New Foundations of Ethical Multiagent Systems

Murukannaiah, P.K.; Ajmeri, Nirav; Jonker, C.M.; Singh, Munindar P.

Publication date
2020

Document Version
Final published version

Published in
Proceedings of the 19th Conference on Autonomous Agents and MultiAgent Systems

New Foundations of Ethical Multiagent Systems

Blue Sky Ideas Track

Pradeep K. Murukannaiah[1], Nirav Ajmeri[2], Catholijn M. Jonker[1], and Munindar P. Singh[2]
[1]Delft University of Technology, Delft, The Netherlands
[2]North Carolina State University, Raleigh, NC
p.k.murukannaiah@tudelft.nl,najmeri@ncsu.edu,c.m.jonker@tudelft.nl,mpsingh@ncsu.edu

ABSTRACT

Ethics is inherently a multiagent concern. However, research on AI ethics today is dominated by work on individual agents: (1) how an autonomous robot or car may harm or (differentially) benefit people in hypothetical situations (the so-called trolley problems) and (2) how a machine learning algorithm may produce biased decisions or recommendations. The societal framework is largely omitted.

To develop new foundations for ethics in AI, we adopt a sociotechnical stance in which agents (as technical entities) help autonomous social entities or principals (people and organizations). This multiagent conception of a sociotechnical system (STS) captures how ethical concerns arise in the mutual interactions of multiple stakeholders. These foundations would enable us to realize ethical STSs that incorporate social and technical controls to respect stated ethical postures of the agents in the STSs. The envisioned foundations require new thinking, along two broad themes, on how to realize (1) an STS that reflects its stakeholders' values and (2) individual agents that function effectively in such an STS.

KEYWORDS

Ethics; values; sociotechnical systems; norms; preferences

ACM Reference Format:
Pradeep K. Murukannaiah[1], Nirav Ajmeri[2], Catholijn M. Jonker[1], and Munindar P. Singh[2] . 2020. New Foundations of Ethical Multiagent Systems. In *Proc. of the 19th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2020), Auckland, New Zealand, May 9–13, 2020,* IFAAMAS, 5 pages.

1 ETHICS IN MULTIAGENT SYSTEMS

The surprising capabilities demonstrated by AI technologies overlaid on detailed data and fine-grained control give cause for concern that agents can wield enormous power over human welfare, drawing increasing attention to ethics in AI.

Ethics is inherently a multiagent concern—an amalgam of (1) one party's concern for another and (2) a notion of justice. To capture the multiagent conception realistically, we model our setting as a *sociotechnical system (STS).* An STS comprises autonomous social entities (*principals*, i.e., people and organizations) and technical entities (*agents*, who help principals, and resources) [13].

What foundations do we need to build STSs that address ethical concerns from multiple perspectives? Since an agent may incorrectly follow its principal's ethical directive or correctly follow an unethical directive, an ethical STS should provide social and technical controls [29–31] to promote ethical outcomes. The STS conception leads us to formulate the problem as the one of specifying (1) an STS to respect a stated systemic ethical posture over its stakeholders' value preferences; and (2) an agent who respects a stated individual ethical posture and functions in that STS.

Existing works on AI and ethics adopt a single-party mindset in topics such as (1) algorithmic accountability [17] and fairness [25], where decisions or recommendations can be biased; and (2) the behavior of agents [16], when facing moral quandaries in hypothetical situations, such as the famous trolley problems [23].

Even MAS-oriented research on ethics largely focuses on analysis of stakeholders' values [24] with the purpose of specifying a single agent. Recent efforts by MAS researchers, e.g., [36], identify limitations of existing approaches, such as goal modeling, suggesting that current models lack important components. In contrast, we advocate realizing (1) STSs that reflect system objectives in their social architecture; and (2) agents that balance moral preferences and help their principals take ethical decisions.

Following [14, 29, 31], we seek to build on recent research on values and principles of justice [41], providing new foundations for ethical multiagent systems along three main research themes.

Model: Developing a model of ethics based on values and norms that supports individual and system-level ethical judgments based on modular criteria we call *ethical postures*.
- *Novelty:* Expanding the scope of multiagent system modeling to include norms and values, incorporating ideas on guilt and inequity aversion [19, 34], and the principles of justice [41].

Analysis: Developing reasoning techniques to help stakeholders identify potential ethical pitfalls in an STS, specifically via (1) formal verification approaches to accommodate ethics in terms of norms and values preferences; and (2) agent-based simulations for assessing the ethicality of STSs and their members..
- *Novelty*: Combining verification and simulation to assess how well an STS respects a system-level ethical posture such as utilitarianism and egalitarianism [35, 40].

Elicitation: Develop techniques to specify an acceptable STS based on value-based negotiation between concerned stakeholders and to elicit value preferences from stakeholders.
- *Novelty:* Enhancing negotiation and deliberation techniques to focus on values and unintrusive learning of value preferences.

2 SOCIOTECHNICAL SYSTEM (STS)

We begin from a description of a sociotechnical system (STS) adapted from Kafalı et al. [29], introducing the necessary concepts underlying our conception. Section 4 discusses additional literature.

Figure 1 shows an STS (right frame) and how we envision such an STS being engineered (left frame).

Proc. of the 19th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2020), B. An, N. Yorke-Smith, A. El Fallah Seghrouchni, G. Sukthankar (eds.), May 9–13, 2020, Auckland, New Zealand.

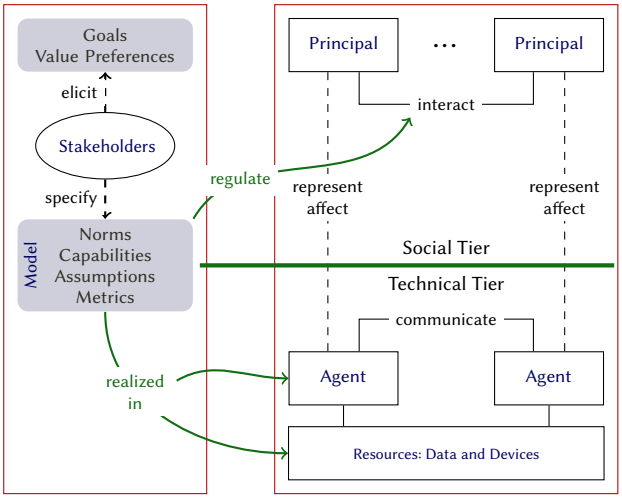


Figure 1: A schematic of a sociotechnical system (STS) [29].

A *stakeholder* in an STS is an autonomous entity (individual or organization) that has an interest in the specification or operations of the STS. For example, the stakeholders of a patient transfer [1] STS include patients, doctors, nurses, and the hospital; for a phone users' [37] STS, the caller, the callee, as well as the people and organizations nearby (e.g., a library is interested in the keeping the phones of people in the library silent) are the stakeholders.

A *principal* is a stakeholder that is active in a system. A principal can choose its actions in the system. Our applications of interest emphasize interactions among principals whereby they exchange information and services, e.g., as in social media, scientific collaboration, and healthcare. A stakeholder who is not a principal would have an interest in the specification of a system but does not participate as a decision maker. For example, in patient transfer, a nurse and physician are principals, but a patient in general is not.

When an STS is operational, its social tier includes principals and its technical tier includes agents and underlying resources, such as databases, services, sensors, and actuators. The agents act on behalf of the principals and their actions affect the principals: in many-to-many relationships, shown as one-to-one for simplicity.

Engineering an STS involves identifying its stakeholders and eliciting their goals (reflecting domain requirements) and value preferences to produce a model that specifies the STS along with its environmental (operating) assumptions and metrics. The specification captures the STS's (1) technical architecture in terms of capabilities, viewed abstractly as actions on resources that participants can perform; and (2) social architecture in terms of the principals' roles and the norms capturing the legitimate expectations between them and the consequences of the actions.

2.1 Ethical Postures

An individual ethical posture refers to how an agent may respond to the value preferences of a principal in the STS who is affected by the agent's actions. An example ethical posture would be to reflect the common intuition that a decision is ethical if it accommodates the preferences of others besides oneself.

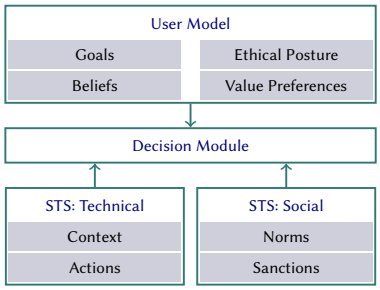
A systemic ethical posture refers to how an STS is specified in light of the value preferences of its stakeholders. Examples of ethical postures include (1) egalitarianism, i.e., to minimize disparity across stakeholders with respect to satisfying their preferences; and (2) utilitarianism, i.e., to maximize aggregate welfare (the greatest good of the greatest number) without regard to any disparities.

The principals who participate in an STS may adopt a different ethical posture from what is incorporated in the STS and different value preferences from those of the stakeholders who specified it.

2.2 Agents

Figure 2 illustrates an agent's representation and decision making. An agent's *user model* describes the agent's user in terms of goals, beliefs, value preferences, and ethical posture. An agent's *user* is a principal on whose behalf the agent acts and interacts. The agent's decisions may affect not only its user but also other principals in the STS. The agent maintains knowledge of the STS in which it functions, including how its principals relate. An agent's *decision module* produces *actions* that reflect the value preferences of its user's ethical posture, the ethical postures of other concerned principals, as well as the systemic ethical posture.

Figure 2: An agent that functions in an STS by representing and reasoning about its technical and social architectures, the goals, values, preferences, and ethical postures of the principals in an STS.



3 RESEARCH CHALLENGES AND QUESTIONS

Our research objective is to create new multiagent foundations for AI ethics. To this end, we advocate addressing the shortcomings of current approaches for AI ethics by, respectively, developing (1) new representations and reasoning approaches about ethics from both individual and system perspectives; (2) new ways of analyzing systems with respect to an ethical posture both statically (verification) and dynamically (simulation); and (3) new ways to elicit value preferences from stakeholders and to assist them in negotiating acceptable STS specifications. A practical approach would have many components. However, we concentrate on novel challenges that we posit as having the highest prospect of reward.

3.1 Model of Ethics for AI

Q_1 Representation. What is an appropriate model and representation of an STS and of an agent from the standpoint of ethics?

Q_2 Decision making. How can we support decision making by an agent in an STS that takes into account the value preferences of its user and other principals as well as the STS?

Motivation. To represent an STS precisely and reusably, we need a sufficiently rich language that supports not only the necessary normative relationships but also provides an ability to capture time

(to support constraints on ordering and occurrence), strength (to use as a basis for determining preferences to handle conflicts), and context (to modulate the outcomes of violation, for instance).

In our conception, an STS is not a separate running entity but is realized through the interactions of the principals, agents, and resources that feature in the STS. Each agent must represent (1) its view of the social architecture of the STS, including the normative relationships in which it participates; (2) its view of the technical architecture of the STS, especially the context (state of the world) and actions that may be performed in it; and (3) the goals and value preferences of its user and other principals. Such a representation enables agents to make decisions that balance the above elements.

Language. Develop a language with a suitable syntax and semantics for specifying a general family of values and norms.

This language should enable the specification of an STS including a model of the specific principals and agents featuring in it. It should support specifying value preferences for each principal where the preferences could be expressed as ordinals or as cardinal values.

Consent is a central construct in ethics and accountability that has not received adequate attention from AI researchers. Consent characterizes when an action by one autonomous party gains legitimacy despite potentially infringing upon the autonomy or authority of another party, memorably called the "moral magic of consent" [6, 28]. As these works and others, e.g., [43], indicate the intuitions about consent are far from established in the legal literature. Two major competing intuitions [43] are that consent reflects (1) a mental action of the consenting party, indicating that it is the exercise of an internal choice; and (2) a communicative act or performative by the consenting party conferring powers on the recipient, indicating that it is the exercise of a normative power [27, 32].

The existing literature on consent focuses on a *retrospective* view (which is to adjudicate on some apparent violation, as in a court of law) but in AI ethics the *prospective* view is arguably more important (since it is about decisions to be made by an agent on the fly).

Decision Making. Realize prosocial agents that model value preferences of not only their respective users but also the other principals affected by the agents' actions.

Specifically, can an agent's decision making reflect its user's ethical posture and the value preferences of the principals affected by its actions? A particular ethical posture is inequity aversion [19], which maps to the informal concept of guilt. When an inequity averse agent doesn't act in accordance with the value preferences of a principal, it accumulates guilt (on behalf of its user). The guilt applies deferentially when it follows or deviates from a norm. Such an agent may anticipate guilt from taking a dubious action, which feeling may discourage the agent from taking that action [34].

3.2 Analysis of Ethicality

Q_3 Verification. How can we verify that an STS specification satisfies the stakeholder requirements with respect to a given systemic ethical posture?

Q_4 Simulation. How can we enable stakeholders of an STS to assess an STS specification in reference to actual or imputed ethical postures of the principals who would realize that STS?

Motivation. As we model ethics, it is important to analyze an STS specification on measures such as liveness (something good happens), safety (nothing bad happens), robustness (how long something good keeps happening), and resilience (how quickly an STS recovers from something bad). Such analyses necessitate the use of (1) formal verification to assess the STS specification, and (2) simulation to foresee an outcome. Bremner et al. [12] present a leading approach for formal verification geared toward ethical reasoning, incorporating beliefs, desires, and values in a framework based on planning. This approach can help advance the present agenda.

Verification of STSs and Agents. Develop new model checking approaches that consider value preferences of stakeholders and work on top of existing probabilistic model checking tools.

Given an STS specification and the knowledge of outcomes promoted by values, an enhanced verification tool would help us understand whether the specification is biased toward certain values. For example, we may identify that a phone ringer agent always prefers safety over privacy. That agent might ring a user's phone loud for a call from a family member. How can we adapt emerging model checking tools for these purposes? A source of complexity in our setting is that we represent both the STS specification and the agents who support its principals. Because of the requirement of autonomy, any norm may be violated [46], though norms provide a basis for accountability. And, in general, we cannot interpret value preferences as expected utilities as is conventional in game theory.

Thus, a research challenge is how to formulate the correctness problems. We anticipate that correctness properties would be assessed (1) separately for an STS and conditional upon an STS for its member agents; and (2) qualitatively with respect to ethical postures of individual agents and of the system.

Although formal verification can help assess an STS specification under general assumptions, social simulations provide us with an avenue to foresee the runtime outcome.

Social Simulation. Enable stakeholders to guide the simulation, and subsequently help them understand the outcomes in an STS if a certain type (or group) of individuals were to interact in it.

For example, if the phone user is traveling extensively for work and is attending meetings, the simulation will help the stakeholder determine that in an STS specification biased toward safety over privacy, the agent will ring the phone loud more frequently than in an STS that balances safety and privacy depending on the user's context; such an agent will thus deviate more often from STS norms and attract more sanctions from agents of other principals.

Can we generate social dilemma scenarios for each user based on an understanding of the user's previous interactions and known value preferences? These social dilemma situations include cases where (1) multiple norms conflict, (2) one or more norms conflict with value preferences of a user, (3) value preferences of one user conflicts with those of other users in the interaction.

3.3 Elicitation of Ethical Systems

Q_5 Learning. How can an agent elicit its users' value preferences?

Q_6 Negotiation. How can we enable stakeholders to create an STS specification that accords with their value preferences?

Motivation. Can agents act in ways that align with the values of principals? To do so, an agent must, first, recognize the value preferences of its principals, which can be extremely challenging. First, asking the principals what values they prefer over others directly (e.g., via a survey) is likely to be futile. As Bostyn et al. [11] show, responses to hypothetical questions on moral preferences (e.g., as in the trolley problem surveys) do not predict the behavior of the participants in real life. Thus, an agent must learn its principals' value preferences by observing what the principals do in real decision scenarios and reasoning about why they did so.

Second, value preferences can be context specific—a principal may prefer one value to another ($v_1 \succ v_2$) in a context but have the opposite preference ($v_2 \succ v_1$) in another context. For example, consider Charlie, a principal who is visually impaired. Charlie prefers safety (a value) to privacy (another value) when he is traveling (a context). Thus, when Charlie is traveling, his agent automatically takes pictures of his surroundings and shares them with his friends. However, if Charlie is traveling with Dave, a trusted friend (another context), there is no need for Charlie's agent to compromise privacy by sharing pictures. Third, although an agent needs to learn its user's values, those values, in turn, may depend on the values of other principals with which the agent and its user interact.

As the examples above suggest, learning value preferences involves recognizing and modeling several nuances. Even for a small set of core values of interest in an application scenario, there can be a large number of value preferences, considering the variety of physical and social contexts in which the preferences apply.

Learning Value Preferences. Learn value preferences by observing (1) the principal's actions; (2) whether the principal approves or disapproves the agent's actions; and (3) whether other principals sanction the agent's actions, positively or negatively.

This problem is fundamentally different from the typical preference learning problems, e.g., [3, 44], whose objective is to learn preferences from pairwise comparisons of items of interest. As we argue above, directly eliciting preferences between value pairs from principals may not yield desirable outcomes. In contrast, we seek to learn value preferences by observing what principals and agents do (as opposed to what they say) in different contexts. Ajmeri et al. [5] show how value preferences can be aggregated to identify a consensus action which is fair to all stakeholders involved.

Knowing the value preferences of stakeholders helps in better facilitating interaction between them. Interest-based negotiation [21] is based on the idea that stakeholders' goals may differ from their positions during negotiation. Thus, satisfying their (imputed) goals is better than giving them what they explicitly ask for.

Prior negotiation protocols for settings related to STSs, e.g., [8, 10], accommodate neither values nor the entire breadth of an STS as conceived here. Existing approaches, e.g., [42], focus on eliminating conflicts among negotiating parties. In contrast, we bring forth conflicts as a basis for negotiation of an STS (during elicitation) and as an input into ethical decision making by an agent (at run time).

Value-Based Negotiation. Support stakeholders with conflicting requirements but similar value preferences in generating an acceptable STS specification.

In our conception of value-based negotiation, each offer comprises an STS specification. A stakeholder can reason about how the current offer contributes to that stakeholder's preferred values to decide the response move: accept, reject, or generate a counteroffer. Facilitating such reasoning requires (1) a normative negotiation framework for the specification of STSs that provides a basis for systematically revising norms to enable the generation of effective offers and counteroffers; and (2) a value-based concession bidding strategy that adapts its offers at run time based on opponent's behavior without predefined utility functions.

4 ETHICS AND RELATED CONSTRUCTS

An AI system is neither merely an algorithm nor a standalone agent, but sociotechnical system representing a society of humans and agents. Accordingly, there is a need and urgency for addressing societal concerns on AI adoption. Ethics is one such concern but it is closely related to other societal concerns on AI, including fairness, accountability, transparency, and privacy. The new foundations we call for can and should address these related concerns as well.

Fairness concerns judgments on the outcomes of machine learning predictors. Research on fairness in AI [18] and how people assess AI fairness [9, 33, 48] focuses on an individual (is it fair to me?) or system (is the system fair as a whole?), but not on a *group*, incorporating the preferences of stakeholders and their social relationships and power dynamics. To achieve group fairness, each agent must support fairness in decision making by understanding contextually relevant norms [4] and reasoning about value preferences of all stakeholders [5], not just of the agent's user.

Accountability is crucial to establishing who is accountable for a decision made by an agent [15]. Prior works understand accountability as either traceability [7, 26] or negative utility [20], but these concepts are neither necessary nor sufficient for capturing accountability because they lack the social-level semantics that undergirds accountability. We seek to capture the normative basis of accountability directly though it supports traceability and sanctioning where appropriate.

Transparency relates to the principle of explicability and concerns traceability [2]. We seek to support these desired principles of responsible AI through traceability of STS negotiation steps as well as explicability of agents' reasoning at runtime.

Privacy is naturally approached from a values perspective [24, 45]. It encompasses values such as confidentiality, disapprobation, and avoiding infringing into others' space [4, 22]. Researchers advocate giving greater control to users on decision making, e.g., for privacy [47]. However, giving control to users raises the question of whether a user's action accords with that user's or other concerned users' values. Social norms are the centerpiece of *contextual integrity* [38, 39], a theory of privacy where violations occur when information flows violate contextual norms.

ACKNOWLEDGMENTS

This research is partially supported by the US DoD through the Science of Security Lablet (SoSL) at NCSU and is part of the research programme Hybrid Intelligence with project number 024.004.022, which is (partly) financed by the Dutch Ministry of Education, Culture and Science (OCW).

REFERENCES

[1] Joanna Abraham and Madhu C. Reddy. 2010. Challenges to Inter-Departmental Coordination of Patient Transfers: A Workflow Perspective. *International Journal of Medical Informatics* 79, 2 (Feb. 2010), 112–122.

[2] AIHLEG. 2019. Ethics Guidelines for Trustworthy AI. Independent High-Level Expert Group on Artificial Intelligence set up by the European Commission. https://ec.europa.eu/newsroom/dae/document.cfm?doc_id=60419.

[3] Nir Ailon. 2012. An Active Learning Algorithm for Ranking from Pairwise Preferences with an Almost Optimal Query Complexity. *The Journal of Machine Learning Research* 13, 1 (Jan. 2012), 137–164.

[4] Nirav Ajmeri, Hui Guo, Pradeep K. Murukannaiah, and Munindar P. Singh. 2018. Robust Norm Emergence by Revealing and Reasoning about Context: Socially Intelligent Agents for Enhancing Privacy. In *Proceedings of the 27th International Joint Conference on Artificial Intelligence (IJCAI)*. IJCAI, Stockholm, 28–34. https://doi.org/10.24963/ijcai.2018/4

[5] Nirav Ajmeri, Hui Guo, Pradeep K. Murukannaiah, and Munindar P. Singh. 2020. Elessar: Ethics in Norm-Aware Agents. In *Proceedings of the 19th International Conference on Autonomous Agents and MultiAgent Systems (AAMAS)*. IFAAMAS, Auckland, 1–9.

[6] Larry Alexander. 1996. The Moral Magic of Consent (II). *Legal Theory* 2, 3 (Sept. 1996), 165–174.

[7] Katerina Argyraki, Petros Maniatis, Olga Irzak, Subramanian Ashish, and Scott Shenker. 2007. Loss and Delay Accountability for the Internet. In *Proceedings of the IEEE International Conference on Network Protocols (ICNP)*. IEEE, Beijing, 194–205.

[8] Reyhan Aydoğan, David Festen, Koen V. Hindriks, and Catholijn M. Jonker. 2017. Alternating Offers Protocols for Multilateral Negotiation. In *Modern Approaches to Agent-based Complex Automated Negotiation*, Katsuhide Fujita, Quan Bai, Takayuki Ito, Minjie Zhang, Fenghui Ren, Reyhan Aydoğan, and Rafik Hadfi (Eds.). Number 674 in Studies in Computational Intelligence. Springer, Cham, 153–167. https://doi.org/10.1007/978-3-319-51563-2_10

[9] Reuben Binns, Max Van Kleek, Michael Veale, Ulrik Lyngs, Jun Zhao, and Nigel Shadbolt. 2018. 'It's Reducing a Human Being to a Percentage': Perceptions of Justice in Algorithmic Decisions. In *Proceedings of the Conference on Human Factors in Computing Systems (CHI)*. ACM, Montreal, 377:1–377:14.

[10] Guido Boella, Patrice Caire, and Leendert van der Torre. 2009. Norm Negotiation in Online Multi-Player Games. *Knowledge and Information Systems* 18, 2 (Feb. 2009), 137–156.

[11] Dries H. Bostyn, Sybren Sevenhant, and Arne Roets. 2018. Of Mice, Men, and Trolleys: Hypothetical Judgment Versus Real-Life Behavior in Trolley-Style Moral Dilemmas. *Psychological Science* 29, 7 (2018), 1084–1093. https://doi.org/10.1177/0956797617752640 PMID: 29741993.

[12] Paul Bremner, Louise A. Dennis, Michael Fisher, and Alan F. T. Winfield. 2019. On Proactive, Transparent, and Verifiable Ethical Reasoning for Robots. *Proc. IEEE* 107, 3 (March 2019), 541–561. https://doi.org/10.1109/JPROC.2019.2898267

[13] Amit K. Chopra, Fabiano Dalpiaz, F. Başak Aydemir, Paolo Giorgini, John Mylopoulos, and Munindar P. Singh. 2014. Protos: Foundations for Engineering Innovative Sociotechnical Systems. In *Proceedings of the 22nd IEEE International Requirements Engineering Conference (RE)*. IEEE Computer Society, Karlskrona, Sweden, 53–62. https://doi.org/10.1109/RE.2014.6912247

[14] Amit K. Chopra and Munindar P. Singh. 2016. From Social Machines to Social Protocols: Software Engineering Foundations for Sociotechnical Systems. In *Proceedings of the 25th International World Wide Web Conference*. ACM, Montréal, 903–914. https://doi.org/10.1145/2872427.2883018

[15] Amit K. Chopra and Munindar P. Singh. 2018. Sociotechnical Systems and Ethics in the Large. In *Proceedings of the AAAI/ACM Conference on Artificial Intelligence, Ethics, and Society (AIES)*. ACM, New Orleans, 48–53. https://doi.org/10.1145/3278721.3278740

[16] Louise A. Dennis, Michael Fisher, Marija Slavkovik, and Matt Webster. 2016. Formal Verification of Ethical Choices in Autonomous Systems. *Robotics and Autonomous Systems* 77 (March 2016), 1–14.

[17] Nicholas Diakopoulos. 2016. Accountability in Algorithmic Decision Making. *Communications of the ACM (CACM)* 59, 2 (Feb. 2016), 56–62.

[18] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. 2012. Fairness Through Awareness. In *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference (ITCS)*. ACM, Cambridge, 214–226.

[19] Ernst Fehr and Klaus M. Schmidt. 1999. A Theory of Fairness, Competition, and Cooperation. *The Quarterly Journal of Economics* 114 (1999), 817–868.

[20] Joan Feigenbaum, Aaron D. Jaggard, and Rebecca N. Wright. 2011. Towards a Formal Model of Accountability. In *Proceedings of the 14th New Security Paradigms Workshop (NSPW)*. ACM, Marin County, California, 45–56.

[21] Roger Fisher, William L. Ury, and Bruce Patton. 1983. *Getting to Yes: Negotiating Agreement Without Giving In* (3rd ed.). Penguin Books, New York.

[22] Ricard López Fogués, Pradeep K. Murukannaiah, Jose M. Such, and Munindar P. Singh. 2017. SoSharP: Recommending Sharing Policies in Multiuser Privacy Scenarios. *IEEE Internet Computing (IC)* 21, 6 (Nov. 2017), 28–36. https://doi.org/10.1109/MIC.2017.4180836

[23] Philippa Foot. 1967. The Problem of Abortion and the Doctrine of Double Effect. *Oxford Review* 5 (1967), 5–15.

[24] Batya Friedman, Peter H. Kahn Jr., and Alan Borning. 2008. Value Sensitive Design and Information Systems. In *The Handbook of Information and Computer Ethics*, Kenneth Einar Himma and Herman T. Tavani (Eds.). John Wiley & Sons, Hoboken, New Jersey, Chapter 4, 69–101.

[25] Sainyam Galhotra, Yuriy Brun, and Alexandra Meliou. 2017. Fairness testing: testing software for discrimination. In *Proceedings of the 11th Joint Meeting on Foundations of Software Engineering, (ESEC/FSE)*. ACM, Paderborn, 498–510.

[26] Andreas Haeberlen. 2010. A Case for the Accountable Cloud. *ACM SIGOPS Operating Systems Review* 44, 2 (April 2010), 52–57.

[27] Wesley Newcomb Hohfeld. 1919. *Fundamental Legal Conceptions as Applied in Judicial Reasoning and other Legal Essays*. Yale University Press, New Haven, Connecticut. A 1919 printing of articles from 1913.

[28] Heidi M. Hurd. 1996. The Moral Magic of Consent. *Legal Theory* 2, 2 (June 1996), 121–146.

[29] Özgür Kafalı, Nirav Ajmeri, and Munindar P. Singh. 2016. Revani: Revising and Verifying Normative Specifications for Privacy. *IEEE Intelligent Systems (IS)* 31, 5 (Sept. 2016), 8–15. https://doi.org/10.1109/MIS.2016.89

[30] Özgür Kafalı, Nirav Ajmeri, and Munindar P. Singh. 2017. Kont: Computing Tradeoffs in Normative Multiagent Systems. In *Proceedings of the 31st Conference on Artificial Intelligence (AAAI)*. AAAI, San Francisco, 3006–3012.

[31] Özgür Kafalı, Nirav Ajmeri, and Munindar P. Singh. 2019. Specification of Sociotechnical Systems via Patterns of Regulation and Control. *ACM Transactions on Software Engineering and Methodology (TOSEM)* 29, 1 (Dec. 2019), 7:1–7:50. https://doi.org/10.1145/3365664

[32] Felix Koch. 2018. Consent as a Normative Power. In *The Routledge Handbook of the Ethics of Consent*, Peter Schaber and Andreas Müller (Eds.). Routledge, London, Chapter 3, 32–43.

[33] Min Kyung Lee. 2018. Understanding perception of algorithmic decisions: Fairness, trust, and emotion in response to algorithmic management. *Big Data & Society* 5, 1 (Jan. 2018), 1–16.

[34] Emiliano Lorini and Roland Mühlenbernd. 2015. The Long-Term Benefits of Following Fairness Norms: A Game-Theoretic Analysis. In *Proceedings of the 18th International Conference on Principles and Practice of Multi-Agent Systems (PRIMA) (Lecture Notes in Computer Science)*, Vol. 9387. Springer, Bertinoro, Italy, 301–318. https://doi.org/10.1007/978-3-319-25524-8_19

[35] John Stuart Mill. 1863. *Utilitarianism*. Longmans, Green and Company, London.

[36] Tim Miller, Sonja Pedell, Antonio A. Lopez-Lorca, Antonette Mendoza, Leon Sterling, and Alen Keirnan. 2015. Emotion-Led Modelling for People-Oriented Requirements Engineering: The Case Study of Emergency Systems. *Journal of Systems and Software* 105 (July 2015), 54–71.

[37] Pradeep K. Murukannaiah and Munindar P. Singh. 2014. Xipho: Extending Tropos to Engineer Context-Aware Personal Agents. In *Proceedings of the 13th International Conference on Autonomous Agents and MultiAgent Systems (AAMAS)*. IFAAMAS, Paris, 309–316. https://doi.org/10.5555/2615731.2615783

[38] Helen Nissenbaum. 2004. Privacy as Contextual Integrity. *Washington Law Review* 79, 1 (Feb. 2004), 119–157.

[39] Helen Nissenbaum. 2011. A Contextual Approach to Privacy Online. *Dædalus, the Journal of the American Academy of Arts & Sciences* 140, 4 (Fall 2011), 32–48.

[40] John Rawls. 1985. Justice as Fairness: Political not Metaphysical. *Philosophy and Public Affairs* 14, 3 (Summer 1985), 223–251.

[41] John Rawls. 1999. *A Theory of Justice* (2nd ed.). Harvard University Press, Cambridge, Massachusetts.

[42] Jéssica Soares dos Santos, Jean de Oliveira Zahn, Eduardo Augusto Silvestre, Viviane Torres da Silva, and Wamberto Weber Vasconcelos. 2017. Detection and Resolution of Normative Conflicts in Multi-Agent Systems: A Literature Survey. *Journal of Autonomous Agents and Multi-Agent Systems (JAAMAS)* 31, 6 (Nov. 2017), 1236–1282.

[43] Hubert Schnüriger. 2018. What is Consent? In *The Routledge Handbook of the Ethics of Consent*, Peter Schaber and Andreas Müller (Eds.). Routledge, London, Chapter 2, 21–31.

[44] Nihar B. Shah and Martin J. Wainwright. 2017. Simple, Robust and Optimal Ranking from Pairwise Comparisons. *The Journal of Machine Learning Research* 18, 1 (Jan. 2017), 7246–7283.

[45] Munindar P. Singh. 2015. Cybersecurity as an Application Domain for Multiagent Systems. In *Proceedings of the 14th International Conference on Autonomous Agents and MultiAgent Systems (AAMAS)*. IFAAMAS, Istanbul, 1207–1212. https://doi.org/10.5555/2772879.2773304 Blue Sky Ideas Track.

[46] Munindar P. Singh and Amit K. Chopra. 2020. Computational Governance and Violable Contracts for Blockchain Applications. *IEEE Computer* 53, 1 (Jan. 2020), 53–62. https://doi.org/10.1109/MC.2019.2947372

[47] Sarah Spiekermann and Lorrie Faith Cranor. 2009. Engineering Privacy. *IEEE Transactions on Software Engineering* 35, 1 (Jan.–Feb. 2009), 67–82.

[48] Allison Woodruff, Sarah E. Fox, Steven Rousso-Schindler, and Jeffrey Warshaw. 2018. A Qualitative Exploration of Perceptions of Algorithmic Fairness. In *Proceedings of the Conference on Human Factors in Computing Systems (CHI)*. ACM, Montreal, 656:1–656:14.