# Guiding Sociotechnical Systems toward Value-Norm Equilibrium

Blue Sky Ideas Track

Nirav Ajmeri
University of Bristol
Bristol, UK
nirav.ajmeri@bristol.ac.uk

Marina De Vos
University of Bath
Bath, UK
cssmdv@bath.ac.uk

Davide Dell'Anna
Utrecht University
Utrecht, The Netherlands
d.dellanna@uu.nl

Pradeep K. Murukannaiah
TU Delft
Delft, The Netherlands
p.k.murukannaiah@tudelft.nl

Vivek Nallur
University College Dublin
Dublin, Ireland
vivek.nallur@ucd.ie

Luis G. Nardin
Mines Saint-Étienne
Saint-Étienne, France
gnardin@emse.fr

Munindar P. Singh
North Carolina State University
Raleigh, USA
singh@ncsu.edu

## ABSTRACT

Values and norms are complementary constructs that undergird prosocial behavior in sociotechnical systems (STSs). Whereas values are intrinsic motivators for prosocial behavior, norms are extrinsic motivators for meeting mutual expectations. An STS is in *equilibrium* when the values of its member actors and the norms that govern it align with each other. Such an equilibrium is not permanent as actors join or leave the STS, and their values and norms evolve. In general, an STS must be guided toward equilibrium by systematically refining the norm specifications and influencing the values of its member actors. We formulate the challenges involved in building systematic methods to detect misalignment and guide the STS toward, and maintain, value-norm equilibrium.

## KEYWORDS

Norms; Values; Prosociality; Alignment; Sociotechnical System

**ACM Reference Format:**
Nirav Ajmeri, Marina De Vos, Davide Dell'Anna, Pradeep K. Murukannaiah, Vivek Nallur, Luis G. Nardin, and Munindar P. Singh. 2026. Guiding Sociotechnical Systems toward Value-Norm Equilibrium: Blue Sky Ideas Track. In *Proc. of the 25th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2026), Paphos, Cyprus, May 25 – 29, 2026*, IFAAMAS, 7 pages. https://doi.org/10.65109/NSZQ3158

## 1 INTRODUCTION

A sociotechnical system (STS) involves social entities (*principals*—humans and organizations) and technical entities (artificial *agents*) [58]. Importantly, social and technical entities work as principal-agent duos [59], where the agent brings the convenience of automation and the principal shoulders accountability for the actions of the duo—the more autonomous an agent, the more accountable the associated principal. We refer to a principal-agent duo as an *actor*.

For an STS to thrive, its actors must exhibit prosocial (socially desirable) behavior [62, 64]; purely selfish behavior by the actors undermines cooperation [36]. While prosocial behavior may be sub-optimal for an actor, it leads to synergistic (win-win) outcomes for the system [11, 37, 54, 71]. Thus, it is important to recognize and nurture the factors that promote prosocial behavior.

Values and norms are two key factors that can promote prosocial behavior in an STS. Values are intrinsic motivators for an actor to act in a manner that is considered good. Schwartz [67] postulates that humans developed values as a shared vocabulary to communicate and cooperate toward societal goals (e.g., security and care). Although values are generally viewed as universal, the relative importance that actors ascribe to values can vary, resulting in (substantial or subtle) differences in the value preferences of actors. In contrast, norms are extrinsic motivators of behavior, usually specified at an organizational level (e.g., legislation, guidelines, contracts, and commitments). They represent mutual expectations among actors, which are essential for actors to cooperate [70]. However, unlike rigid rules, actors can choose to comply with or deviate from norms, provided that they can justify their behavior [69].

The intrinsic-extrinsic contrast can yield values and norms that support or work against each other, depending on whether they are aligned. We state that an STS is in *value-norm equilibrium* when its values and norms are aligned and in *value-norm disequilibrium* otherwise. In equilibrium, actors can adopt behaviors that are value-aligned and norm-compliant. A state of near-equilibrium can be sufficient for an STS to function effectively; in fact, rebellion and disobedience may be desirable in an effective STS [8, 49].

In disequilibrium, an actor's behavior may satisfy their value preferences but violate norms, or vice versa. This misalignment can be catastrophic, leading to systemic instability, widespread norm violations, mistrust in institutions, and eventual breakdown of the STS. An STS can be in disequilibrium because (1) its member actors have disparate value preferences; (2) its member actors are not aware of their value preferences until they have to act on that



*Proc. of the 25th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2026), C. Amato, L. Dennis, V. Mascardi, J. Thangarajah (eds.), May 25 – 29, 2026, Paphos, Cyprus.* © 2026 International Foundation for Autonomous Agents and Multiagent Systems (www.ifaamas.org). https://doi.org/10.65109/NSZQ3158

preference; or (3) the norms do not reflect the value preferences of all member actors (i.e., are not value-aligned). Even after an equilibrium is reached, this state may not be permanent in an STS. Value preferences may change due to actors joining or leaving the STS, and norms may evolve due to external factors (e.g., technological advancements, socioeconomic shifts, and new regulations).

The key research challenge that we pose is: *How can we guide an STS towards value-norm equilibrium?* Parts of this puzzle have been studied extensively under value-sensitive design [74], value inference [50], value alignment [46, 55], computing norms [25], norm violation and sanctioning [1, 3, 35, 60, 72]. The dynamic adaptation of norms, includes norm revision [6, 10, 28, 33, 34, 77], emergence [26, 54, 56, 65], and learning [75, 76]. Other research addresses how norms and values are related [13, 14, 44, 66]. However, how to computationally model this relationship to guide an STS toward value-norm equilibrium is largely unexplored.

To unpack this research challenge, Section 2 describes a conceptual model of the interplay between norms and values in an STS. Sections 3 and 4 outline the challenges of (1) measuring misalignment, and (2) mechanisms for guiding an STS toward equilibrium. For each challenge, we also describe our vision on addressing it.

## 2 CONCEPTUAL MODEL

Figure 1 shows our conceptualization of an STS. Each actor (principal-agent duo) is a member of the organization that governs the STS and has value preferences. The agent represents the values of its principal, which it may learn from its ongoing interactions [52, 63]. The STS has a set of norms specified in the organization [70]. These norms can be initially established via negotiation [9, 15] or learned via norm identification and mining [2, 12, 29, 53].

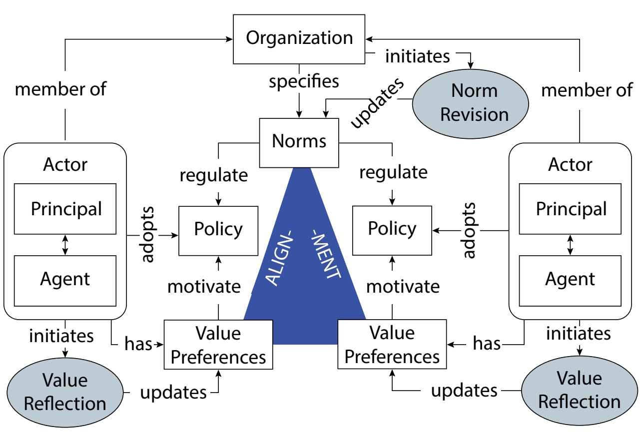


**Figure 1: Interplay between values and norms in an STS.**

Each actor in the STS seeks to acquire a policy (how to act) that aligns with the actor's value preferences and complies with the organizational norms. However, often a policy may not be able to achieve both objectives, i.e., a policy may align with a value preference but deviate from a norm [18, 19, 47], or comply with a norm but diverge from a value preference [4], illustrating a misalignment. These circumstances provide an opportunity for *norm revision* and *value reflection* to guide the STS toward equilibrium top-down (organization level) and bottom-up (actor level), respectively.

Norm revision is a process to modify the norms governing an STS [73]. Norm revision permeates the norm life cycle [45], occurring whenever norms are (1) created or dropped, (2) recognized and enforced, and (3) accepted, modified, and internalized. The norm revision process depends on a part of the STS to identify the need and initiate the revision, which varies depending on the architecture of the STS (see [78] for design options). Existing studies emphasize the process of norm revision, but not on the factors that lead to it. Norm revision can be informed by a variety of factors, e.g., the values of the actors or the "owners" (if any) of the organization [41–43], the observation of the actors complying with or violating the norms [32, 57], the effectiveness of the actors in achieving organizational objectives, and behavioral patterns.

Conversely, value reflection is internal to each actor. Reflection can result in revision (changing the preference order between values or changing the strength of a value preference) or make a latent value preference explicit to the actor.

## 3 MEASURING MISALIGNMENT

The foundation for guiding an STS toward value-norm equilibrium is to characterize, detect, and quantify the extent of misalignment. Misalignment is a crucial feedback signal that motivates norm revision by the organization or value reflection by (some) actors.

At the actor level, misalignment between norms and values can arise since norms regulate the actors' policies, which promote or demote values [30]. For instance, a norm may prohibit an actor from executing an action that promotes one of its preferred values. At the system level, the efficacy of norms is defined not only by their compliance rate, but by their ability to guide the STS toward a state in line with the aggregated values of the actors [32, 48]. Table 1 shows several examples of value-norm misalignment.

**Table 1: Examples of misalignment involving an Actor (Ind) and a Society (Soc), considering actor's action (IA) and value (IV), and societal norm (SN), value (SV), and behavior (SB).**

| Misalignment Type | Misalignment Comp | Description and Example |
|---|---|---|
| Ind | IA, IV | An actor performs actions inconsistent with their values, *e.g., a person gets vaccinated despite their religious principles opposing vaccination.* |
| Ind-Soc | IA, SN | An actor violates a norm, *e.g., a person doesn't wear a mask during a pandemic.* |
| | IV, SN | Norms that conflict with an actor's value preferences, *e.g., mandatory vaccination.* |
| | IV, SV | A member of a minority group, *e.g., a lung patient works with people who don't wear masks.* |
| Soc | SN, SV | Outdated norms, *e.g., requirement to be vaccinated is inapplicable when a pandemic ends.* |
| | SN, SB | Lack of adoption of a norm, *e.g., a strict mask norm that most passengers ignore.* |

Moving toward value-norm equilibrium requires metrics that can be monitored, compared, and used to trigger value reflection and norm revision. We identify three challenges in this direction.

### Challenge 1. Common Frame of Representation

The literature lacks explicit and comparable representations of norms and values. Norms are typically formalized in, e.g., deontic logic, institutional frameworks, or as rules [16, 27, 40]. An actor's values, on the contrary, are often latent, preference-based, and inferred from behavior. However, there is no framework that unifies normative specifications with inferred or expressed values of actors.

***Vision***. We advocate for a research agenda on *computational value-norm interfaces* to design representational frameworks that make norms and values mutually legible.

**Indicators** of misalignment, where observable events (e.g., norm deviations, norm contestations, shifts in trust toward institutions) function as proxies for the identification of value-norm misalignment. A single deviation may simply be noise, but deviation patterns and contestations would indicate misalignment.

**Proactive contestation methods** that actors can use to deliberately test the boundaries of norms through low-risk, value-promoting deviations. Such methods, coupled with justifications, can transform norm violations into feedback for norm design.

**Affective and psychological signals** as first-class abstractions in norm monitoring. Emotional responses to enforcement, perceived fairness, and experienced burden can augment purely behavioral indicators, enabling richer models of value-norm fit.

Overcoming this challenge yields a suite of indicators that connect normative specifications to value-centric observations.

### Challenge 2: Alignment Metrics

The next challenge is to aggregate and quantify the heterogeneous indicators (Challenge 1) into meaningful metrics of value-norm alignment. Existing norms related metrics focus on compliance rates or enforcement costs. Although necessary, these metrics are not sufficient to capture whether the values align with the norms in an STS. This indicates a lack of metrics that (1) distinguish high compliance due to internalization or fear of sanctions, (2) capture the degree to which norms promote (or demote) values, and (3) describe the misalignment between norms and values over time.

***Vision***. We envision STSs that are instrumented with *real-time value-norm misalignment metrics*, which yield *misalignment indices* that quantify value-norm coherence at multiple levels.

**Actor-level misalignment** indices that capture the degree to which an actor's actions satisfy its value preferences and norms.

**Group- and system-level misalignment** indices that aggregate the misalignment value preferences of actors in a group while accounting for diversity (e.g., minority values vs. majority values).

**Temporal misalignment** indices that describe characteristics of misalignment between values and norms over time, such as oscillation, trend, and stability.

We envision these indices to be used in adaptive frameworks (e.g., multi-agent reinforcement learning). Actors and organizations can incorporate these indices into their objective functions, trading off short-term efficiency against long-term value-norm coherence. This calls for multi-objective optimization and constraints-based formulations, turning value-norm alignment into a core design requirement rather than a side effect.

### Challenge 3: Value-Norm Tipping Point

This challenge concerns the dynamics of misalignment. Specifically, it refers to the magnitude and persistence of the misalignment needed to trigger value reflection and norm revision.

Although social tipping points have been studied in complex systems [23, 39], the thresholds for value-norm equilibrium in an STS have not been formalized. The gradual misalignment between norms and values can increase to a degree that leads to, e.g., (1) sudden collapse of norm adoption (e.g., mass violation of an outdated regulation), (2) rapid normative reform, or (3) value change (e.g., repeated exposure to new norms). We currently lack theories and tools to predict, detect and exploit these tipping points.

***Vision***. We identify two critical *value-norm tipping points*:

- $\mathcal{T}_{V\to N}$: threshold at which value shifts trigger norm revision. This threshold might be crossed, e.g., when actors lose confidence in the organization's ability to protect a critical value (e.g., security), causing a systematic reduction of norm adoption.
- $\mathcal{T}_{N\to V}$: threshold at which norms induce a lasting change in values. This threshold might be crossed, e.g., when a norm is successful (i.e., rapidly internalized and adopted by actors), leading to changes in their motivations and value preferences.

Monitoring proximity to $\mathcal{T}_{V\to N}$ and $\mathcal{T}_{N\to V}$ enables STSs to make informed decisions about when to revise norms and when to resist pressure to change, and actors decide when to reflect on their values preferences, rather than merely react to crises.

In this direction, we advocate for research on designing formal and simulation models as well as resilience mechanisms to estimate and detect the *proximity* to the value-norm tipping points, and implement *preemptive* and *stabilizing* actions, when necessary.

**Formal models** (e.g., evolutionary game-theoretic models) to capture how misalignment indices evolve under different norm enforcement regimes and patterns of value change.

**Simulation models** (e.g., agent-based models) to estimate where the tipping points might lie under various assumptions on sanctioning, information diffusion, and institutional responsiveness.

**Resilience mechanisms** (e.g., adaptive sanctioning, meta-norms, or temporary suspension protocols) to steer the system away from undesirable tipping points (e.g., norm collapse) or toward desirable ones (e.g., widespread internalization of norms).

## 4 VALUE-NORM EQUILIBRIUM PROCESSES

We now turn to the dynamics that drive an STS toward or away from an equilibrium. We envision these dynamics to be mediated by two complementary processes: (1) *value reflection*, a bottom-up process in which actors revise (or make explicit) their values in response to experience and normative pressure, and (2) *norm revision*, a top-down process through which organizations modify the norms that govern interactions in an STS. These processes are tightly coupled: as actors interact, their values shape emergent patterns of behavior and norms; norms in turn influence which behaviors are (positively or negatively) sanctioned, and thus shape value change over time. The core engineering challenge is to model, monitor, and control this bi-directional interplay to promote prosocial behavior and system stability. We identify three challenges along these lines.

### Challenge 4: Adaptation Timescale

The literature offers multiple approaches to learning and adapting norms [5, 54, 68]. However, few studies explicitly model the joint adaptation of norms and values—in particular, the rate and timing of such changes. In an STS, these adaptations exist in at least three timescales: (1) environmental, representing the pace of socio-economic or technological changes; (2) normative, representing how quickly norms are revised and internalized; and (3) value based, representing how quickly actors' value preferences evolve. If norm and value adaptation lag behind environmental changes, norms may become irrelevant or mistrusted, and values inappropriate. Conversely, if norms or values change too quickly, actors and organizations may fail to coordinate, leading to instability.

***Vision.*** We call for research on integrating control and dynamic systems theories in solutions of value-norm co-adaptation, enabling reasoning about rates, timing, and feedback strength.

**Model rate of mutual adaptation** to characterize a sustainable speed at which norms and values can change in response to each other and the environment without destabilizing the system.

**Timing and lag structure mechanisms** for monitoring and managing lead-lag relationships between environmental and policy changes, and adaptations in norms and values, accounting for delays in detection and information propagation (e.g., emergence detection from local, noisy observations [61]).

**Feedback strength between norms and values** to characterize and tune how norms influence value change and how value shifts feed back into norm adaptation to avoid runaway feedback loops or permanent stagnation.

This challenge opens opportunities for control-theoretic MAS, adaptive systems, simulations, and multiagent learning communities to design and analyze mechanisms that keep value-norm equilibrium within safe, stable, and productive regimes.

### Challenge 5: Value-Driven Norm Revision

Several theoretical works have emphasized the importance of norm revision [7, 22, 38]. Computational studies have focused on aspects such as changing regulative norms [16, 17, 20, 31, 33, 57] and sanctions [21, 24, 34]. As Dechesne et al. [32] recognize, legal norms are introduced with the aim of better aligning the world with values considered important. However, despite their clear link to values, norm revision mechanisms are often driven only by effectiveness or norm compliance, and not by value (mis)alignment.

The challenge advocates for developing computational frameworks to enable an organization to revise its norms and an actor to reflect on its values based on value-norm misalignment metrics.

***Vision.*** We envision value-driven norm revision mechanisms in which misalignment metrics drive a process from diagnosis to candidate revisions, to deliberation, to implementation.

**Alignment-aware diagnosis** uses misalignment indices (Section 3) to identify the norms or actors that contribute to misalignment, and distinguish transient and persistent misalignment patterns.

**Norm-space exploration** defines candidate changes to norms (e.g., modifying conditions, sanctions, priorities) and develops search or optimization methods that propose norm revisions to reduce misalignment.

**Value-sensitive deliberation** employs tools from computational social choice and argumentation to aggregate heterogeneous value preferences and to select among candidate revisions in ways that are procedurally fair, robust to manipulation, and transparent to affected actors.

This challenge creates opportunities for the normative MAS, computational social choice, argumentation, and mechanism-design communities to design norm-revision processes that are both computationally grounded and value-driven.

### Challenge 6: Facilitating Value Reflection

Current literature offers methods for value learning and alignment [50, 51, 63], but it largely lacks mechanisms of intentional, structured value reflection that are triggered by normative pressures and by persistent failure to jointly satisfy values and norms. Without value reflection, organizations may be forced to constantly adjust norms to their members' preferences, risking overfitting to transient or idiosyncratic demands. Conversely, if actors never reflect and change their value preferences, persistent misalignment can only be addressed through increasingly rigid or complex norms. Thus, the challenge is to recognize when and how to initiate value reflection in light of changes to the norms and the environment.

***Vision.*** We call for actors equipped with explicit value-reflection modules that monitor conflicts between values, norms, and behavior and trigger structured value-reflection processes when necessary.

**Misalignment-based triggers** define the criteria (e.g., repeated absence of policies that satisfy both values and norms above a threshold) to initiate reflection and design rules to adapt these thresholds over time.

**Structured reflection** provides mechanisms to re-evaluate the preference order between values and strengths of value preferences, including querying the principal when necessary.

**Integration into decision-making** ensures that updated value configurations propagate consistently into policy selection, learning objectives, and explanations, so that reflection has clear behavioral consequences and can itself be audited.

This challenge offers opportunities for the (multi-objective) learning, explainability, and human-AI interaction communities to design reflective agents that adjust values in a controlled, transparent, and human-in-the-loop manner, supporting value-norm equilibrium without undermining autonomy and pluralism.

## 5 CONCLUSION

Values and norms are two behavioral motivators which, when misaligned, destabilize an STS. We presented a research agenda on (1) quantifying the misalignment and recognizing when it is problematic, and (2) mechanisms to guide the STS toward equilibrium, where values and norms are aligned. Our agenda, while ambitious, is pragmatic. Whereas engineering normative and ethical MAS is at the core of our agenda, our vision builds on several foundational areas of AAMAS, including representation (Challenge 1), multi-objective search and optimization (Challenge 2), evolutionary game theory, mechanism design, and simulation (Challenge 3), multiagent learning (Challenge 4), argumentation and computational social choice (Challenge 5), and human-agent interaction (Challenge 6).

## ACKNOWLEDGMENTS

This work was initiated at Dagstuhl Seminar 25271 "Policy Modeling and Reasoning in Sociotechnical Systems." https://www.dagstuhl.de/25271. MPS thanks US NSF grant IIS-2116751. NA thanks UKRI EPSRC Grant No. EP/Y028392/1: AI for Collective Intelligence (AI4CI). LGN thanks ANR-FAPESP grant ANR-22-CE23-0018-01, FAPESP 2022/03454-1. MDV thanks UKRI Grant No.: EP/S023437/1: Centre for Doctoral Training in Accountable, Responsible and Transparent AI (ART-AI). PKM and DDA thank the Hybrid Intelligence Center (https://www.hybrid-intelligence-centre.nl/), a 10-year programme funded by the Dutch Ministry of Education, Culture and Science through the Netherlands Organisation for Scientific Research, under Grant No. (024.004.022).

## REFERENCES


[1] Rishabh Agrawal, Nirav Ajmeri, and Munindar P. Singh. 2022. Socially Intelligent Genetic Agents for the Emergence of Explicit Norms. In *Proceedings of the 31st International Joint Conference on Artificial Intelligence (IJCAI)*. IJCAI, Vienna, 10–14.

[2] Stéphane Airiau, Sandip Sen, and Daniel Villatoro. 2014. Emergence of conventions through social learning: Heterogeneous learners in complex networks. *Autonomous Agents and Multi-Agent Systems* 28 (2014), 779–804.

[3] 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

[4] 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, 16–24. https://doi.org/10.5555/3398761.3398769

[5] Giulia Andrighetto, Marco Campenni, Federico Cecconi, and Rosaria Conte. 2010. The Complex Loop of Norm Emergence: A Simulation Model. In *Simulating Interacting Agents and Social Phenomena*. Springer, New York, 19–35.

[6] Giulia Andrighetto, Sergey Gavrilets, Michele Gelfand, Ruth Mace, and Eva Vriens. 2024. Social norm change: drivers and consequences. , 20230023 pages.

[7] Giulia Andrighetto, Guido Governatori, Pablo Noriega, and Leendert W. N. van der Torre (Eds.). 2013. *Normative Multi-Agent Systems*. Number 4 in Dagstuhl Follow-Ups. Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik, Wadern, Germany. http://drops.dagstuhl.de/opus/portals/dfu/index.php?semnr=13003

[8] Thomas Arnold, Gordon Briggs, and Matthias Scheutz. 2022. Only Those Who Can Obey Can Disobey: The Intentional Implications of Artificial Agent Disobedience. In *Autonomous Agents and Multiagent Systems. Best and Visionary Papers: AAMAS 2022 Workshops, Virtual Event, May 9–13, 2022, Revised Selected Papers*. Springer-Verlag, Berlin, Heidelberg, 130–143. https://doi.org/10.1007/978-3-031-20179-0_9

[9] Alexander Artikis. 2009. Dynamic Protocols for Open Agent Systems. In *Proceedings of the 8th International Conference on Autonomous Agents and Multiagent Systems (AAMAS)*. International Foundation for Autonomous Agents and Multiagent Systems, Budapest, 97–104.

[10] Duangtida Athakravi, Domenico Corapi, Alessandra Russo, Marina De Vos, Julian Padget, and Ken Satoh. 2012. Handling change in normative specifications. In *International Workshop on Declarative Agent Languages and Technologies*. Springer, Valencia, 1–19.

[11] Emmanuelle Auriol and Stefanie Brilon. 2010. The good, the bad, and the ordinary: Anti-social behavior in profit and non-profit organizations. In *Proceedings of the German Development Economics Conference*. Verein für Socialpolitik, Ausschuss für Entwicklungsländer, Göttingen, Hannover, 1–41.

[12] Robert Axelrod. 1986. An evolutionary approach to norms. *American political science review* 80, 4 (1986), 1095–1111.

[13] Cristina Bicchieri. 2005. *The grammar of society: The nature and dynamics of social norms*. Cambridge University Press, Cambridge.

[14] Charles E. Bidwell. 1966. Values, norms, and the integration of complex social systems. *The Sociological Quarterly* 7, 2 (1966), 119–136.

[15] Guido Boella, Leendert Van Der Torre, and Harko Verhagen. 2006. Introduction to normative multiagent systems. *Computational & Mathematical Organization Theory* 12, 2 (2006), 71–79.

[16] Eva Bou, Maite López-Sánchez, and Juan A. Rodríguez-Aguilar. 2006. Adaptation of Autonomic Electronic Institutions Through Norms and Institutional Agents. In *Proceedings of the 7th International Workshop Engineering Societies in the Agents World (ESAW) (Lecture Notes in Computer Science, Vol. 4457)*, Gregory M. P. O'Hare, Alessandro Ricci, Michael J. O'Grady, and Oguz Dikenelli (Eds.). Springer, Dublin, 300–319. https://doi.org/10.1007/978-3-540-75524-1_17

[17] Eva Bou, Maite López-Sánchez, Juan A. Rodríguez-Aguilar, and Jaime Simão Sichman. 2008. Adapting Autonomic Electronic Institutions to Heterogeneous Agent Societies. In *Proceedings of the 1st International Workshop on Organized Adaption in Multi-Agent Systems (OAMAS) (Lecture Notes in Computer Science, Vol. 5368)*, George A. Vouros, Alexander Artikis, Kostas Stathis, and Jeremy V. Pitt (Eds.). Springer, Estoril, Portugal, 18–35. https://doi.org/10.1007/978-3-642-02377-4_2

[18] Jan M. Broersen, Mehdi Dastani, Joris Hulstijn, Zhisheng Huang, and Leendert W. N. van der Torre. 2001. The BOID architecture: conflicts between beliefs, obligations, intentions and desires. In *Proceedings of the 5th International Conference on Autonomous Agents (AGENTS)*, Elisabeth André, Sandip Sen, Claude Frasson, and Jörg P. Müller (Eds.). ACM, Montreal, 9–16. https://doi.org/10.1145/375735.375766

[19] Jan M. Broersen, Mehdi Dastani, and Leendert W. N. van der Torre. 2001. Resolving Conflicts between Beliefs, Obligations, Intentions, and Desires. In *Proceedings of the 6th European Conference on Symbolic and Quantitative Approaches to Reasoning with Uncertainty (ECSQARU) (Lecture Notes in Computer Science, Vol. 2143)*, Salem Benferhat and Philippe Besnard (Eds.). Springer, Toulouse, 568–579. https://doi.org/10.1007/3-540-44652-4_50

[20] Jordi Campos, Maite López-Sánchez, Maria Salamó, Pedro Avila, and Juan A. Rodríguez-Aguilar. 2013. Robust Regulation Adaptation in Multi-Agent Systems. *ACM Transactions on Autonomous and Adaptive Systems* 8, 3 (2013), 13:1–13:27. https://doi.org/10.1145/2517328

[21] Henrique Lopes Cardoso and Eugénio C. Oliveira. 2009. Adaptive Deterrence Sanctions in a Normative Framework. In *Proceedings of the 2009 IEEE/WIC/ACM International Conference on Intelligent Agent Technology (IAT)*. IEEE Computer Society, Milan, 36–43. https://doi.org/10.1109/WI-IAT.2009.123

[22] Cristiano Castelfranchi. 2016. A Cognitive Framing for Norm Change. In *Proceedings of the International Workshop on Coordination, Organizations, Institutions, and Norms in Agent Systems (COIN@AAMAS/IJCAI'15) (Lecture Notes in Computer Science, Vol. 9628)*, Virginia Dignum, Pablo Noriega, Murat Sensoy, and Jaime Simão Sichman (Eds.). Springer, Buenos Aires, 22–41. https://doi.org/10.1007/978-3-319-42691-4_2

[23] Juan C. Castilla-Rho, Rodrigo Rojas, Martin S. Andersen, Cameron Holley, and Gregoire Mariethoz. 2017. Social tipping points in global groundwater management. *Nature Human Behaviour* 1 (2017), 640–649. https://doi.org/10.1038/s41562-017-0181-7

[24] Roberto Centeno, Holger Billhardt, and Ramón Hermoso. 2011. An Adaptive Sanctioning Mechanism for Open Multi-agent Systems Regulated by Norms. In *Proceedings of the 23rd IEEE International Conference on Tools with Artificial Intelligence*. IEEE Computer Society, Boca Raton, FL, 523–530. https://doi.org/10.1109/ICTAI.2011.85

[25] Amit Chopra, Leendert van der Torre, Harko Verhagen, and Serena Villata. 2018. *Handbook of normative multiagent systems*. College Publications, London.

[26] Rosaria Conte, Giulia Andrighetto, and Marco Campennì (Eds.). 2013. *Minding Norms: Mechanisms and dynamics of social order in agent societies*. Oxford University Press, Oxford.

[27] Rosaria Conte, Cristiano Castelfranchi, and Frank Dignum. 1998. Autonomous Norm Acceptance. In *Intelligent Agents V, Agent Theories, Architectures, and Languages, 5th International Workshop, ATAL '98, Paris, France, July 4-7, 1998, Proceedings (Lecture Notes in Computer Science, Vol. 1555)*, Jörg P. Müller, Munindar P. Singh, and Anand S. Rao (Eds.). Springer, Berlin, Heidelberg, 99–112. https://doi.org/10.1007/3-540-49057-4_7

[28] Domenico Corapi, Alessandra Russo, Marina De Vos, Julian Padget, and Ken Satoh. 2011. Normative design using inductive learning. *Theory and Practice of Logic Programming* 11, 4-5 (2011), 783–799.

[29] Stephen Cranefield, Felipe Meneguzzi, Nir Oren, and Bastin Tony Roy Savarimuthu. 2016. A Bayesian Approach to Norm Identification. In *Proceedings of the 22nd European Conference on Artificial Intelligence (ECAI) (Frontiers in Artificial Intelligence and Applications, Vol. 285)*, Gal A. Kaminka, Maria Fox, Paolo Bouquet, Eyke Hüllermeier, Virginia Dignum, Frank Dignum, and Frank van Harmelen (Eds.). IOS Press, The Hague, 622–629. https://doi.org/10.3233/978-1-61499-672-9-622

[30] Karen da Silva Figueiredo and Viviane Torres da Silva. 2013. An Algorithm to Identify Conflicts Between Norms and Values. In *Proceedings of the 2013 International Workshop on Coordination, Organizations, Institutions, and Norms in Agent Systems (COIN@AAMAS/PRIMA) (Lecture Notes in Computer Science, Vol. 8386)*, Tina Balke, Frank Dignum, M. Birna van Riemsdijk, and Amit K. Chopra (Eds.). Springer, St. Paul, MN, 259–274. https://doi.org/10.1007/978-3-319-07314-9_14

[31] Mehdi Dastani, John-Jules Ch. Meyer, and Nick A. M. Tinnemeier. 2012. Programming norm change. *Journal of Applied Non-Classical Logics* 22, 1-2 (2012), 151–180. https://doi.org/10.1080/11663081.2012.682784

[32] Francien Dechesne, Gennaro Di Tosto, Virginia Dignum, and Frank Dignum. 2013. No smoking here: values, norms and culture in multi-agent systems. *Artificial intelligence and law* 21 (2013), 79–107.

[33] Davide Dell'Anna, Natasha Alechina, Fabiano Dalpiaz, Mehdi Dastani, and Brian Logan. 2022. Data-driven revision of conditional norms in multi-agent systems.

*Journal of Artificial Intelligence Research* 75 (2022), 1549–1593.

[34] Davide Dell'Anna, Mehdi Dastani, and Fabiano Dalpiaz. 2020. Runtime revision of sanctions in normative multi-agent systems. *Autonomous Agents and Multi-Agent Systems* 34 (2020), 1–54.

[35] Thiago Freitas dos Santos, Nardine Osman, and Marco Schorlemmer. 2024. Is this a violation? Learning and understanding norm violations in online communities. *Artificial Intelligence* 327 (2024), 104058.

[36] Ernst Fehr and Urs Fischbacher. 2003. The Nature of Human Altruism. *Nature* 425 (2003), 785–791. https://doi.org/10.1038/nature02043

[37] Yibin Feng, Tianqi Song, Yugin Tan, Zicheng Zhu, and Yi-Chieh Lee. 2025. Multi-Agent Systems Shape Social Norms for Prosocial Behavior Change. In *Companion Publication of the 2025 Conference on Computer-Supported Cooperative Work and Social Computing (CSCW Companion '25)*. Association for Computing Machinery, Bergen, Norway, 320–325. https://doi.org/10.1145/3715070.3749246

[38] Christopher K. Frantz and Gabriella Pigozzi. 2018. Modelling norm dynamics in multi-agent systems. In *Handbook of Normative Multiagent Systems*. College Publications, Rickmansworth, 73–141.

[39] Sergey Gavrilets, Denis Tverskoi, Nianyi Wang, Xiaomin Wang, Juan Ozaita, Boyu Zhang, Angel Sánchez, and Giulia Andrighetto. 2024. Co-evolution of behaviour and beliefs in social dilemmas: estimating material, social, cognitive and cultural determinants. *Evolutionary Human Sciences* 6 (2024), e50.

[40] Guido Governatori and Antonino Rotolo. 2010. Changing legal systems: legal abrogations and annulments in Defeasible Logic. *Log. J. IGPL* 18, 1 (2010), 157–194. https://doi.org/10.1093/JIGPAL/JZP075

[41] Samaneh Heidari. 2022. *Agents with Social Norms and Values: A framework for agent based social simulations with social norms and personal values*. Ph.D. Dissertation. Utrecht University.

[42] Samaneh Heidari, Maarten Jensen, and Frank Dignum. 2020. Simulations with values. In *Proceedings of the 14th Social Simulation Conference*. Springer, Stockholm, 201–215.

[43] Samaneh Heidari, Nanda Wijermans, and Frank Dignum. 2019. Agents with Dynamic Social Norms. In *Proceedings of the 20th International Workshop on Multi-Agent-Based Simulation (MABS) (Lecture Notes in Computer Science, Vol. 12025)*, Mario Paolucci, Jaime Simão Sichman, and Harko Verhagen (Eds.). Springer, Montreal, 112–124. https://doi.org/10.1007/978-3-030-60843-9_9

[44] Philipp M. Hetzer. 2011. *The evolution of fairness preferences, altruistic punishment, and cooperation*. Ph.D. Dissertation. ETH Zurich.

[45] Christopher D. Hollander and Annie S. Wu. 2011. The Current State of Normative Agent-Based Systems. *Journal of Artificial Societies and Social Simulation* 14, 2 (2011), 1–24. https://doi.org/10.18564/JASSS.1750

[46] Jiaming Ji, Tianyi Qiu, Boyuan Chen, Jiayi Zhou, Borong Zhang, Donghai Hong, Hantao Lou, Kaile Wang, Yawen Duan, Zhonghao He, Lukas Vierling, Zhaowei Zhang, Fanzhi Zeng, Juntao Dai, Xuehai Pan, Hua Xu, Aidan O'Gara, Kwan Ng, Brian Tse, Jie Fu, Stephen Mcaleer, Yanfeng Wang, Mingchuan Yang, Yunhuai Liu, Yizhou Wang, Song-Chun Zhu, Yike Guo, Yaodong Yang, and Wen Gao. 2025. AI Alignment: A Contemporary Survey. *Comput. Surveys* 58, 5, Article 132 (Nov. 2025), 38 pages. https://doi.org/10.1145/3770749

[47] Martin J. Kollingbaum and Timothy J. Norman. 2003. Norm Adoption and Consistency in the NoA Agent Architecture. In *Proceedings of the 1st International Workshop on Programming Multi-Agent Systems (PROMAS) (Lecture Notes in Computer Science, Vol. 3067)*, Mehdi Dastani, Jürgen Dix, and Amal El Fallah Seghrouchni (Eds.). Springer, Melbourne, 169–186. https://doi.org/10.1007/978-3-540-25936-7_9

[48] Roger X. Lera-Leri, Enrico Liscio, Filippo Bistaffa, Catholijn M. Jonker, Maite Lopez-Sanchez, Pradeep K. Murukannaiah, Juan A. Rodriguez-Aguilar, and Francisco Salas-Molina. 2024. Aggregating value systems for decision support. *Knowledge-Based Systems* 287 (2024), 111453. https://doi.org/10.1016/j.knosys.2024.111453

[49] Peter R Lewis, Harry Goldingay, and Vivek Nallur. 2014. It's good to be different: Diversity, heterogeneity, and dynamics in collective systems. In *IEEE Eighth International Conference on Self-adaptive and Self-organizing Systems Workshops (SASOW)*. IEEE, London, 84–89. Type: Conference proceedings.

[50] Enrico Liscio, Roger Lera-Leri, Filippo Bistaffa, Roel I.J. Dobbe, Catholijn M. Jonker, Maite Lopez-Sanchez, Juan A. Rodriguez-Aguilar, and Pradeep K. Murukannaiah. 2023. Value Inference in Sociotechnical Systems. In *Proceedings of the 2023 International Conference on Autonomous Agents and Multiagent Systems (AAMAS '23)*. IFAAMAS, London, 1774–1780.

[51] Enrico Liscio, Luciano C. Siebert, Catholijn M. Jonker, and Pradeep K. Murukannaiah. 2025. Value Preferences Estimation and Disambiguation in Hybrid Participatory Systems. *Journal of Artificial Intelligence Research* 82 (April 2025), 32. https://doi.org/10.1613/jair.1.14958

[52] Enrico Liscio, Michiel van der Meer, Luciano C. Siebert, Catholijn M. Jonker, and Pradeep K. Murukannaiah. 2022. What Values should an Agent Align with? An Empirical Comparison of General and Context-Specific Values. *Autonomous Agents and Multi-Agent Systems* 36, 1 (2022), 23. https://doi.org/10.1007/s10458-022-09550-0

[53] Moamin A. Mahmoud, Mohd Sharifuddin Ahmad, Mohd Zaliman M. Yusoff, and Salama A. Mostafa. 2018. A Regulative Norms Mining Algorithm for Complex Adaptive System. In *Proceedings of the 3rd International Conference on Soft Computing and Data Mining (SCDM) (Advances in Intelligent Systems and Computing, Vol. 700)*, Rozaida Ghazali, Mustafa Mat Deris, Nazri Mohd Nawi, and Jemal H. Abawajy (Eds.). Springer, Johor, Malaysia, 213–224. https://doi.org/10.1007/978-3-319-72550-5_21

[54] Mehdi Mashayekhi, Nirav Ajmeri, George F. List, and Munindar P. Singh. 2022. Prosocial Norm Emergence in Multiagent Systems. *ACM Transactions on Autonomous and Adaptive Systems (TAAS)* 17, 1–2 (June 2022), 3:1–3:24. https://doi.org/10.1145/3540202

[55] Jack McKinlay, Marina De Vos, Janina A. Hoffmann, and Andreas Theodorou. 2025. Understanding the Process of Human-AI Value Alignment. arXiv:2509.13854 [cs.CY] https://arxiv.org/abs/2509.13854

[56] Andreasa Morris-Martin, Marina De Vos, and Julian Padget. 2019. Norm Emergence in Multiagent Systems: A Viewpoint Paper. *Autonomous Agents and Multi-Agent Systems (JAAMAS)* 33, 6 (2019), 706–749.

[57] Andreasa Morris-Martin, Marina De Vos, Julian A. Padget, and Oliver Ray. 2023. Agent-directed Runtime Norm Synthesis. In *Proceedings of the 22nd International Conference on Autonomous Agents and Multiagent Systems (AAMAS)*, Noa Agmon, Bo An, Alessandro Ricci, and William Yeoh (Eds.). ACM, London, 2271–2279. https://doi.org/10.5555/3545946.3598905

[58] Pradeep K. Murukannaiah, Nirav Ajmeri, Catholijn M. Jonker, and Munindar P. Singh. 2020. New Foundations of Ethical Multiagent Systems. In *Proceedings of the 19th International Conference on Autonomous Agents and Multiagent Systems (AAMAS)*. IFAAMAS, Auckland, 1706–1710. https://doi.org/10.5555/3398761.3398958 Blue Sky Ideas Track.

[59] Pradeep K. Murukannaiah and Munindar P. Singh. 2020. From Machine Ethics to Internet Ethics: Broadening the Horizon. *IEEE Internet Computing* 24, 3 (2020), 51–57. https://doi.org/10.1109/MIC.2020.2989935

[60] Luis G. Nardin, Tina Balke-Visser, Nirav Ajmeri, Anup K. Kalia, Jaime S. Sichman, and Munindar P. Singh. 2016. Classifying sanctions and designing a conceptual sanctioning process model for socio-technical systems. *The Knowledge Engineering Review* 31, 2 (2016), 142–166.

[61] Eamonn O'Toole, Vivek Nallur, and Siobhán Clarke. 2017. Decentralised detection of emergence in complex adaptive systems. *ACM Transactions on Autonomous and Adaptive Systems* 12, 1 (2017), 1–31. https://doi.org/10.1145/3019597 Type: Journal article.

[62] Ana Paiva, Filipa Correia, Raquel Oliveira, Fernando Santos, and Patrícia Arriaga. 2021. Empathy and Prosociality in Social Agents. In *The Handbook on Socially Interactive Agents: 20 Years of Research on Embodied Conversational Agents, Intelligent Virtual Agents, and Social Robotics Volume 1: Methods, Behavior, Cognition* (1 ed.). ACM, New York, 385–432. https://doi.org/10.1145/3477322.3477334

[63] Stuart Russell. 2022. Artificial Intelligence and the Problem of Control. *Perspectives on digital humanism* 19 (2022), 1–322.

[64] Fernando P. Santos. 2024. Prosocial dynamics in multiagent systems. *AI Magazine* 45, 1 (2024), 131–138. https://doi.org/10.1002/aaai.12143

[65] Bastin Tony Roy Savarimuthu and Stephen Cranefield. 2011. Norm creation, spreading and emergence: A survey of simulation models of norms in multi-agent systems. *Multiagent and Grid Systems* 7, 1 (2011), 21–54.

[66] Shalom H Schwartz. 1977. Normative influences on altruism. In *Advances in experimental social psychology*. Vol. 10. Elsevier, Cambridge, MA, 221–279.

[67] Shalom H. Schwartz and Wolfgang Bilsky. 1987. Toward a universal psychological structure of human values. *Journal of personality and social psychology* 53, 3 (1987), 550.

[68] Sandip Sen and Stéphane Airiau. 2007. Emergence of norms through social learning. In *Proceedings of the 20th International Joint Conference on Artifical Intelligence (IJCAI'07)*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1507–1512.

[69] Amika M. Singh and Munindar P. Singh. 2023. Norm Deviation in Multiagent Systems: A Foundation for Responsible Autonomy. In *Proceedings of the 32nd International Joint Conference on Artificial Intelligence (IJCAI)*. IJCAI, Macau, 289–297. https://doi.org/10.24963/ijcai.2023/33

[70] Munindar P. Singh. 2013. Norms As a Basis for Governing Sociotechnical Systems. *ACM Transactions on Intelligent Systems and Technology (TIST)* 5, 1, Article 21 (Dec. 2013), 23 pages.

[71] Divya Sundaresan, Akhira Watson, Eleni Bardaka, Crystal Chen Lee, Christopher B. Mayhorn, and Munindar P. Singh. 2025. Prosociality in Microtransit. *Journal of Artificial Intelligence Research (JAIR)* 82 (Jan. 2025), 77–110. https://doi.org/10.1613/jair.1.16777

[72] Sz-Ting Tzeng, Nirav Ajmeri, and Munindar P. Singh. 2024. Norm Enforcement with a Soft Touch: Faster Emergence, Happier Agents. In *Proceedings of the 23rd International Conference on Autonomous Agents and Multiagent Systems (AAMAS)*. IFAAMAS, Auckland, 1837–1846. https://doi.org/10.5555/3635637.3663046

[73] Edna Ullmann-Margalit. 1990. Revision of norms. *Ethics* 100, 4 (1990), 756–767. http://www.jstor.org/stable/2381777

[74] Till Winkler and Sarah Spiekermann. 2021. Twenty years of value sensitive design: a review of methodological practices in VSD projects. *Ethics and Information Technology* 23, 1 (March 2021), 17–21. https://doi.org/10.1007/s10676-018-9476-2

[75] Jessica Woodgate and Nirav Ajmeri. 2025. Combining Normative Ethics Principles to Learn Prosocial Behaviour. In *Proceedings of the 24th International Conference on Autonomous Agents and Multiagent Systems (AAMAS)*. IFAAMAS, Detroit, 2789–2791. https://doi.org/10.5555/3709347.3744013

[76] Jessica Woodgate, Paul Marshall, and Nirav Ajmeri. 2025. Operationalising Rawlsian Ethics for Fairness in Norm-Learning Agents. In *Proceedings of the 39th AAAI Conference on Artificial Intelligence (AAAI)*. AAAI, Philadelphia, 2382–26390. https://doi.org/10.1609/aaai.v39i25.34837

[77] Elena Yan, Luis G. Nardin, O. Boissier, and Jaime S. Sichman. 2025. A regulation adaptation model for multi-agent systems. In *Proceedings of the 28th European Conference on Artificial Intelligence (ECAI)*, Inês Lynce, Nello Murano, Mauro Vallati, Serena Villata, Federico Chesani, Michela Milano, Andrea Omicini, and Mehdi Dastani (Eds.). IOS Press, Bologna, 3671–3678. https://doi.org/10.3233/FAIA251245

[78] Elena Yan, Luis G. Nardin, Olivier Boissier, and Jaime S. Sichman. 2026. A unified view on regulation management in multi-agent systems. In *Coordination, Organizations, Institutions, Norms, and Ethics for Governance of Multi-Agent Systems XVIII (Lecture Notes in Computer Science, Vol. 16253)*, Sz-Ting Tzeng, Davide Dell'Anna, and Jaime S. Sichman (Eds.). Springer Nature Switzerland, Cham, 55–74. https://doi.org/10.1007/978-3-032-17542-7_4