

Characterizing Artificial Socio-Cognitive Technical Systems

Rob Christiaan¹, Aditya Ghose², Pablo Noriega³, and Munindar P. Singh⁴

¹ Vrije Universiteit Amsterdam; The Netherlands

² University of Wollongong; Australia

³ IIIA-CSIC; Spain

⁴ North Carolina State University; USA

Abstract. This paper is an invitation to examine a class of socio-technical systems—artificial socio-cognitive (ASCS)—whose distinctive nature is that they may involve humans as well as artificial agents who interact in a regulated milieu. We propose a characterization of these ASCS and build on that characterization to describe how these systems evolve.

1 Motivation

In recent years we have witnessed the appearance of several socio-technical systems like Facebook, eBay, Amazon Turk that have produced significant changes in the way everyday social coordination takes place. Changes that involve not only new types of coordination but, as is evident in the first two cases, coordination at a massive global extent. While these three examples may be paradigmatic of the highly visible and successful systems, the phenomenon includes a large amount of systems that share a number of cognate features and have a similar or potentially similar social and technological impact.

We believe it is worth taking a systematic look at these systems. This paper outlines a modest contribution in that direction. We point towards a characterization of these systems by introducing some terminological distinctions and an abstract descriptive framework. We, then outline the key elements of a particular aspect that we believe has not received yet systematic attention: the process through which these systems are created or updated.

Our interest is about socio-technical systems (STS) but will limit our scope to a particular kind that we call “artificial socio-cognitive systems” or ASCS. These matters were first discussed in [10] and developed in [11]. To illustrate some ideas we shall refer to electronic commerce and public health care, and use the system presented in [9] for more specific references.

2 A vocabulary for characterization

Socio-cognitive systems come in many different forms. Most are built around an IT platform that offers coordination capabilities, but the intents, structures and functionalities are typically widely divergent. Within the ambit of socio-cognitive systems, one would

find systems as diverse as the ecologies built around social media platforms such as Facebook, Twitter or LinkedIn, B2C trading platforms such as eBay, B2B platforms such as Ariba or NASDAQ, crowdsourcing applications like Ushahidi or the Amazon Turk, mixed-level participatory social simulation, multiplayer online games, public health systems, or military command and control systems. There are also unlikely instances such as a flashmob or a criminal syndicate that leverages an IT infrastructure for coordination.

For researchers interested in developing a better understanding of such systems, it is useful to explore what is common to these systems, and how they differ. A careful characterization of such systems using a common set of principles, or a common vocabulary can help. There is a long history of similar, principled approaches to characterize broad classes of systems. Kenneth Arrow's [3] seminal work on characterizing social decision processes (an equally broad and variegated class as the one of interest here) is perhaps the earliest example. Arrow offered a set of postulates that formalized what are arguably intuitive requirements for social decision processes (specifically preference aggregation functions) and went on to establish the well-known result that no function existed that satisfied all of these properties. A similar exercise was undertaken by Alchourron, Gärdenfors and Makinson [1] to systematize what was at that time a highly varied class of belief revision operators.

Our intent here is similar. What we present below represents the first steps towards a comprehensive set of properties for characterizing the systems of interest. We will distinguish between two classes of properties: *definitional* and *architectural*. Definitional properties will offer a vocabulary for discriminating artificial socio-cognitive systems from other socio-technical systems. We shall use architectural properties to characterize sub-classes of such systems.

Definitional properties:

Broadly speaking, our aim is to study systems that involve several rational participants who come together to perform a collective activity that they cannot accomplish on their own and such action does not occur directly between individuals but is mediated by technological artefacts. The following properties are a first attempt towards a top-down abstract characterisation of ASCS:

- *System* An artificial socio-cognitive system is composed by two (“first class”) entities: a *social space* and the *agents* who act within that space. The system exists in the real world and there is a boundary that determines what is inside the system and what is out.
- *Agents* Agents are entities who are capable of acting within the social space. They exhibit the following characteristics:
 - *Socio-cognitive* Agents are presumed to base their actions on some internal decision model. The decision-making behaviour of agents, in principle, takes into account social aspects because the actions of agents may be affected by the social space or other agents and may affect other agents and the space itself [4].
 - *Opaque* The system, in principle, has no access to the decision-making models, or internal states of participating agents.

- *Mixed* Agents may be human or software entities (we'll simply call them "agents" or "participants where it is not necessary to distinguish).
 - *Heterogeneous* Agents may have different decision models, different motivations and respond to different principals.
 - *Autonomous* Agents are self-motivated, not necessarily competent or benevolent, hence they may fail to act as expected or demanded of them.
- *Social space*. This is the environment or milieu where agent interactions take place. It should have the following properties:
- *Open* Agents may enter and leave the social space and *a priori*, it is not known (by the system or other agents) which agents may be active at a given time, nor whether new agents will join or leave at some point or not.
 - *Perceivable* All interactions and events within the shared social space are mediated by technological artefacts—that is, as far as the system is concerned there are no direct interactions between agents outside the system and only those actions that are mediated by a technological artefact that is part of the system may have effects in the system—and although they might be described in terms of the five senses, they can collectively be considered percepts.
 - *Constrained* In order to coordinate actions, the space includes (and governs) regulations, obligations, norms or conventions that agents are in principle supposed to follow.
 - *Persistent* The social space may change over time in two ways: either by events that happen while the systems is enacted and by the actions of agents; or through some system functionalities that are part of the system design and are triggered while the system is enacted.⁵

We see these systems as *socio-technical* systems because of the participation of humans and computational components [14], although they are better understood in the sense of [13] where software agents may also be involved. The term *artificial* is used to evoke the existence of some external design of the system and the term *socio-cognitive* to suggest that in order to characterise or deploy them we need to “ ‘understand’ and reproduce features of the human social mind like commitments, norms, mind reading, power, trust, ‘institutional effects’ and social macro-phenomena” [4]. Because of the assumption of intrinsic constraint on agent interactions, the above assumptions characterise a type of *normative multiagent system* [2].

Architectural properties:

These are properties that one would expect that all ASCS should, However have they would be achieved with different means and expressed in different ways and degrees. Thus we want to make them explicit for two main reasons, one is help us to classify systems into classes and get hold of appropriate expressive and functional means for modelling or deploying particular classes.

⁵ Persistence, is a matter of convention. As Sec.4 illustrates, there may be situations when stakeholders may decide to change an active system (“the system as is”) into a new one (the “system to be”) that is different enough to deserve that exogenous intervention and labelling. While many features of the old system (including agents and commitments) “persist” in the new one, the old system itself is acknowledged to end.

- *Information decentralization*: The extent to which information is decentralized varies across these systems. An email system would represent full decentralization (discounting settings where the email network manager obtains privileged access to all emails exchanged). In a public health system, patients own data that pertains to them, but clinicians have access to the records of all patients that they attend to, while administrators and insurers have access to even larger classes of medical records. A social media platform such as Facebook provides limited (policy-regulated) access to data for specific users, while the providers of the platform have privileged access to all data from all users. Information decentralization in military command-and-control session would be similar.
- *Governance/control*: Systems vary in the extent of autonomy afforded to constituent agents. An email system offers considerable autonomy, as would a flashmob. A B2B market offers a slightly lower level of autonomy, by requiring participants to conform to a set of market rules. A military command-and-control system would traditionally offer very little autonomy to the personnel under its command, but modern “network-centric warfare” offers considerably greater autonomy.
- *Fluidity of norms*: A flashmob or an email system involves a minimal set of norms, but these are also relatively static. The market rules governing B2B or B2B market providers also tend to change relatively infrequently. Social media platforms frequently revise norms (specially those governing privacy, reacting to shifting user perceptions). Many online multi-player games support configurable games-for these the norms are clearly highly fluid.
- *Transparency*: Market providers (specially B2B providers) are obliged to be very transparent in terms of the information available and the norms governing their behaviour (stock markets, for instance, need to comply with stringent transparency requirements imposed by market regulators). Public health systems are not always required to be transparent, both with the norms that govern their operations and the information they retain. Patient information clearly cannot be shared, but public health systems are often reluctant to share operational data for fear of being shown to be inefficient. Military command-and-control systems are by definition non-transparent.
- *Accountability*: Public health systems and stock markets are typically held to very high standards of accountability. A social media platform might be required to be accountable to some degree in the event of privacy breaches or other adverse events. On the other extreme, an email system or a flashmob represent examples of systems with very lax accountability requirements.
- *Nature of identity*: A stock market (or other B2B market providers), a public health system and a military command-and-control system would insist on the true identity of each participating agent. A social media platform might permit the same agent to assume multiple identities. A flashmob would, in general, be not particularly concerned with the identities of the participating agents.

3 The WIT framework

From the definitional properties mentioned above, one may see ASCS as systems where it is possible to *govern* the interaction of agents that are situated in a physical or artifi-

cial “world” by means of technological artefacts. The key element, which is not usually included in other accounts of socio-technical systems, is the “governance” or “institutional” part that mediates between the “world” and the technological artefacts. The realization that one needs to account for the relationships between the institutional aspects of an ASCS and its associated technological and working aspects, motivates an abstract characterization of an ASCS from the point of view of each of those aspects. The relationships between these three components is explained in the following “notion” and illustrated in Fig. 1.⁶

Notion 1 The WIT framework: *An artificial socio-cognitive system is composed by three interrelated elements:*

- View 1: The world system, \mathcal{W} , as the agents (both human and software) see it and relate to it.*
- View 2: An ideal institutional system, \mathcal{I} , that stipulates the way the system should behave.*
- View 3: The technological artefacts, \mathcal{T} , that implement the ideal system and run the applications that enable users to accomplish collective actions in the real world, \mathcal{W} , according to the rules set out in \mathcal{I} .*

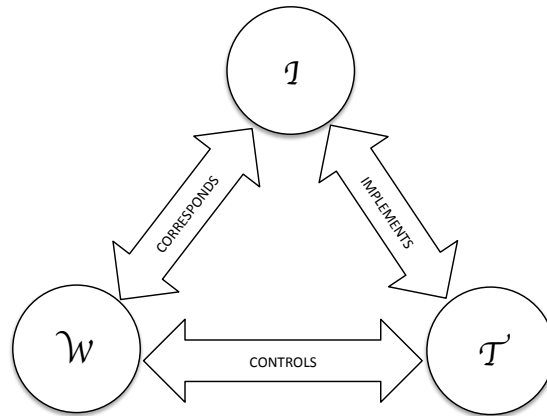


Fig. 1: The WIT trinity view of artificial socio-cognitive systems: The normative institutional system, \mathcal{I} ; the technological artefacts that implement it, \mathcal{T} , and the actual world where the system is used, \mathcal{W} . After [10].

These three views are interrelated through three binary relationships:

- *The institutional world corresponds with the real world through what is known as a “counts-as” relationship [12, 7] by which (brute) facts and (brute) actions*

⁶ Note that our discussion about innovation (Sec.4) shows how what gives rise to a new ASCS, is the formulation within the existing ASCS of the institutional part of that new ASCS.

in the real world correspond to institutional facts and actions in the institutional world \mathcal{I} provided these comply with the institutional conventions. Suppose Alice sells her property to Bob by signing off a deed in Bob's favour. The ownership of property is an institutional fact; its sale is an institutional action. The presence of ink of a certain pattern on a piece of paper is a brute fact; placing that ink with a pen is a brute action. The signed transfer deed counts as proof of ownership; signing the deed counts as a sale. A brute action creates (or revises) a brute fact; in appropriate circumstances, it counts as an institutional action and creates (or revises) institutional facts.

- *The conventions prescribed in the institutional world have their counterpart in the technological world in the sense that institutional conventions constitute a specification of the requirements of the system that is implemented in \mathcal{T} .*
- *The system, as implemented in \mathcal{T} , is what enables interactions (through a proper interface) in \mathcal{W} .*

It should be noted that each of these three binary relationships needs to satisfy certain integrity conditions:

- The *corresponds* relationship needs: (i) to guarantee that the objects and concepts involved in the descriptions and functioning in \mathcal{I} are properly associated with entities in \mathcal{W} ; i.e., that there is a bijection between terms in the languages in \mathcal{I} and objects and actions in \mathcal{W} . (ii) that the identity of agents in \mathcal{W} is properly reflected in their counterparts in \mathcal{I} and is preserved as long as the agents are active in the system, (iii) that the agents that participate in \mathcal{W} have the proper entitlements to be subject to the conventions that regulate their interactions and in particular to fulfil in \mathcal{W} those commitments that they establish in \mathcal{I} , and (iv) that the commitments that are established according to \mathcal{I} are properly reflected in \mathcal{W} .
- The *implements* relationship needs to be a faithful programming of the institutional conventions so that actions and effects are well programmed, norms are properly represented and enforced, etc.
- Finally, the *controls* relationship needs to make sure that: (i) the technological artefacts work properly (communication is not scrambled, data bases are not corrupted, etc.) and (ii) inputs and outputs are properly presented and captured in \mathcal{W} , according to the implementation of the corresponding processes in \mathcal{I} . (iii) Algorithms and data structures in \mathcal{T} behave as the conventions in \mathcal{I} prescribe.

This WIT framework is used in [11] to clarify how changes in the social space are accounted for in each of the three views. The framework is further used to discuss how the normative view in \mathcal{I} is mirrored in \mathcal{T} through an institutional specification language (in \mathcal{I}) that is programmed with data structures and operations that are implemented as software in \mathcal{T} .

4 Describing the innovation process

Next, we elaborate the abstract WIT structure above via an envisioned reference architecture and methodology. Specifically, consider the situation where some stakeholders

come together to create or modify a sociotechnical system. In what follows we shall presume that there is already an existing ASCS, \mathcal{S}_0 (the *STS-as-is*) and that those participants who are in its \mathcal{W} component are in the process of designing a new \mathcal{S}_1 (the *STS-to-be*) by specifying its \mathcal{I}_1 component (Fig. 2).

We assume that the stakeholders express some requirements that they would like the STS-to-be (adapting the terminology of requirements engineering [15]) to address. In general, the stakeholders may not express their requirements explicitly, but their requirements might be elicited through discussion or understood through (e.g., ethnographic) observation.

These requirements would necessarily be framed in terms of the institution to which the stakeholders belong (\mathcal{I}_0). Based on these requirements, the stakeholders would enact a potentially complex negotiation to create a specification for the STS-to-be. This specification in our terms would be a normative specification and thus serves as an abstract institution. That is, an STS-to-be does not initially include any principals or technical entities (i.e., resources or infrastructure). Principals, defined as socially autonomous entities, would decide whether to adopt various roles in this institution; their adoption would succeed provided they meet requirements such as the qualifications imposed by the institution. The principals would introduce the necessary technical entities, i.e., resources and infrastructure (\mathcal{T}_1). The idea here is that the technical entities needed for the STS-to-be must come from somewhere. In our stance on socio-technical systems, this means that there must be a principal behind each technical entity. In introducing the requisite technical entities, the principals would instantiate and make the institution concrete, thereby instantiating the STS-to-be.

Following Chopra et al. [6], we distinguish between stakeholders and principals. *Stakeholders* are social entities involved in the design process that begins from requirements elicitation and ends in the creation of a specification of an STS. *Principals* are social entities who adopt roles in an institution being instantiated to produce an STS. The principals are not only a changing set but frequently are drawn from a different population than the stakeholders. That is, principals may have somewhat different requirements from the (original) stakeholders, though the principals are able to adapt the given STS to meet their requirements.

The above process incorporates some subtleties, which we explain via examples. A simple imagined scenario is of e-commerce. People already trade goods and they live in a world where institutional concepts such as ownership, transfer of ownership, and money are defined and infrastructure is available for shipping and delivery, and payment. Suppose that some prospective buyers and sellers or even one imaginative person (as in the case of eBay's founding) comes up with requirements for buying and selling. These requirements are framed in terms of the stakeholders of e-commerce, buyers, sellers, market enabler, and banks. The stakeholders could all be identified from the start or a few of them (e.g., buyers, sellers, and market enabler) might begin interacting and then recruit additional stakeholders (e.g., banks). The stakeholders figure out a solution that would support e-commerce via auctions. This solution is a normative specification of the STS-to-be. Let us say the market enabler (e.g., eBay) adopts the key unique role in this STS-to-be and provides the technical infrastructure. One or more banks join in for payment processing. Gradually the other roles are adopted by other principals, and

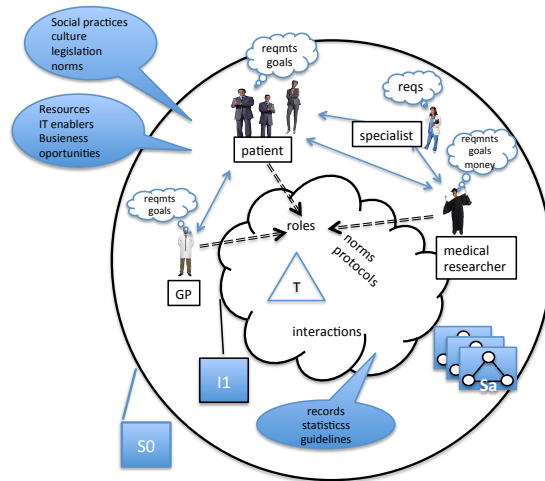


Fig. 2: The STS-as-is (\mathcal{S}_0) and the definition of the institutional component (\mathcal{I}_1) of the STS-to-be.

the STS is fully instantiated. This STS carries out its specific interactions, e.g., placing a bid and determining a winner, according to its normative specification. The world outside remains present, e.g., to provide a venue for sanctioning via lawsuit if one of the principals claims another principal to have violated some norm.

An example from the health care domain is where the stakeholders are clinical researchers, general practitioners (GPs), and patients, who wish to promote medical research. They would work within the institution of the existing health care STS, wherein concepts such as physician and patient are defined along with relationships such as treats and advises. In that STS, the stakeholders would identify a specification of an STS-to-be in which patients are recruited to participate in various medical studies.

In typical cases, the STS-to-be would not be created from scratch but would be a modification of an existing STS. For example, if clinical researchers want to conduct a new category of medical studies where patients carry sensors on their bodies or medical studies that engage families of patients, they would invite physicians as additional stakeholders and design new interactions whereby a patient may be recruited to wear a wrist band or a patient and the patient's spouse may both be recruited to survey them on their family's diet.

Summing up the above vision of requirements modeling and realization, we see that the process begins from an institution that functions as the social substrate or world for an STS-to-be. In this view, the existing world is indeed the STS-as-is [15]. The process produces a normative specification of an STS-to-be, which upon instantiation becomes the STS-to-be. The STS-to-be need not cause any structural changes to the STS-as-is. For example, commerce remains defined and appropriate in the old world even after

eBay, and the health care system continues to function despite the introduction of a clinical trials STS. However, if the requirements addressed by the new STS-to-be are significant, successful operation of the STS-to-be would affect the STS-as-is. For example, less commerce may occur in person when e-commerce is successful and fewer clinical trials would be planned and executed through a traditional way of recruiting patients.

We take the view that the technical infrastructure is not silently or secretly provided. That is, behind any component of the technical infrastructure, just as of any resource, there must be a principal. For example, for e-commerce, we require that the market enable (e.g., eBay, the company) provides the market website, including functionality for sellers listing items, buyers placing bids, and the market enabler determining the winning bid. Similarly, the clinical trials STS would require a principal providing the infrastructure, which could be the clinical trials company or the hospital where the patient is seen. Requiring a principal behind the infrastructure ensures that we can impose requirements on the infrastructure and make sure that there is a principal within the STS who is accountable for such requirements. In typical cases, some of the infrastructure would come from the STS-as-is. For example, the network connectivity needed for e-commerce is provided by the Internet Service Provider of each buyer and seller. It is not within the scope of the e-commerce market interactions but its functioning is a necessary assumption for success of the STS.

5 Closing remarks

Towards a better characterization of ASCS. In Sec. 2 we introduced a few architectural properties to classify systems. A good compilation of examples and their scoring according to that list of properties should provide insights on the correctness and completeness of that list.

Another direction worth exploring (using the WIT framework) is the interplay between the means to specify the social space and its governance, on one side and, on the other, the technological “platform” that implements those means [11]. Two lines of exploration are obvious:

- Focusing on the institutional view of ASCS, look into distinct classes of applications—like on-line role-playing games, participatory simulation environments, prediction markets—and identify the expressive features that are common to the STS that fall in the class.
- In contrast, one may look into the class of socio-cognitive systems that may be built with particular existing platforms—for example *wiki*-based collaborative environments, *Amazon Turk*-based micro tasking aggregation systems, *Repast*-based simulations or *CryEngine*-based games—and abstract from these the affordances and the expressive and operational power of the devices that implement them.

Methodological outlook . The discussion of the process of moving from an existing STS to a new one in Sec. 4 is motivated by the aspiration of building ASCS in a *principled manner*. While the discussion in that section was limited to the identification of the

main components and activities in the innovation process, we wanted to give an indication of the actual complexity of that process and the need to make a thorough analysis on which methodological guidelines may be founded.

One particular aspect that we believe deserves serious consideration in this respect, is to identify criteria to qualify as “adequate” the design and implementation processes of an ASCS, alongside the methodologies that guarantee that those criteria are met. (see [8] for a related argument).

Practical significance of ASCS There are two reasons (beyond their characterization) for the empirical study of of ASCS. One is to provide an objective basis for theoretical and technological developments. The other is to understand—from economic, sociological, political and anthropological perspectives—how value is created through ACSC and how that value can be acquired for the benefit of society. This task is, evidently, a rather obvious challenge for interdisciplinary research.

An emerging scientific field. We share the view of Castelfranchi [5], that we are on the threshold of a new society where ASCS will be a pervasive reality. It is one that we do not fully understand and one of which we are becoming citizens through our use of ASCS. It is perhaps not an exaggeration to claim that it may be worth developing a scientific view of this reality and consequently develop the conceptual and theoretical constructs to explain what is happening and to have a crisper view of what may come next. Maybe, in a way not all that dissimilar to the *zeitgeist* of the early fifties that gave birth to artificial intelligence—with its “mind as processor” model for individual rationality—we are witnessing a new *zeitgeist* that may give birth to a new *artificial social intelligence*.

Acknowledgments

Pablo Noriega received support from the European Network for Social Intelligence, SINTELNET (FET Open Coordinated Action FP7-ICT-2009-C Project No. 286370) and Generalitat of Catalunya grant 2009-SGR-1434. Munindar Singh was partially supported by the U.S. Department of Defense (National Security Agency) under the Science of Security Lablet grant.

References

1. Carlos Alchourron, Peter Gärdenfors, and David Makinson. On the logic of theory change: partial meet contraction and revision functions. *Journal of Symbolic Logic*, 50:510–530, 1985.
2. Giulia Andrighetto, Guido Governatori, Pablo Noriega, and Leendert W. N. van der Torre, editors. *Normative Multi-Agent Systems*, volume 4 of *Dagstuhl Follow-Ups*. Schloss Dagstuhl - Leibniz-Zentrum fuer Informatik, 2013.
3. Kenneth J. Arrow. *Social Choice and Individual Values*, volume 12. Yale University Press, 2012.

4. Cristiano Castelfranchi. InMind and OutMind; Societal Order Cognition and Self-Organization: The role of MAS. Invited talk for the IFAA-MAS “Influential Paper Award”. AAMAS 2013. Saint Paul, Minn. US. <http://www.slideshare.net/sleeplessgreenideas/castelfranchi-aamas13-v2?ref=httpMay2013>.
5. Cristiano Castelfranchi. Making visible the invisible hand. the mission of social simulation. In D. F. Adamatti, G. P. Dimuro, and H. Coelho, editors, *Interdisciplinary Applications of Agent-Based Social Simulation and Modeling*, pages 1–314. IGI Global. Hershey, PA, 2014.
6. Amit K. Chopra, Fabiano Dalpiaz, F. Başak Aydemir, Paolo Giorgini, John Mylopoulos, and Munindar P. Singh. Protos: Foundations for engineering innovative sociotechnical systems. In *Proceedings of the 18th IEEE International Requirements Engineering Conference (RE)*, pages 1–10, Karlskrona, Sweden, 2014. IEEE Computer Society.
7. Andrew Jones and Marek Sergot. A formal characterization of institutionalized power. *Logic Journal of the IGPL*, 4(3):427–446, 1996.
8. Andrew J. I. Jones, Alexander Artikis, and Jeremy Pitt. The design of intelligent socio-technical systems. *Artif. Intell. Rev.*, 39(1):5–20, 2013.
9. Samhar Mahmoud, Gareth Tyson, Simon Miles, Adel Taweel, Tjeerd Vanstaa, Michael Luck, and Brendan Delaney. Multi-agent system for recruiting patients for clinical trials. In *Proceedings of the 2014 International Conference on Autonomous Agents and Multi-agent Systems*, AAMAS ’14, pages 981–988. International Foundation for Autonomous Agents and Multiagent Systems, 2014.
10. Pablo Noriega, Amit K. Chopra, Nicoletta Fornara, Henrique Lopes Cardoso, and Munindar P. Singh. Regulated MAS: Social Perspective. In Giulia Andrighetto, Guido Governatori, Pablo Noriega, and Leendert W. N. van der Torre, editors, *Normative Multi-Agent Systems*, volume 4 of *Dagstuhl Follow-Ups*, pages 93–133. Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik, Dagstuhl, Germany, 2013.
11. Pablo Noriega, Julian Padget, Harko Verhagen, and Mark d’Inverno. The challenge of artificial socio-cognitive systems. In *Pre-proceedings of Coordination, Organizations, Institutions, and Norms in Agent Systems COIN@AAMAS 2014, Paris, May 2014*, Lecture Notes in Computer Science, page tbd. Springer, 2014.
12. John R. Searle. What is an institution? *Journal of Institutional Economics*, 1(01):1–22, 2005.
13. Munindar P. Singh. Norms as a basis for governing sociotechnical systems. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 5(1):21:1–21:23, December 2013.
14. Eric Trist. The evolution of socio-technical systems. *Occasional paper, Ontario Ministry of Labour*, 2, 1981.
15. Axel van Lamsweerde. From system goals to software architecture. In *Formal Methods for Software Architectures*, volume 2804 of *Lecture Notes in Computer Science*, pages 25–43. Springer, 2003.