Towards a Framework for Socio-Cognitive Technical Systems

Pablo Noriega[1], Julian Padget[2(✉)], Harko Verhagen[3], and Mark d'Inverno[4]

[1] IIIA-CSIC, Barcelona, Spain
pablo@iiia.csic.es
[2] Department of Computer Science, University of Bath, Bath, UK
j.a.padget@bath.ac.uk
[3] Stockholm University, Stockholm, Sweden
verhagen@dsv.su.se
[4] Goldsmiths, University of London, London, UK
dinverno@gold.ac.uk

Abstract. This paper is an invitation to carry out science and engineering for a class of socio-technical systems where individuals — who may be human or artificial entities — engage in purposeful collective interactions within a shared web-mediated social space. We put forward a characterization of these systems and introduce some conceptual distinctions that may help to plot the work ahead. In particular, we propose a tripartite view (*WIT Trinity*) that highlights the interplay between the institutional models that prescribe the behaviour of participants, the corresponding implementation of these prescriptions and the actual performance of the system. Building on this tripartite view we explore the problem of developing a conceptual framework for modelling this type of systems and how that framework can be supported by technological artefacts that implement the resulting models. The last section of this position paper outlines a list of challenges that we believe are worth facing. This work draws upon the contributions that the MAS community has made to the understanding and realization of the concepts of coordination, norms and institutions from an organisational perspective.

1 Introduction

"Social coordination" is a many-faceted phenomenon that has been the subject of attention in a number of scientific communities: from economics to social anthropology, from biology to computer science. The arrival of the internet and the massive adoption of social networks and other web-enabled practices have lead the notion of social coordination to acquire new meaning and, in reference to such on-line situations, an unprecedented and substantial economic and social importance. Hence, we put forward this position paper in order to start a debate about the research agenda (i) by making a first attempt to identify the key features that characterize the space of artificial socio-cognitive technical systems (SCTS) (ii) outlining an intentional architecture for SCTS, and (iii) sketching

A. Ghose et al. (Eds.): COIN 2014, LNAI 9372, pp. 164–181, 2015.
DOI: 10.1007/978-3-319-25420-3_11

some ideas, informed by some possible application domains, for a software engineering approach to help realize SCTS, utilizing the many contributions of the COIN community.

We are witnessing the birth of a new sort of tools that, anchored to human cognitive capabilities, aim to support human-like social interactions in a virtual space where the frontiers between the physical and the artificial are increasingly difficult to determine. There is an opportunity to observe with a scientific eye how this process is taking place and articulate an understanding that gives grounds to a serious assessment of its positive and negative aspects and, perhaps, to its evolution. On the other hand, there is also a technological opportunity to address the creation of those new tools in a principled way. Needless to say that behind those opportunities there are ethical concerns that should be taken into account.

This paper aims to be a step towards realising those two opportunities. Hence, its focus is on social coordination within a particular kind of systems that enable individuals — who may be human or artificial entities — to interact in a shared web-mediated social space in a purposeful fashion. We shall call them *(artificial) socio-cognitive technical systems* (SCTS). Our goal is to provide foundations for an understanding of these systems and in time establish a principled methodology for their construction. The immediate outcome in this paper is the introduction of some conceptual distinctions for that purpose. The ancillary objective of this paper is to point the way towards future actions.

This is a position paper in which our key contributions are:

1. An intentional definition of SCTS (Sect. 2), with two essential distinct components: socio-cognitive agents and the social space where these interact;
2. A "tripartite view" (Sect. 3) that attempts to explain the interplay among the three complementary aspects of an SCTS: the institutional, the technological and the "real-world";
3. An identification of those features that are required to model a social space for SCTS that has at least three properties or *affordances* (see Sect. 4): (i) Awareness, by which participants perceive their context (ii) Coordination, by which collective action is enabled and (iii) Validity which establishes a set of correspondences between the elements of our tripartite description of SCTS;
4. How the relationship between the model of an SCTS and its implementation is mediated by a metamodel and a platform (Sect. 5), and, finally
5. A call to arms (Sect. 6)

2 A Superficial Exploration of SCTS

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.

This crude characterisation may be clarified by making explicit some underlying assumptions:

Notion 1. *A* socio-cognitive technical system (SCTS) *is a multiagent system that satisfies the following assumptions:*

A.1 ***System.*** *A socio-cognitive technical 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.*

A.2 ***Agents.*** *Agents are entities who are capable of acting within the social space. They exhibit the following characteristics:*

- **A.2.1** ***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 [7].*
- **A.2.2** ***Opaque Socio-cognitive Agents.*** *The system, in principle, has no access to the decision-making models, or internal states of participating agents.*
- **A.2.3** ***Mixed.*** *Agents may be human or software entities (we shall call them all "agents" or "participants" where it is not necessary to distinguish).*
- **A.2.4** ***Heterogeneous.*** *Agents may have different decision models, different motivations and respond to different principals.*
- **A.2.5** ***Autonomous.*** *Agents are not necessarily competent or benevolent, hence they may fail to act as expected or demanded of them.*

A.3 ***Persistence.*** *The social space may change either as effect of the actions of the participants, or as effect of events that are caused (or admitted) by the system.*

A.4 ***Perceivable.*** *All interactions 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.*

A.5 ***Openness.*** *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 at some point or not.*

A.6 ***Constrained.*** *In order to coordinate actions, the space includes (and governs) regulations, obligations, norms or conventions that agents are in principle supposed to follow.*

We may think of these systems as *socio-technical* systems because of the participation of humans and software components [23], although they are better understood in the sense of [18] or even [22] where software agents may be involved. We use the term *artificial* because we want to stress the fact that there is some external design of the system and the term *socio-cognitive* to stress the fact the we glimpse some notion of social intelligence. Because of the assumption of intrinsic constraint on action (***A.6***), in standard multiagent systems terminology, the above assumptions characterize a type of *normative multiagent system* [3].

Jones et al. [16] refer to this type of system as an *intelligent socio-technical system*. While in this characterization, the adjective "intelligent" denotes an assumption of rationality, they also assert that these systems involve entities that "*interact with each other against a social, organisational or legal background*" (as in ***A.2*** above). Analogously, Castelfranchi calls them *socio-cognitive technical systems* to stress the fact that in order to characterize 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*" [7]. It is in this spirit that we adopt the term; however, we would like to stress the fact that in this paper we are mostly concerned by the fact that these systems are designed and built with some purpose in mind and occasionally refer to them as *artificial* socio-cognitive systems to capture the essence of these last two interpretations and omit the "technical" label to avoid redundancy.

Although it would be premature to propose a broad taxonomy of artificial socio-cognitive systems, it is nevertheless possible to identify application domains where these systems are or will be paradigmatic. For example, serious on-line games, massive multiplayer on-line role playing games, mixed-level participatory simulation of social systems, open innovation environments as well as other crowd-based applications, on-line electronic markets, policy support systems, or on-line alternative dispute resolution, to name a few. Such an empirical approach would be essential if one aspires to a serious characterization of SCTS. An argument for the need of empirical research on existing SCTS is formulated below (Subsect. 6.2). The pursuit of a proper characterization of SCTS (and its empirical foundations) was articulated in [9].

The research programme for SCTS that we envision should eventually enable us to design new such systems using a principled approach. We propose to address the general problem, first by delimiting the universe to an explicit set of features that may allow us to decide whether a given system — existing or in design — belongs to that universe, and second, developing an abstract understanding of what is common to these systems. These two steps would provide foundations for SCTS formalisms, tools and methodologies.

3 The WIT Trinity: A Tripartite View of Artificial Socio-cognitive Systems

Keeping the assumptions ***A.1–6*** and examples in mind, one may advance an intuitive description of SCTS as systems where it is possible to *govern* the interaction of agents that are situated in a physical or artificial world by means of technological artefacts. The key element in this description is in the "governance" part that mediates between the world and the technological artefacts. It is an aspect worth distinguishing in SCTS because of the need to control the activity of complex individuals that is at the root of SCTS (***A.2*** and ***A.6***). In order to elucidate how such governance is achieved we propose the following tripartite view of SCTS (the *WIT Trinity)*[1]:

[1] We abuse the term "trinity" to stress the fact that every SCTS has these three views, that each of these views has several characteristic features but that the three views are interrelated in an indissoluble way in order to constitute *the* SCTS.

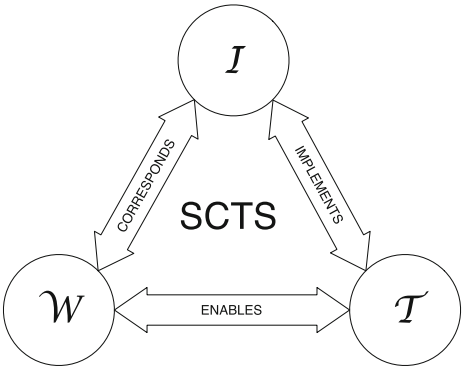


Fig. 1. The WIT trinity: The ideal system, $\mathcal{I}$; the technological artefacts that implement it, $\mathcal{T}$, and the actual world where the system is used, $\mathcal{W}$. After [19].

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 according to the rules set out in $\mathcal{I}$.

These three views are interrelated through three binary relationships (as depicted in Fig. 1). The institutional world corresponds with the real world by a sort of "counts-as" relationship [15,21] by which (brute) facts and (brute) actions in the real world correspond to institutional facts and actions in the institutional world $\mathcal{I}$ only when these comply with the institutional conventions, in which case the institutional effects of those institutional actions carry over to have effects in the real world.[2]

Secondly, the conventions prescribed in the institutional world have their counterpart in the technological world in the sense that institutional conventions

[2] Note that $\mathcal{W}$ is not the *entire* real-world, it is only the fragment of the physical reality that affects and is affected by the SCTS. Thus, if we think of *Amazon* as an SCTS the $\mathcal{W}$ (of *Amazon*) corresponds only to the reality around those online transactions that take place on line between a company call *Amazon.com*, buyers and sellers of books through the system that supports these transactions. In other words, there are events that happen in the word that may or may not be relevant for *Amazon* depending on what $\mathcal{I}$ (of *Amazon*) stipulates, for instance; the real-world event "new dollar / euro exchange rate" is in $\mathcal{W}$ (of *Amazon*) –or "meaningful" or relevant in *Amazon*–only if payments may be made in either of those two currencies. Likewise, a move in an online chess game is part of the game (*is* in $\mathcal{W}$), if and only if it is communicated and acknowledged through the on-line system ($\mathcal{T}$) and complies with the rules of chess defined in $\mathcal{I}$ (it is a proper chess move and is made on time, for example).

constitute a specification of the requirements of the system that is implemented in $\mathcal{T}$.

In turn, the system, as implemented in $\mathcal{T}$ is what enables interactions (through a proper interface) in $\mathcal{W}$, so the agents in $\mathcal{W}$ control the artefacts in $\mathcal{T}$, but also, we contend, this relationship is symmetric, in that by virtue of the percepts delivered via $\mathcal{T}$, the artefacts in $\mathcal{T}$ effect some control over the agents 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) the identity of agents in $\mathcal{W}$ to be properly reflected in their counterparts in $\mathcal{I}$ and preserved as long as the agents are active in the system, (iii) the agents that participate in $\mathcal{W}$ to 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) the commitments that are established according to $\mathcal{I}$, to be 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 *enables* 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.

3.1 The Shared State of an Artificial Socio-Cognitive System

We emphasize that, in the preceding discussion, we are suggesting that the three views correspond to the same SCTS. In other words, when we make reference to an SCTS, we always refer to an entity that exists in the real world, works by means of some technological artefacts and behaves according to some institutional conventions. We also state that the three views are interrelated. However, we may go a step further and establish the actual correspondence between the three views. For that purpose we rely on the notion of *shared state*.

The intuition behind shared state is that at any point in time, what happens in the world and enters the system produces some effects in the computational system that become effective in the world. In other words, that the *state of the world*, as far as the system is concerned, changes if and when an *attempted* action in $\mathcal{W}$ is *validated* by $\mathcal{I}$, and then the code in $\mathcal{T}$ *processes* the input that happens in $\mathcal{W}$ and outputs the effects in $\mathcal{W}$.

We may use the WIT Trinity of SCTS to get a clearer picture of how interactions of agents within the system change the shared state. Figure 2 illustrates how interactions among individuals take place within a socio-cognitive system.

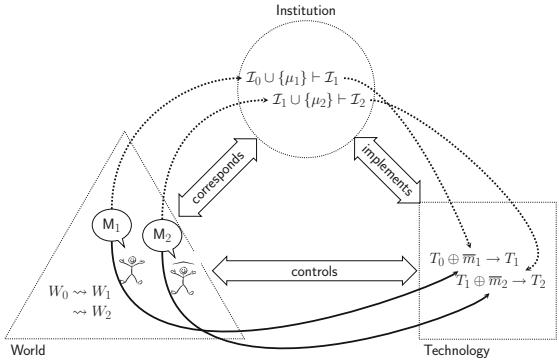


Fig. 2. Shared state in a socio-cognitive system

First let us focus on $\mathcal{W}$. Take two agents a_1 and a_2, in $\mathcal{W}$, who are about to interact within the system, each through its own interface device. Notice that, since these individuals are real — human or software agents — and are present in the part of the real world involved with the system, then the objects that exist, the facts that are true and whatever changes take place *in that part of the real world*, are the same for both agents, and for every other agent that is in the system at that point in time. Technically speaking, the agents *share the state* of $\mathcal{W}$. Now let the first agent (a_1) take an action M_1 in $\mathcal{W}$. Provided that M_1 is a *feasible* action, that action changes some facts in $\mathcal{W}$, and the state of the world changes from $\mathcal{W}_0$ to $\mathcal{W}_1$. Now, if a_2 takes a new feasible action M_2 the world changes to a new shared state $\mathcal{W}_2$. Second, from a computational perspective, inputs M_1 and M_2 correspond to messages $\bar{m}_1$ and $\bar{m}_2$ that when *processed* in $\mathcal{T}$, produce changes in the data structures and values of variables in $\mathcal{T}$, hence new successive *shared computational states*, $\mathcal{T}_1$ and $\mathcal{T}_2$. Finally, a similar thing happens in $\mathcal{I}$ when an institutional action μ_1, (that corresponds to action M_1 and is implemented as message $\bar{m}_1$) takes the system from an *institutional state* $\mathcal{I}_0$ where certain formulas are admitted, to a new shared institutional state $\mathcal{I}_1$ with new admitted formulas, if and when μ_1 is an institutionally *admissible* action, and likewise for a proper μ_2. In other words, we have now established a more abstract notion of an SCTS by introducing three complementary components:

- A tripartite understanding of artificial socio-cognitive systems.
- The notion of state (of the world, computational, institutional), the use of valid interactions as the sole way of changing that state and the existence of a set of conventions that determine when an interaction is valid and, if so, how it changes the state.
- Three mappings between the three views of the system: (i) mappings between actions, messages and formulas, (ii)mappings between states of the world, system and institution and (iii) mappings between three notions of validity of interactions: feasible, processable and admissible.

These constructs can be made precise, although such task is beyond the scope of this paper, but even this crude description brings to light three crucial features that an SCTS must provide in order to control sophisticated interactions. First, an agent needs to be *aware* of the state of the world in order to decide what to do at some point. Moreover, in order to attempt an action, that agent needs to *coordinate* with other agents with whom it is interacting or would like to interact. Finally, the system needs to support a proper notion of *validity*, so that the "isomorphisms" described above between the evolution of the states of $\mathcal{W}$, $\mathcal{I}$ and $\mathcal{T}$ are operational.

4 Designing the Social Space

In Sect. 2, we characterized SCTS as collective processes involving several socio-cognitive agents (human or not) who engage in web-enabled interactions within a shared social space. We now want to move a step ahead and see how an SCTS can be designed or modelled. For that purpose and based on the previous discussion, we need to account for a way of dealing with the evolution of the shared state. Keeping in mind the distinctions between system, participants and social space (***A*.1**) and the fact that agents are opaque to the system (***A*.2.2**), we may limit our attention to the social space. Moreover, because of the correspondences implicit in the WIT Trinity, we may limit the discussion to the features of the social space in $\mathcal{I}$ and then extend that understanding to $\mathcal{T}$ and $\mathcal{W}$. In other words, if we want to design SCTS, what are the features we need in the social space so one can determine what is a state of the system and what is involved in performing a valid action. We propose to achieve this through what we call "affordances" (in the spirit of Norman [20]) needed to *model* an SCTS.[3]

Notion 2. *An* affordance *(of the social space of an SCTS) is a property of the social space that supports effective interactions of agents within an SCTS.*

At the end of the previous section, we postulated three *affordances* of every SCTS:

1. *Awareness*, which provides participating entities access to those elements of the shared state of the world that should enable them to decide what to do
2. *Coordination*, so that the actions of individuals are conducive to the collective endeavour that brings them to participate in the SCTS and
3. *Validity* that preserves the proper correspondences of the tripartite view.

There may be others, but we identify these because they contribute directly firstly, to the establishment of individual perception of (common) social situations, secondly to the realization of the mechanisms for collective action and thirdly to the correctness of the activity as a whole.

[3] Recall Norman's barrel. It is a water-tight cylinder with an intended affordance for holding liquids but it also provides affordances of a table or a hiding place. Similarly, the features we enumerate below have an intended affordance but others affordances may be achieved (for free) depending on the way they are specified or implemented.

It is evident that *awareness* and *coordination* — and other affordances as well — may be achieved by a variety of means. Consequently, one could use a way to make explicit the particular means through which these properties are achieved in a given SCTS; first because there may be reasons to choose among different particular means and second because participants — and technological artefacts — need to conform to the particular means used for modelling the given SCTS. For this purpose we, first, take a look at features that are involved in the achievement of the essential affordances. Next we postulate the notion of a *metamodel* as a way of describing the particular means that are used to generate those features.

A glance at some families of SCTS mentioned earlier (games, simulation, crowd-based systems, electronic markets,...) suggest concrete features that appear to be necessary for the modelling of most SCTS:

1. **Ontology.** The point of this feature is to establish the objects that describe and populate the social space. Some objects may be generic to a metamodel (norm, scene, workspace,..) or to a family of SCTS (weapons in first person shooter games, contract in prediction markets, etc.), others are specific to the application domain of the particular SCTS (sword, bid,...).
2. **Primitive Actions and Events.** How percepts are represented. For example, offering a picture for sale in an auction, bidding for it and declaring a bid invalid; reading the room temperature.
3. **Activities.** The possibility of organising atomic actions into repetitive activities through protocols, social semantics, a set of norms, etc. (to represent a bidding round or mapping crisis events of a city).
4. **Subspaces and their Interrelationships.** Constructs to describe (i) activities that involve only part of the participants who share a substate of the system that is not necessarily accessible to other participants, (ii) how these activities are interrelated and (iii) whether or not agents may be active in more than one activity at a given time (e.g., sequential scenes in a play, simultaneous auctions in *eBay*).
5. **Social Structure.** Roles (author and reviewer) and relationships among roles (authors cannot review their papers); groups (ad.hoc: task force; standard: jury; board of directors) and organisational structures (team, department).
6. **Social Devices.** Means for (i) tagging the behaviour of individuals, so that participants may become aware of particular qualities (trust, social standing) or (ii) processes for modifying it (ostracism, whitewashing, fines and incentives).
7. **Regulatory System.** Norms, normative consequence, enforcement mechanisms and procedures, norm life-cycle management, etc. (see [19] for a thorough discussion of normative affordances and features).
8. **Dynamics of the System.** How to measure the performance of the system and the means to make the system change over time.
9. **Types of Agents.** Means to choose the composition of the class of participants and specially to include as part of the system design those agents (or

their roles) whose decision-making model is defined or is in control of the system itself. Two types are most usual: *external* agents that are opaque to the system and *internal* who act on behalf of the system who is responsible for their behaviour. For example, in games: "players" (usually human) and "non-player characters" (software agents deployed by the system designer).

10. **Languages and Information Framework**. Needed to express the specific instantiation of features (for protocols, norms,...) and to store the design and enactment data (local and global states of the system, agent profiles, performance indicators, etc.).

These examples of features are meant to suggest how to make explicit the means required for designing or modelling an SCTS. With the following descriptions we aim to make more precise what we understand by "the means for modelling" and "modelling" an SCTS.[4]

Notion 3. *A* metamodel *(for SCTS) is a collection of languages, data structures and operations that when instantiated produce a model of an SCTS (and its internal agents, if any), through features that achieve the affordances of awareness and coordination in a social space.*

Consequently, a model is simply a "good" description of a socio-cognitive system:

Notion 4. *A* model *of an artificial socio-cognitive system $\mathcal{S}$ is the instantiation of a metamodel for SCTS, such that the correspondence between the view of $\mathcal{S}$ in $\mathcal{W}$ matches the view of $\mathcal{S}$ in $\mathcal{I}$.*

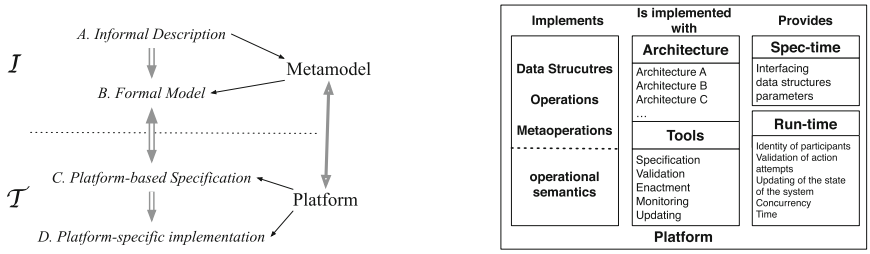
Note that this "matching" entails that the integrity requirements of the three relationships are in fact correctly achieved. In particular (i) the *counts-as* relationship is correctly established by participants having the proper entitlements and an appropriate bijection between terms in $\mathcal{I}$ and objects and potential actions in $\mathcal{W}$, (ii) the model is faithfully implemented in $\mathcal{T}$ and (iii) the input/output flow between $\mathcal{T}$ and $\mathcal{W}$ is not corrupted. Note also that while we have kept the discussion in $\mathcal{I}$, in the next section we connect $\mathcal{I}$ with $\mathcal{T}$ by clarifying the relationship between the ideal model of an SCTS and the actual implementation of that SCTS that is underneath the achievement of (ii).

5 Metamodels and Platforms

In our characterization of metamodel (Notion 3) we did not commit to implementation and formalisation although both are desirable properties. As far as

[4] We adapt to SCTS the standard use of *model* as an abstract representation of a real entity and *metamodel* as the abstract representation of models. See for example this use in UML: "...[an abstract syntax that defines] modelling concepts, their attributes and their relationships, as well as the rules for combining these concepts to construct partial or complete ... models." (superstructure version 2.2 (2009-02-03), p1).

implementation is concerned, it would be rather convenient to have a cohesive collection of technological artefacts (a platform) that includes a specification language to make a precise definition of the model. Then, other artefacts of the platform would produce a run-time implementation of the model that controls inputs and outputs that preserve the validity conditions of the shared context, as postulated in Sect. 3. Thus, the "implement" relationship depicted in Fig. 1 may be elucidated by the diagram in Fig. 3a.



(a) The implementation process of an SCTS.

(b) Contents and functions of an SCTS platform.

Fig. 3. Metamodel and platform

In a top-down reading of the diagram, one starts with an informal understanding of the system (A) that will be implemented (D). Ideally, one would expect to have a formal model (B), which corresponds to the exact version of the SCTS that one would like to have in $\mathcal{I}$ so that the effects of the actions on $\mathcal{W}$ have the exact effect $\mathcal{W}$ prescribed in $\mathcal{I}$. However, the transition from an informal representation of an SCTS to a formal model is far from straightforward [16]. One way out is to rely on the metamodel to connect (A) and (B) since, ideally, it provides the abstract constructs to describe (A) in precise terms. The metamodel also provides a bridge between (B) and (D) when it is linked to a *platform* that includes a specification language such that the metamodel instantiations specified with it (C) generate faithful implementations of the formal models (B).[5]

A bottom-up reading of the diagram suggests a symmetric path where one starts with an existing platform and intends to determine formal and computational properties of the models that can be implemented with it (such would be the case of SCTS constructed using, for example the *Amazon Turk* or mash-ups of *Facebook* and *Ushahidi*).

[5] This point is aptly made in Jones et al. [16] (Step 1, Step 2. Phase 1, and Step 3) where they argue for a rigorous analysis of the expressiveness of the formalisms and their operationalisation, in order to arrive to a proper specification (C). We acknowledge that those same issues — as well as the computational considerations of their Step 2, Phase2 — are all present in the "top-down" design and the choice of the metamodel.

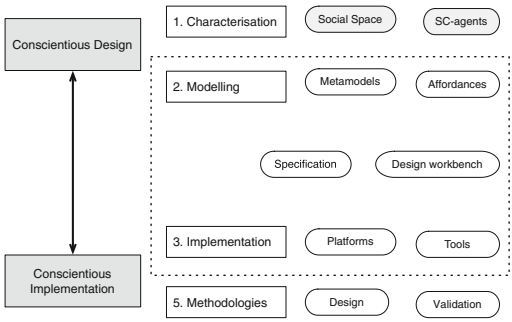


Fig. 4. Main research blocks on SCTS

There are some metamodels for social coordination motivated by work on open multiagent systems. The following have been in development for a number of years and have a cohesive collection of technological artefacts that support them and have been used to design or implement SCTS of reasonable complexity: ANTE [6], EI/EIDE [11], InstAL [10], MOISE/JaCaMo [17] OCeaN [13], OperA/OperettA [2] and THOMAS/ROMAS [14]. It is outside the scope of this paper to make a systematic analysis of these but an illustrative comparison of ANTE, OCeaN and EI/EIDE is available in [12].

6 A Call to Arms

This paper looks at artificial socio-cognitive technical systems from a broad and superficial perspective, as an attempt to open a path into a new field. Although it is too early to draft something as precise as a research programme, Fig. 4 maps a rough itinerary suggested by the previous discussion.

6.1 Technical Challenges

Validity as an affordance. When we introduced the notion of *affordance* (Notion 2), we stated that *validity* is an essential affordance of the social space, in addition to awareness and coordination; an assertion based on the preceding discussion of shared context. In the discussion of the notion of *model* (Notion 4), we stated that a model is *valid* if it preserves the "counts-as" relationship (and by transitivity of the tripartite diagram, its implementation is supposed to uphold that validity in the real world). In other words we wish to sustain the implicit claim that validity is a *supervened affordance* of the social space. A claim that should first be made precise and then made operational. Informally, the argument is as follows: from a top-down perspective, one would need to prove that the normative components of the metamodel define models whose validity can be demonstrated; and from a bottom-up perspective, the kernel of the proof is in the bridge between the *platform* and the *metamodel*, since one may take the position that an action in $\mathcal{W}$ is valid in $\mathcal{T}$ (is accepted as an input), and should be valid in $\mathcal{I}$ only if the *metamodel* is a faithful formalisation of the *platform*.

Affordances and features. We also side-stepped – in Sect. 4 – two issues that are central to the notion of metamodel:

1. The first is ontological. It is the problem of determining whether a list of features is a good way to support the affordances of SCTS. On one hand, we have incidental indication that all the features we mentioned are present in one way or another in the families of examples we have mentioned along the paper, and some objective indication that most are needed to implement the type of SCTS that the seven frameworks mentioned in Sect. 5, in as much as most of these features are directly accessible (i.e. features may be expressed and implemented with their basic constructs and artefacts), and may otherwise be paraphrased. However, a serious effort on an extensional description of SCTS is needed to avoid the latent *petitio* of this argument.
2. The second is methodological. Whichever way this "completeness" is achieved or demonstrated, the problem of choosing a collection of features and a good form of description and implementation for those features needs to be resolved for the design of a metamodel (and its corresponding platform), and then the actual instantiation has to be decided when modelling a particular SCTS.

Metamodel specialisation. The previous remark directs attention at a significant design challenge: how specialised should a metamodel be? There is no obvious reason that we can find that prevents the creation of a single metamodel for all SCTS but neither is there an obvious reason that we can find to claim that developing such an archetype would be advantageous.

Experience with the seven metamodels listed in Sect. 5 confirm the procrustean curse of formalisms and implementations: every time one models an SCTS with one of those frameworks, the SCTS is "tortured" into the particular features afforded directly by the framework. We presently lack a systematic comparison of frameworks that assesses their advantages and limitations and provides sound guidelines for choosing one or another, or to approach the question of whether a unifying framework would be that ultimate metamodel.

On the other hand, the same reservations about the procrustean curse would suggest the possibility of moving in the opposite direction. That is, develop metamodels (and platforms) that are well-adapted to particular types of SCTS: a metamodel for games, another one for participatory social simulations, yet another one for crowd-based SCTS, and so on. The question then is, where should the specialisation stop? A metamodel for games or a metamodel for first-person shooter games and one for MMORGs and one for serious games? Again, we lack enough empirical analysis of families of SCTS and a robust understanding of affordances, features and metamodels to venture even a tentative answer, but these are open questions that, we believe, may be fruitfully explored.

Metamodel/platform interplay. In Sect. 5 we pointed out the *Whorfian* [24] relationship between the conceptual framework that supports the formulation of a model of an SCTS and the artefacts that are used to implement it (i.e., the expressiveness of the conceptual metamodels and the facilities provided by platforms that serve to implement particular SCTS). In some families of SCTS, there

is a predominance of the platform over the metamodel fostered by the wealth of cases for which an existing platform is a good match (for example the *Amazon Turk*[6] or MMORG engines, like *RedDwarf server*), or fostered by the versatility of the basic functionalities of a platform (e.g. *Facebook* used as the input for crowd-sourcing the draft of the Moroccan Constitution). On the other hand the experience with current metamodels is that the platform that supports them is not necessarily an integral implementation. Although in many cases the actual features of the metamodels are immediately expressible in the platform, many times they can be achieved only through paraphrases.

The trade-off is not always clear and we believe that it is worth exploring ways to find a balance of platform and metamodel expressiveness by examining the problem from both sides. One possibility (mentioned above) may be to develop a more "generic" metamodel that addresses all properties with a variety of formalisms that may be assembled or instantiated in order to model specific SCTS. Figure 5 is a toy candidate for the type of generic metamodel that involves all the properties we listed in Sect. 4. Another approach to the interplay of metamodel and platform is to construct a sound conceptual model for mashing-up available artefacts and platforms in order to provide proper foundations to those components and, by extension, to the resulting mash-up.

The dynamics of actual SCTS. In this paper we have discussed SCTS as if they were static objects that exist in an abstract reality that is limited to those events, facts and actions that are directly relevant to the state of that SCTS. This simplification is wholly inappropriate when observing and designing actual SCTS. In that situation, a framework for SCTS needs to address two significant aspects: First, the social context where the SCTS is designed and operates, and second, how to account for the changes that a given SCTS may undergo beyond the evolution that has been programmed into it at design-time. A first discussion of these issues can be found in [9] but, evidently, these are no minor challenges.

Separation of concerns. We hold the assumption (***A.1***) that agents and social space are different components of an SCTS. This separation is useful for a conceptual analysis of SCTS, but it may also be valuable from a design point of view. An illustration of this value is the advantages of designing non-player characters (NPC), or in general BDI agents [5] within a norm-regulated environment. Likewise, the separation of design and implementation — achieved by having a metamodel and platform — gives designers the possibility of choosing the tools that implement their ideas, rather than choosing the problems that are implementable by the tools. The degree and tooling of those types of separation deserve, we believe, a systematic analysis.

Reinventing the wheel. Because of the intrinsic interdisciplinary character of social coordination in SCTS, there is a natural propensity to approach the subject from a particular perspective — ours being software development and regulated MAS — without paying due attention to the questions, principles, theories

[6] https://www.mturk.com.

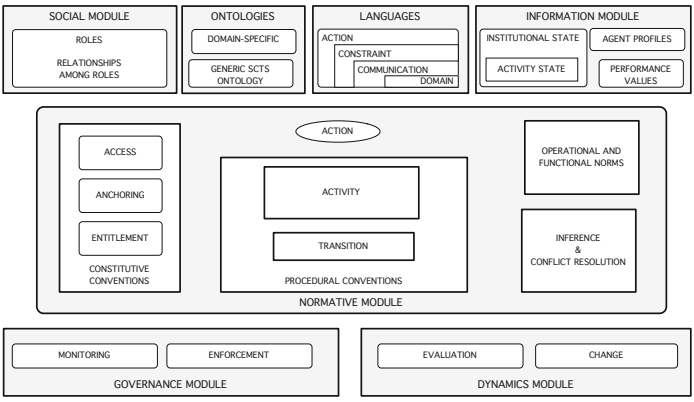


Fig. 5. A "generic" metamodel for SCTS. Each feature contains several formalisms and their supporting artefacts that are tailored to the peculiarities of a given SCTS

and artefacts that have been and are being developed in the theoretical fields of the inter-discipline. As Jones et. al. propose in [16], a serious use of the pertinent developments of other converging disciplines is not only useful but essential, if one intends to develop a principled approach to the description and design of SCTS.

Towards a conscientious design of SCTS. This meandering of SCTS is motivated by the inevitability of socio-cognitive systems and therefore the need to become aware of the social significance of these systems and the responsibility that scientists and engineers have in the design and deployment of artificial socio-cognitive technical systems. The challenge is to develop precise notions and the associated methodological guidelines and tools to design systems in a conscientious way. This entails, first a clear understanding the inherent values, how to operationalise them and then how to assess that they are properly reflected in the design and the deployed system. A tentative blueprint of the inherent issues may address three dimensions:

1. *Thoroughness.* This is achieved when the system is technically correct, requirements have been properly identified and faithfully implemented. This entails the use of appropriate formalisms, accurate modelling and proper use of tools.
2. *Mindfulness.* This describes supra-functional features that provide the users with awareness of the characteristics of the system and the possibility of selecting a satisfactory tailoring to individual needs or preferences. Thus, features that should be accounted for should include ergonomics, governance, coherence of purpose and means, identification of side-effects, no hidden agency, and the avoidance of unnecessary affordances.
3. *Responsibility.* This is true both towards users and to society in general. It requires a proper empowerment of the principals to honour commitments

and responsiveness to stakeholders legitimate interests. Hence, features like its scrutability, transparency and accountability alongside a proper support of privacy, a "right to forget"; proper handling of identity and ownership, attention to liabilities and proper risk allocation, and support of values like justice, fairness and trustworthiness.

6.2 A Wider View

The motivation behind this work is the realisation that the MAS community and the COIN community in particular is well-positioned to address the challenges that SCTS brings and harness the possibilities of developing a principled methodology for the study and development of SCTS. The space for innovation is still to be plotted but it is undoubtedly vast and some milestones are already visible.

Empirical study of SCTS. This task should be approached for two kinds of reasons. One is to provide an objective basis for theoretical and technological developments, and (as argued in [9]) formulate a characterization of SCTS in the spirit of Kenneth Arrow's [4] or Alchourron, Gardenfors and Makinson [1]. The other is to understand — from economic, sociological, political and anthropological perspectives — how value is created through SCTS and how that value can be acquired for the benefit of society. This task is, evidently, a rather obvious challenge for interdisciplinary research.

Technological developments. Little needs to be argued about the social significance of platforms that are already available for developing SCTS and how some of their original or intended applications have become massive social phenomena and considerable economic successes. This is not likely to cease in the near future and consequently there is a substantial opportunity for innovation in tools, methodologies and applications. Specially if the emphasis on "principled" design is taken to heart.

Synergies. A systematic study of SCTS will most likely require the convergence of several disciplines. The topic of social coordination is currently being inspected (within the SINTELNET project) from different standpoints: games, social simulation, analytical sociology, cognitive and social psychology, formalisms for informal phenomena, crowd-based applications, institutional theory and philosophy of law. These activities are already fostering collaborations with a strong synergistic component. This experience points in the direction of new academic communities that are likely to spawn conferences and periodic publications and eventually develop curricula and training.

An emerging scientific field. We share the view of Castelfranchi [8], that we are on the threshold of a new society where SCTS 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 SCTS. 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 comes 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* — with "social coordination" as the core of socio-cognitive rationality.

Acknowledgements. The authors wish to acknowledge the support of SINTELNET (FET Open Coordinated Action FP7-ICT-2009-C Project No. 286370) in the writing of this paper. In addition, d'Inverno acknowledges the support of the FP7 Technology Enhanced Learning Program Project: Practice and Performance Analysis Inspiring Social Education (PRAISE).

References

1. Alchourron, C., GŁrdenfors, P., Makinson, D.: On the logic of theory change: partial meet contraction and revision functions. J. Symbolic Logic **50**, 510–530 (1985)
2. Aldewereld, H., Dignum, V.: OperettA: organization-oriented development environment. In: Dastani, M., El Fallah Seghrouchni, A., Hübner, J., Leite, J. (eds.) LADS 2010. LNCS, vol. 6822, pp. 1–18. Springer, Heidelberg (2011)
3. Andrighetto, G., Governatori, G., Noriega, P., van der Torre, L.W.N. (eds.): Normative Multi-Agent Systems. Dagstuhl Follow-Ups, vol. 4. Schloss Dagstuhl - Leibniz-Zentrum fuer Informatik, Wadern (2013)
4. Arrow, K.J.: Social Choice and Individual Values, vol. 12. Yale University Press, New Haven (2012)
5. Balke, T., De Vos, M., Padget, J.: Normative run-time reasoning for institutionally-situated bdi agents. In: Cranefield, S., van Riemsdijk, M.B., Vázquez-Salceda, J., Noriega, P. (eds.) COIN 2011. LNCS, vol. 7254, pp. 129–148. Springer, Heidelberg (2012)
6. Cardoso, H.L., Urbano, J., Brandão, P., Rocha, A.P., Oliveira, E.: ANTE: agreement negotiation in normative and trust-enabled environments. In: Demazeau, Y., Müller, J.P., Rodríguez, J.M.C., Pérez, J.B. (eds.) Advances on PAAMS. AISC, vol. 155, pp. 261–264. Springer, Heidelberg (2012)
7. Castelfranchi, C.: InMind and OutMind; Societal Order Cognition and Self-Organization: The role of MAS. Invited talk for the IFAAMAS "Influential Paper Award". AAMAS 2013. Saint Paul, Minn, US, May 2013. http://www.slideshare.net/sleeplessgreenideas/castelfranchi-aamas13-v2?ref=http
8. Castelfranchi, C.: Making visible "the invisible hand" the mission of social simulation. In: Adamatti, D.F., Dimuro, G.P., Coelho, H. (eds.) Interdisciplinary Applications of Agent-Based Social Simulation and Modeling, pp. 1–314. IGI Global, Hershey (2014)
9. Christiaanse, R., Ghose, A., Noriega, P., Singh, M.P.: Characterizing artificial socio-cognitive technical systems. In: Herzig, A., Lorini, E. (eds.) Proceedings of the European Conference on Social Intelligence (ECSI-2014), Barcelona, Spain, November 3–5, 2014. CEUR Workshop Proceedings, vol. 1283, pp. 336–346. CEUR-WS.org (2014)

10. Cliffe, O., De Vos, M., Padget, J.: Answer set programming for representing and reasoning about virtual institutions. In: Inoue, K., Satoh, K., Toni, F. (eds.) CLIMA 2006. LNCS (LNAI), vol. 4371, pp. 60–79. Springer, Heidelberg (2007)
11. d'Inverno, M., Luck, M., Noriega, P., Rodriguez-Aguilar, J.A., Sierra, C.: Communicating open systems. Artif. Intell. **186**, 38–94 (2012)
12. Fornara, N., Cardoso, H.L., Noriega, P., Oliveira, E., Tampitsikas, C., Schumache, M.I.: Modelling agent institutions. In: Ossowski, S. (ed.) Agreement Technologies. Law, Governance and Technology Series, vol. 8, pp. 277–307. Springer, Dordrecht (2013)
13. Fornara, N., Vigan, F., Verdicchio, M., Colombetti, M.: Artificial institutions: a model of institutional reality for open multiagent systems. Artif. Intell. Law **16**(1), 89–105 (2008)
14. Garcia, E.: Engineering Regulated Open Multiagent Systems. AI Communications (in press)
15. Jones, A.I.J., Sergot, M.: A formal characterization of institutionalized power. Logic J. IGPL **4**(3), 427–446 (1996)
16. Jones, A.I.J., Artikis, A., Pitt, J.: The design of intelligent socio-technical systems. Artif. Intell. Rev. **39**(1), 5–20 (2013)
17. Kitio, R., Boissier, O., Hübner, J.F., Ricci, A.: Organisational artifacts and agents for open multi-agent organisations: "giving the power back to the agents". In: Sichman, J.S., Padget, J., Ossowski, S., Noriega, P. (eds.) COIN 2007. LNCS (LNAI), vol. 4870, pp. 171–186. Springer, Heidelberg (2008)
18. Nikolic, I., Ghorbani, A.: A method for developing agent-based models of socio-technical systems. In: 2011 IEEE International Conference on Networking, Sensing and Control (ICNSC), pp. 44–49. IEEE (2011)
19. Noriega, P., Chopra, A.K., Fornara, N., Cardoso, H.L., Singh, M.P.: Regulated MAS: social perspective. In: Andrighetto, G., Governatori, G., Noriega, P., van der Torre, L.W.N. (eds.) Normative Multi-Agent Systems. Dagstuhl Follow-Ups, vol. 4, pp. 93–133. Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik, Dagstuhl (2013)
20. Norman, D.A.: Affordance, conventions, and design. Interactions **6**(3), 38–43 (1999)
21. Searle, J.R.: What is an institution? J. Inst. Econ. **1**(01), 1–22 (2005)
22. Singh, M.P.: Norms as a basis for governing sociotechnical systems. ACM Trans. Intell. Syst. Technol. (TIST), 1–21 (2013, in press)
23. Trist, E.: The Evolution of Socio-technical Systems. Occasional Paper, vol. 2. Ontario Ministry of Labour, Ontario (1981)
24. Whorf, B.L.: The relation of habitual thought and behavior to language. In: Carroll, J.B. (ed.) Language, Thought, and Reality: Selected Writings of Benjamin Lee Whorf, pp. 134–159. MIT Press (1956). ISBN 0-262-73006-5