenance at expected intervals; if a garage applied the software updates as expected by the manufacturer; and so on.

Adaptation. The first specification of an STS is unlikely to be its final one. Data gathered about STS enactments and changing requirements would inform the adaptation of the STS. For example, it may be discovered that a certain model of self-driving car is able to handle an unforeseen situation better than other models because of the sophistication of its control algorithm. Then the STS may be changed to incorporate the relevant aspects of the algorithms in the appropriate safety standards. Adaptation is about feeding back outcomes, including descriptive norms, back into the design of prescriptive norms.

4 Accountability and Its Pitfalls

Autonomy and accountability are fundamental concepts in understanding sociotechnical systems. Autonomy means each principal is free to act as it pleases; accountability means that a principal may be called upon to account for its actions. In general, balancing autonomy and accountability is crucial for ensuring that an STS would not devolve into the extremes of chaos or tyranny.

We understand an agent's autonomy not merely in cognitive terms, that is, as a matter of its intelligence and capabilities, but also the ability to do the unexpected in violation of applicable norms, social or personal. Innovation presupposes the willingness to deviate from norms. Therefore, accountability becomes all important in light of autonomy, for to be accountable to someone means that you can get called up to explain your actions.

Accountability is classically understood, e.g., in political theory (Grant and Keohane 2005) and healthcare (Emanuel 1996), in terms of the standing of one party—the *account-taker*—to expect certain behavior from another—the *account-giver*. That is, any accountability relationship is inherently a normative relationship. In fact, norms and accountability relationships are inseparable. However, computer science approaches on accountability lose a lot of its core intuitive basis.

4.1 Traceability

Some approaches labeled "accountability," e.g., (Argyraki et al. 2007; Haeberlen 2010), address traceability of actions: traceability is an important mechanism for holding someone accountable, but is neither necessary nor sufficient for accountability. Traceability is not necessary because accountability holds even without adequate traceability and could be adequate depending upon assumptions of trustworthiness. For example, a patient may hold a hospital accountable for loss of privacy even if the loss was caused by an untraceable attacker or by equipment failure. Traceability is not sufficient because even perfect traceability can be circumvented through external interactions. For example, if Alice circumvents traceability by getting Bob to act on her behalf, she remains accountable for the norms she violates.

4.2 Utilities

Approaches based on utility, e.g., (Feigenbaum et al. 2011; Feigenbaum, Jaggard, and Wright 2011; Conitzer et al. 2017) treat accountability as the negative utility accrued by the accountable party for failing to act as expected. Consider the following example to understand the shortcomings of the above view. A nurse Don is prohibited from giving a Schedule III Controlled Substance to a patient Charlie without a prescription from a physician Alice. Let us suppose Don risks losing his bonus if he violates the prohibition. First, negative payoffs may serve as a deterrent, but in providing an assurance mechanism, they remove accountability. In essence, instead of norm *N*, Don is accountable for the norm "*N*, but if you violate *N*, then penalty." Don need no longer give an account for violating *N* provided he pays the penalty. Second, seeing that Charlie is flat-lining, Don may know that the probability of punishment is zero, but that does not mean Don is not accountable for administering controlled drugs. Third, sanctioning (including rewarding (Radcliffe-Brown 1934)) an accountable party is a process that is subsequent to accountability, not incorporated in its definition (Emanuel and Emanuel 1996; Grant and Keohane 2005). Indeed, Don could gain acclaim (a social reward) if his quick action saves Charlie's life.

4.3 Algorithms in Normative Settings

Bostrom and Yudkowsky (2014) give the example of a financial company that uses an AI algorithm-based process for approving mortgage applications. The algorithm does not explicitly rely on any race-related criterion in granting loans to applicants, and yet it turns out that the approval process increasingly favors approving applications from white applicants and rejecting those from black applicants. A rejected applicant brings a lawsuit against the company alleging discrimination. The company argues otherwise based on the design of its algorithm. An investigation finds that the algorithm rejects worthy black applicants. Bostrom and Yudkowsky point out that it may be difficult, even impossible, to figure out the behavior of an advanced AI and use this example to motivate desirable "social" properties of AI algorithms, such as transparency and predictability.

Below, we examine the different ways in this narrative may unfold once an agency starts investigating the outcomes of the algorithm. For easier reference, let's adopt these names: the financial company, Finn; Finn's algorithm, Algo; and the regulatory agency, HUD. Suppose HUD finds the outcomes produced by Algo questionable, because in aggregate they fly in the face of non-discrimination norms, those norms being codified as law. HUD calls Finn to account. HUD may determine that Finn did not use any ethnicity-related criterion in Algo and did not violate any fair-lending norms. Let's suppose HUD determines that Finn had no foreknowledge that Algo would produce the undesired or that Finn was negligent in any manner. Based upon these facts, HUD absolves Finn of any wrongdoing.

We make the following observations before proceeding.

- In any organization of autonomous principals, norms serve as the standard of correct behavior.
- Even when principals complying with the norms, questionable outcomes may be produced.
- Some principal should be accountable for the outcomes (and, therefore, the mechanisms that participate in producing it).
- Accountability is distinct from blame, which is closer to the notion of a (negative) sanction. Finn is de facto accountable to HUD because it is the subject of norms but found to be not blameworthy.

5 Ethics in the Large

Instead of focusing on problems of ethics and narrowly on agents, we should focus on *ethics in the large*; that is, we should focus on making computational the sociotechnical systems in which agents are embedded and their governance. The norms of an STS are its *objective* ethics and yield an operational basis for decision-making within the system.

We think of the norms of an STS as being its ethics for all practical purposes. When a principal adopts a role in an STS, the STS's norms become a primary focus of the principal's attention and decision-making. No driving school instructs its pupils in grand ethical dilemmas; all of them do however focus on drilling into their pupils the mundane norms of driving, e.g., to keep your hands on the steering wheel at all times and do not use a phone while driving.

Formulating appropriate norms for any complex STS would seldom be straightforward and may involve considerable deliberation and argumentation as various stakeholders try to advance their own agendas. It is important that the norms be clear or even objectively evaluable so as to avert disputes about their satisfaction or violation.

5.1 Autonomy

In the foregoing, we have narrowed the problem of ethical decision-making by agents as prima facie acting in a norm-compliant manner. And accountability means that even they are noncompliant, they may not be blamed. Further, governance means that noncompliant behaviors may be instituted as required behavior by changing the constituted norms if those behaviors are seen to promote the underlying values. Such modifications to the constitution go back to our point about autonomy and innovation. Whereas the norms serve as the ethics, in principle, agents are free to act as they please. Of course, they may also have to bear the brunt of negative sanctions, if any.

Autonomy in AI is typically conceived of in terms of cognitive capacity and the ability of an agent to perform complex tasks without supervision, reflecting perhaps in the complexity of the agent itself. In our conception, an autonomous agent has no meaningful autonomy unless it is connected via normative relationships and accountability to other agents that circumscribe autonomy. These relationships are public and they exist outside the agent by fact of public communication between agents, regardless of whether an agent may believe or intend (Singh 1998). In other words, autonomy is a social phenomenon in our thinking, not a cognitive one. To be autonomous is to be accountable and to be accountable is to be autonomous.

5.2 Ethical STSs

We claim that the idea of being *ethical* applies to STSs as well. There is a history in human discourse of distinguishing the system from an individual and claiming, e.g., that the system was unfair so that the individual had no choice but to do something unethical in a certain situation. That is, the problem originated in the system. This tension is a source of poignancy in literature, such as Les Misérables. We can thus talk about ethical healthcare systems. For example, many would find that a system that leaves people uninsured for all practical purposes because of preexisting conditions would be unethical. The frame of reference for this evaluation is potentially a set of values that either do not coincide with the set of values of this healthcare system under consideration. Or, perhaps their values align, and the shortcomings fall to the design of the healthcare system being poorly characterized in term of its norms. This point leads us to a conception of ethical STSs.

Definition 1 *An STS S is ethical at time t from the point of view of values V if and only if S's outcomes align with V at t.*

This definition has several interesting features. One, it emphasizes outcomes over norms. Norms are crucially instrumental to outcomes, giving an operational basis for actions by agents, but in the end, it is the actual outcomes that matter. The values V provide the frame of reference. V could be the values of S itself but they could as well be the values of another STS T. Such a conception enables looking at an STS from the point of view of another. S and T may be both ethical from their own respective points of view but S may be unethical from T's viewpoint and T from S's. And, finally, an STS is not always ethical or unethical. It may be ethical today but because of inadequate governance (that lets values and outcomes become misaligned) lapse into being unethical. Analogously, an unethical STS may later become ethical because of responsive governance.

5.3 Adaptability and Emergence

If all agents are constrained to be ethical in the same way, then there is less room for *innovation* in ethics. Any ethical standard has a sociocultural context (Dignum 2017), which itself is continually changing. One only has to look at the many norms that have changed over the last two centuries, e.g., concerning slavery, women's rights, and gay rights.

Adaptation is a process of alignment, that is, the minimization of deviation, between the values and norms and outcomes. This yields a design perspective: What changes would we need to make to a system to produce this alignment? For example, will adding resources help produce alignment? Alternatively, are the specified norms undesirable, perhaps too weak (do not constrain enough, no penalties for violations, and so on), or too strong (too constraining or deterring)?

Let's resume our loan-granting algorithm example. The outcome was questionable but the fair lending norms were not violated. In that case, HUD may propose altering the non-discrimination norms to accommodate factors so that the apparent loophole is closed. If the norms change, Finn must either modify Algo to accommodate the altered norms or risk noncompliance. Additionally, HUD may propose norms as part of governance activities that commit financial companies to monitor outcomes and notify the agency every quarter. Doing so would contribute to the value of fair lending.

If the outcome is deemed acceptable, Finn may continue using Algo. There is a variation where the outcome's status remains questionable. In such a case, HUD may propose just the monitoring norms. We make the following observations.

- Norms can change to guide the likely outcomes in a desirable direction. A change in norms is an adaptation of the broader sociotechnical system in which computational mechanisms operate.
- Computational mechanisms, e.g., algorithms may need to change in response or the accountable parties risk noncompliance.

6 Conclusion

Since the inception of AI as a field of study and in much of the discourse on autonomous agents, the emphasis has been on cognition, on intelligence—sense the environment, reason about and learn from it, and select the most appropriate course of action for the current moment. As agents are increasingly visible in the social sphere, it is important to model STSs in which they are embedded and develop the concomitant methodologies, infrastructures, and tools that help bring computational sociotechnical systems to life. The question of ethics is not limited to agents alone; it applies crucially and equally to STSs. Focusing on STSs provides many benefits. No agent exists in isolation. And in contrast to the difficult of ethics in general, the concrete ethics of an STS as codified in its norms provide a clear operational basis for action by agents. STS provide a basis for decision-making and accountability. We gave a definition of ethical STS in terms of values and outcomes.

A pertinent question for AI research is how can we computationally support the governance of ethical STSs. Traditional operating systems support the execution of unitary machines—more a case of top-down management than of governance in our conception (Pitt, Schaumeier, and Artikis 2012; Singh 2013). What kind of "governance system" (GS) would be needed to support an STS? An STS, unlike an application that runs on an operating system is not a unitary machine; the stakeholders are themselves part of the STS. We anticipate that the GS would support public deliberation methods such as argumentation and tools that help analyze and aggregate arguments. The stakeholders would in turn be supported by tools that help them understand the "distance" between norms and outcomes. Such tools would apply data mining to form models of the descriptive norms from the outcomes. Additional tools would help connect outcomes to values. Voting methods may support decision-making among the stakeholders. These decisions could be about, e.g., how the norms needs to be changed. Needless to say, we would also need advances in formal languages to represent norms. The future is wide open.

Acknowledgments

Chopra was supported by the EPSRC grant EP/N027965/1 (*Turtles*). Singh thanks the US Department of Defense for partial support under the Science of Security Lablet.

References

Argyraki, K.; Maniatis, P.; Irzak, O.; Ashish, S.; and Shenker, S. 2007. Loss and delay accountability for the Internet. In *Proceedings of the IEEE International Conference on Network Protocols (ICNP)*, 194–205.

Artikis, A.; Sergot, M. J.; and Pitt, J. V. 2009. Specifying norm-governed computational societies. *ACM Transactions on Computational Logic* 10(1).

Bonnefon, J.; Shariff, A.; and Rahwan, I. 2016. The social dilemma of autonomous vehicles. *Science* 352(6293):1573–1576.

Bostrom, N., and Yudkowsky, E. 2014. The ethics of artificial intelligence. In *The Cambridge Handbook of Artificial Intelligence*. 316–334.

Chopra, A. K., and Singh, M. P. 2016. From social machines to social protocols: Software engineering foundations for sociotechnical systems. In *Proceedings of the 25th International World Wide Web Conference*, 903–914.

Cialdini, R. B.; Reno, R. R.; and Kallgren, C. A. 1990. A focus theory of normative conduct: Recycling the concept of norms to reduce littering in public places. *Journal of personality and social psychology* 58(6):1015–1026.

Conitzer, V.; Sinnott-Armstrong, W.; Borg, J. S.; Deng, Y.; and Kramer, M. 2017. Moral decision making frameworks for artificial intelligence. In *Proceedings of the 31st AAAI Conference on Artificial Intelligence*, 4831–4835.

Danks, D., and London, A. J. 2017. Algorithmic bias in autonomous systems. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence*, 4691–4697.

Dennis, L.; Fisher, M.; Slavkovik, M.; and Webster, M. 2016. Formal verification of ethical choices in autonomous systems. *Robotics and Autonomous Systems* 77(Supplement C):1–14.

Diakopoulos, N. 2016. Accountability in algorithmic decision making. *Communications of the ACM* 59(2):56–62.

Dignum, V. 2017. Responsible autonomy. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence*, 4698–4704.

Emanuel, E. J., and Emanuel, L. L. 1996. What is accountability in health care? *Annals of Internal Medicine* 124(2):229–239.

Emanuel, L. L. 1996. A professional response to demands for accountability: Practical recommendations regarding ethical aspects of patient care. *Annals of Internal Medicine* 124(2):240–249.

Feigenbaum, J.; Hendler, J.; Jaggard, A. D.; Weitzner, D. J.; and Wright, R. N. 2011. Accountability and deterrence in online life (extended abstract). In *Proceedings of the 3rd International Web Science Conference*, 7:1–7:7. Koblenz: ACM Press.

Feigenbaum, J.; Jaggard, A. D.; and Wright, R. N. 2011. Towards a formal model of accountability. In *Proceedings of the 14th New Security Paradigms Workshop (NSPW)*, 45–56.

Grant, R. W., and Keohane, R. O. 2005. Accountability and abuses of power in world politics. *American Political Science Review* 99(1):25–43.

Haeberlen, A. 2010. A case for the accountable cloud. *ACM SIGOPS Operating Systems Review* 44(2):52–57.

Jones, A. J. I., and Sergot, M. J. 1993. On the characterisation of law and computer systems: The normative systems perspective. In Meyer, J.-J. C., and Wieringa, R. J., eds., *Deontic Logic in Computer Science: Normative System Specification*. Chichester, UK: John Wiley and Sons. chapter 12, 275–307.

Kafalı, Ö.; Ajmeri, N.; and Singh, M. P. 2016. Revani: Revising and verifying normative specifications for privacy. *IEEE Intelligent Systems* 31(5):8–15.

Kafalı, Ö.; Ajmeri, N.; and Singh, M. P. 2017. Kont: Computing tradeoffs in normative multiagent systems. In *Proceedings of the 31st Conference on Artificial Intelligence (AAAI)*, 3006–3012. San Francisco: AAAI.

King, T. C.; Li, T.; Vos, M. D.; Dignum, V.; Jonker, C. M.; Padget, J.; and van Riemsdijk, M. B. 2015. A framework for institutions governing institutions. In *Proceedings of the Fourteenth International Conference on Autonomous Agents and Multiagent Systems*, 473–481. IFAAMAS.

Meyer, J.-J. C., and Wieringa, R. J., eds. 1993. *Deontic Logic in Computer Science: Normative System Specification*. Chichester, United Kingdom: Wiley.

Modgil, S.; Oren, N.; Faci, N.; Meneguzzi, F.; Miles, S.; and Luck, M. 2015. Monitoring compliance with E-contracts and norms. *Artificial Intelligence and Law* 23(2):161–196.

Moor, J. H. 2006. The nature, importance, and difficulty of machine ethics. *IEEE Intelligent Systems* 21(4):18–21.

Pitt, J.; Schaumeier, J.; and Artikis, A. 2012. Axiomatization of socio-economic principles for self-organizing institutions: Concepts, experiments and challenges. *ACM Transactions on Autonomous and Adaptive Systems* 7(4):39:1–39:39.

Radcliffe-Brown, A. R. 1934. Social sanction. In Seligman, E. R. A., ed., *Encyclopaedia of the Social Sciences*, volume XIII. Macmillan Publishers. 531–534. Reprinted in *Structure and Function in Primitive Society*, chapter 11, pages 205–211, The Free Press, Glencoe, Illinois, 1952.

Richardson, H. S. 2014. Moral reasoning. In Zalta, E. N., ed., *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, winter 2014 edition.

Russell, S.; Dewey, D.; and Tegmark, M. 2015. Research priorities for robust and beneficial artificial intelligence. http://futureoflife.org/static/data/documents/research_priorities.pdf.

Sen, A. 2011. *The Idea of Justice*. Harvard University Press.

Singh, M. P. 1998. Agent communication languages: Rethinking the principles. *IEEE Computer* 31(12):40–47.

Singh, M. P. 2013. Norms as a basis for governing sociotechnical systems. *ACM Transactions on Intelligent Systems and Technology (TIST)* 5(1):21:1–21:23.

Singh, M. P. 2015. Cybersecurity as an application domain for multiagent systems. In *Proceedings of the 14th International Conference on Autonomous Agents and MultiAgent Systems (AAMAS)*, 1207–1212. Istanbul: IFAAMAS. Blue Sky Ideas Track.

Von Wright, G. H. 1963. *Norm and Action: A Logical Enquiry*. International Library of Philosophy and Scientific Method. New York: Humanities Press.

Von Wright, G. H. 1999. Deontic logic: A personal view. *Ratio Juris* 12(1):26–38.