Identifying affordances for modelling second-order emergent phenomena with the $\mathcal{WIT}$ framework

Pablo Noriega[1], Jordi Sabater-Mir[1], Harko Verhagen[2], Julian Padget[3], and Mark d'Inverno[4]

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

Abstract. We explore a means to understand second order emergent social phenomena (*EP*2), that is, phenomena that involve groups of agents who reason and decide, specifically, about actions – theirs or others' – that may affect the social environment where they interact with other agents. We propose to model such phenomena as socio-cognitive technical systems that involve, on one hand, agents that are imbued with social rationality (thus socio-cognitive) and, on the other hand, a social space where they interact. For that modelling we rely on the WIT framework that defines such socio-cognitive technical systems as a trinity of aspects (the social phenomenon, the simulation model and the implementation of that model). In this paper we centre our attention on the use of *affordances* as a useful construct to model socio-cognitive technical systems. We use the example of reputation emergence to illustrate our proposal.

1 Introduction

There is a rich discussion within the COIN[5] community about the properties and uses of open regulated multiagent systems that may be brought to bear upon the modelling of second order emergent phenomena (*EP*2). Such social phenomena involve agents that not only decide about their own actions but also about the actions of others and on the effect those actions have in the social environment where they interact. Although some *EP*2 have been explained as complex systems, it has been argued that agent-based simulation modelling may prove useful not only for explaining emergent features but also to understand motivational, strategic and organisational features that are ascribed

[5] COIN is the acronym Coordination, Organisations, Institutions and Norms, which has been adopted by a community of researchers, mostly within multiagent systems, who focus on these four topics. The COIN community typically organises two workshops each year leading to an annual volume of collected papers, published by Springer LNCS. The first COIN workshop took place in 2005 alongside AAMAS in Utrecht.

to the individuals involved in these phenomena and the outcomes of their activity within a given social environment.

The $\mathcal{WIT}$ framework is one way to analyse and describe those multiagent systems. The $\mathcal{WIT}$ framework postulates that coordination support frameworks for open regulated MAS are the amalgam of three aspects: (i) $\mathcal{W}$: a socio-technical system that constitutes actual coordination of a particular collective activity in the real *world*; (ii) $\mathcal{I}$: an abstract or *institutional* specification of the conventions that articulate the interactions in that system; and (iii) $\mathcal{T}$: the *technological* elements that implement the institutional conventions and enable the use of the system in practice. The $\mathcal{WIT}$ framework postulates also the type of relationships that should exist between those three aspects and how to characterise classes of socio-cognitive technical systems by linking $\mathcal{I}$ with $\mathcal{T}$ through the correspondence between metamodels for agents and social spaces and the platforms that implement those metamodels.

We claim that the use of the $\mathcal{WIT}$ framework provides the relevant foundations to deal effectively with the problem of modelling *EP*2. In this paper we use a specific example of the emergence of reputation to make a first step in this direction. Namely, when rumours about the behaviour of an individual circulate within a group, the reputation of that individual may change. Wwhen members of the group perceive that change, they may react by sending messages that reinforce or attenuate reputation change. Therefore, as in other *EP*2, the perceived signals influence the behaviour of individuals, which in turn influences how that reputation evolves.

Informed by the $\mathcal{WIT}$ framework, here we focus our attention on the abstract features that are needed to model both socio-cognitive agents and their social space. In particular we use the $\mathcal{WIT}$ framework (section 2) to elucidate the *affordances* required for modelling *EP*2. We approach this goal by working through three levels of refinement, each level being more specific than the previous. At the first, we put forward a primary list of *affordances* required for a generic *EP*2 (section 4). At the second we choose a second order emergent phenomenon – reputation – to explore, and informed by the primary list and the characteristics of the phenomenon we build a second, more specific, list of affordances (section 4.2). Finally, at the third level, we focus on a specific scenario that utilises the social phenomenon analysed at the second level. Again, using the primary and secondary lists, we build a third list that considers the particularities of the scenario (section 4.3). We conclude with a brief discussion of future work (section 5).

2 Socio-cognitive Technical systems. The $\mathcal{WIT}$ framework.

A socio-cognitive technical system (SCTS) is an open regulated multiagent system where agents – that may be human or software – interact in a shared virtual (online) space. We distinguish SCTS from other MAS by making explicit some assumptions about the agents that participate and the form that participation takes. To make this more precise we reproduce in Notion 1 the definition set out in [15]. We then use that as a starting point to put forward three (new) associated notions:

Notion 2: The social space in which a SCTS is situated and in particular the state of that social space that participants may perceive;

Notion 3: How the views that characterise the $\mathcal{WIT}$ framework can capture perspectives on SCTS, while providing a potentially helpful separation of concerns, as well as drawing attention to the interfaces between $\mathcal{W}$, $\mathcal{I}$ and $\mathcal{T}$;

Notion 4: How the "correct" interaction between $\mathcal{W}$, $\mathcal{I}$ and $\mathcal{T}$ leads to a definition of a coherent SCTS.

Notion 1 (SCTS) *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 [4].*
- **A.2.2** ***Opaque*** *The system, in principle, has no access to the decision-making models, or internal states of participating agents.*
- **A.2.3** ***Hybrid*** *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 only those actions that are mediated by a technological artefact that is part of the system may have effects in the system. Note that although such actions 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.*

¶

SCTS abound, and some typical examples are: (i) classical hybrid online social systems like *Facebook* [16], (ii) socio-cognitive technical systems like online public procurement systems and electronic institutions for various kinds of trading (e.g.

EverLedger's diamond provenance system) [1,9], (iii) massive on-line role playing games [27], and (iv) agent based simulation systems [27], in particular the like of those we discuss in sections 3 and 4.

A key feature of all SCTS, that is common to these examples, is that they are state-based systems, in the following sense:

Notion 2 (State of the social space) *A SCTS involves autonomous entities that interact in a common restricted environment that we call the social space, so that:*

B.1 *At any point in time the social space is in a "state" that consists of all the facts that hold in the social space at that point in time. Such state is unique and, therefore, common to all participants.*

B.2 *The state of the social space changes either through the actions of individuals that comply with the conventions that regulate the SCTS, or through events that are acknowledged by the STSC, even if those actions be unempowered or non-compliant.*[6]

¶

In order better to characterise SCTS and develop guidelines for their design, we proposed an abstract framework – the $\mathcal{WIT}$ *framework* [15] – whose distinctive contribution is the realisation that every SCTS can be understood as a composition of three "aspects": an actual functioning system in the real world ($\mathcal{W}$), the institutional description of the system ($\mathcal{I}$) and the technological artefacts that support the operation of the system ($\mathcal{T}$). This realisation provides a separation of concerns for each aspect that is convenient for description and design of SCTS (for an illustration of these claims see [16]). In Section 3, we show how these ideas apply to simulation systems.

As we suggested above, one can see the system that simulates a particular second order emergence phenomenon as a particular SCTS. In this case, $\mathcal{I}$ would the specification of a model of the given phenomenon, $\mathcal{T}$ the implementation of that specification and $\mathcal{W}$ would be the simulated emergent phenomenon. Thus, in $\mathcal{W}$ one deals with issues concerned with the proper implementation of data structures and algorithms; as well as the interfaces that allow the visualisation of the simulated phenomenon. In $\mathcal{I}$one is concerned with the expressiveness of the formalism used to model emergent phenomena and whether the understanding that one has of the social phenomenon is faithfully transcribed in that formalism. Finally, $\mathcal{W}$ is the simulated phenomenon one wishes to study and therefore one is concerned with the means to define the variable behaviour of agents (human or artificial) and the exogenous events and how to interpret outcomes of those interactions.

These intuitions are firmed up in the next set of definitions (cf. [15]). Notion 3 says that the three views may be characterised by their core ontologies, a *compatibility* relationship and their particular notion of state. Notion 4 states that the three compatibility notions are "aligned" so that the state of the three aspects evolve coherently.

[6] We mean exogenous events that affect the behaviour of the system *in a relevant way* and should therefore be accounted for in the description and implementation of the system. For example, rainfall, a new exchange rate, the passage of time.

Notion 3 ($\mathcal{WIT}$ views) *The* $\mathcal{WIT}$ framework characterisation *of a SCTS* $\mathcal{S}$ *is the triad* $\langle \mathcal{W}, \mathcal{I}, \mathcal{T} \rangle$*, where:*

C.1 $\mathcal{W} = \langle W, \succ \rangle$*, is the view of* $\mathcal{S}$ *as a running system situated in the (real) world. It comprises:*
C.1.1 *A domain ontology* W*, that captures the intuition that only certain facts, events and actions that happen in the physical world are relevant for the system;*
C.1.2 *The* $\mathcal{W}$*-compatibility relationship,* $\succ$*, corresponds to the intuition that relevant actions are "feasible" in* $\mathcal{W}$*, only if the proper conditions hold, and if a relevant action is feasible its effects will be relevant as well;*
C.1.3 $(\mathcal{S}_{\mathcal{W}t})$*, the state of* $\mathcal{W}$ *at time* t *, is the set of all facts that are relevant in* $\mathcal{W}$ *at time* t*:*

$$\mathcal{S}_{\mathcal{W}t} = \{\alpha \mid W \succ \alpha\} \tag{1}$$

C.2 $\mathcal{I} = \langle I, \propto \rangle$*, the institutional view of* $\mathcal{S}$ *is the abstract representation of the system and the conventions that govern the actions that may take place in* $\mathcal{W}$ *and their effects. It comprises:*
C.2.1 *An institutional ontology* I *that captures the intuition that the institutional representation of* $\mathcal{S}$ *involves an ontology formed by "institutional" assertions and actions that correspond to the relevant facts, events and actions in* W*;*
C.2.2 *The* $\mathcal{I}$*-compatibility relationship* $\propto$ *picks up the intuition that attempted institutional actions will be "admissible" in* $\mathcal{I}$*, only if they comply with the prevailing conventions; and when an attempted action is admitted, its effects will be admitted in* $\mathcal{I}$ *as well.*
C.2.3 *The state of* $\mathcal{I}$ *at time* t*, is the set of all expressions that are admitted ("hold") in* $\mathcal{I}$ *at time* t*:*

$$\mathcal{S}_{\mathcal{I}t} = \{\psi \mid W \propto \psi\} \tag{2}$$

C.3 $\mathcal{T} = \langle T, \bowtie \rangle$*, the technological view of* $\mathcal{S}$ *is the implementation of the system according to* $\mathcal{I}$ *that receives inputs from and produces outputs in* $\mathcal{W}$*. It includes:*
C.3.1 *a collection of data structures of the implementation of* $\mathcal{S}$ *whose values change when an "acceptable" input is processed in* $\mathcal{T}$*.*
C.3.2 *The* $\mathcal{T}$*-compatibility relationship* $\bowtie$ *catches the intuition that the values of some variables change when the system processes an acceptable input.*
C.3.3 *The state of* $\mathcal{T}$ *at time* t*, is the set of values of the relevant variables in* $\mathcal{T}$ *at time* t*:*

$$\mathcal{S}_{\mathcal{T}t} = \{\phi \mid W \bowtie \phi\} \tag{3}$$

¶

An important feature of the $\mathcal{WIT}$ characterisation is that one would like to express that only those actions that are *compatible* with the conventions of the system can change the state of the system. For that purpose we need to establish some sort of alignment between actions in $\mathcal{W}$, $\mathcal{I}$ and $\mathcal{T}$and use the three *compatibility relationships* $(\succ, \propto, \bowtie)$ to indicate that the corresponding state changes if only if the attempted action

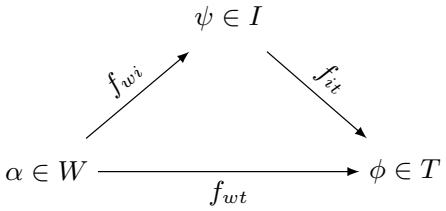

Fig. 1. The three "bijections" of Notion 4

is compatible with the prevailing state of the context. In particular, we postulate that if an SCTS is properly specified and deployed, the three $\mathcal{WIT}$ views are "coherent" in the sense that their corresponding states evolve as intended. In other words, when an action is attempted, in $\mathcal{W}$ –which is expressed as an attempted input in $\mathcal{T}$ – its effects in $\mathcal{W}$ should be the ones prescribed in $\mathcal{I}$, which ought to be the ones that are computed in $\mathcal{T}$ and are reflected in $\mathcal{W}$, as pictured in Fig. 1. The following notion approximates such alignment:[7]

Notion 4 (Coherence) *Let f_{wi}, f_{it} and f_{wt} be three "bijections" between the $\mathcal{WIT}$ views of a SCTS $\mathcal{S}$; and let α, ψ and ϕ be actions in $\mathcal{W}$, $\mathcal{I}$ and $\mathcal{T}$, respectively, such that $\psi = f_{wi}(\alpha)$ and $\phi = f_{it}(\psi)$ and $\phi = f_{wt}(\alpha)$.*

The $\mathcal{WIT}$ views are coherent *iff for every time t,*

$$(\mathcal{S}_{\mathcal{W}t} \succ \alpha) \Leftrightarrow (\mathcal{S}_{\mathcal{I}t} \propto \psi) \Leftrightarrow (\mathcal{S}_{\mathcal{W}t} \bowtie \phi) \tag{4}$$

¶

It is worth noting that beyond the mapping of actions and effects that support the coherence of the three views, there are other relationships between views as depicted in Fig. 2. The following remarks give an indication of what these relationships stand for. Although we will not deal with these matters in detail here, we should note that they support design and methodological concerns (as suggested in [16]). In that spirit we illustrate the interrelationship between views in Sec. 3.

D.1 We call the $\mathcal{I}$ view *institutional* following the usage of Searle [23]. Thus we expect to have a bottom-up "corresponds" relationship from $\mathcal{W}$ to $\mathcal{I}$ that serves to create the "institutional reality". This is usually achieved through "constitutive norms" that transform (and legitimise) relevant brute facts and actions into the "corresponding" institutional facts and actions.

[7] In Notion 4 we postulate that the views are coherent when they are sort of *isomorphic*. This is an elusive concept in the sense that unless one has a precise specification of each view it is impossible to define the intended "bijections". However, the alignment can be made precise when one has a precise description of the domain language used in $\mathcal{W}$, the corresponding action, norm and communication languages used in $\mathcal{I}$; and, in turn how those are transcribed into actual code in $\mathcal{T}$ through some specification language. See [9,14] for an example.

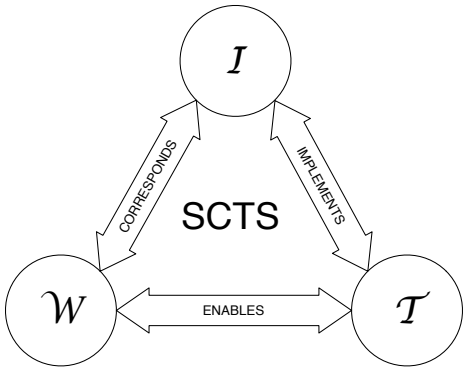

Fig. 2. The WIT trinity: The ideal system, $\mathcal{I}$; the technological artefacts that implement it, $\mathcal{T}$, and the actual world where the system is used, $\mathcal{W}$.

D.2 The intended coherence between the two aspects also entails a top-down "corresponds" relationship that converts (or anchors) the institutional effects of institutional facts and actions into the corresponding relevant brute facts and action. Thus, it also works as a prescriptive relationship (from $\mathcal{I}$ to $\mathcal{W}$).

D.3 Notice that the "corresponds" relationships presume that the representation in $\mathcal{I}$ of $\mathcal{W}$ is adequate (all relevant entities are properly represented and all pertinent institutional entities are properly reflected in relevant brute entities).

D.4 Once $\mathcal{I}$ is understood as a *prescription* of the intended behaviour of $\mathcal{S}$, it is used to *specify* the software that implements it. Thus the top-down "implements" relationship. Conversely, the actual behaviour of the implemented system in $\mathcal{T}$ should comply with the institutional conventions in $\mathcal{I}$.

D.5 Notice that the "implements" correspondence presumes that the specification is accurate and the implementation correct.

D.6 $\mathcal{T}$ enables $\mathcal{W}$ because Notion 1 postulates that all STSC are online systems. Thus, every relevant event that takes place in $\mathcal{W}$ and any action that is attempted in $\mathcal{W}$ may affect the state of the $\mathcal{S}$ only, when wrapped as a message, it is deemed a valid input in $\mathcal{T}$. Conversely, changes in the state of $\mathcal{S}$ become actual brute facts in $\mathcal{W}$ if they are presented as outputs from $\mathcal{T}$.

D.7 Notice, finally, that those input-output connections between $\mathcal{W}$ and $\mathcal{T}$ presume that information is not lost or corrupted, that interfaces are ergonomic and correct, and that transfer of information is made according to the conventions stipulated in $\mathcal{I}$.

3 Simulation of *EP*2 with the $\mathcal{WIT}$ framework

In broad terms, we want to build simulation systems to study second-order emergent social phenomena. As discussed in section 4, these phenomena involve individuals that may recognise that a macro phenomenon is emerging and, as a consequence, this phenomenon and the emergence process itself can be intentionally supported, initiated,

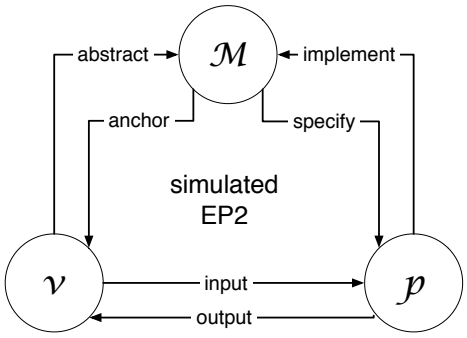

Fig. 3. A refinement of the WIT trinity for the simulation of second order emergence phenomena. The $\mathcal{W}$ view becomes a simulated virtual world $\mathcal{V}$ where one studies the emergent phenomenon, $\mathcal{I}$ becomes the conceptual model of the social phenomenon, $\mathcal{T}$ is, now, the implementation of the model that runs the virtual world and interrelationships between views are instrumental.

changed or contrasted by the same individuals. In other words, individual agents decide what to do in view of their own motivations and preferences but also taking into account what others may or may not do and the effects of their own actions and the actions of others. Thus, in order to model *EP*2, we need socio-cognitive agents. Moreover, since these agents do not act in a void but in a social environment that provides them with cues, opportunities and means to interact with other agents, we also need a persistent, regulated social space. In fact, since all the assumptions we postulate in Notion 1 apply to the systems where of second order phenomena emerge, we may use the $\mathcal{WIT}$ framework to characterise these systems.

Indeed, once we commit to a SCTS representation of the social phenomenon, a rough $\mathcal{WIT}$ characterisation is straightforward: the $\mathcal{I}$ view is the abstract model of a social phenomenon (we'll refer to it as $\mathcal{M}$) and $\mathcal{T}$ is the corresponding working computational model (we'll call it $\mathcal{P}$). Finally, $\mathcal{W}$ is the simulated (virtual) environment ($\mathcal{V}$) where one inputs experimental data and observes the social phenomenon.[8]

We may get a more refined characterisation of simulation systems by qualifying the relationships between $\mathcal{V}$, $\mathcal{M}$ and $\mathcal{P}$. Figure 3 adapts our original $\mathcal{WIT}$ trinity (Fig. 2) to simulation, and splits in two each of the relationships between views in order to clarify the character of those relationships when the framework is used for simulation.

The process of design and construction of a simulation is (as usual) a cyclic process that normally starts (i) with a vague understanding of the phenomenon that is (ii) reflected in an abstract model ($\mathcal{M}$), which is in turn (iii) implemented ($\mathcal{P}$) to produce (iv) the virtual world ($\mathcal{V}$) where actual simulation runs take place, and subsequently (v) the simulated phenomenon is progressively refined by testing the implementation of

[8] Experimental data inputs consist of an initial state —including a population of agents with their own profiles and data— that is loaded into $\mathcal{P}$, and then events —generated somehow— and actions taken by agents. By extension, the presence of human actors in $\mathcal{V}$ would make this a participatory simulation.

the model through the virtual world. What is distinctive about the WIT approach can be summarised along the following three lines:

1: Metamodels and affordances. We postulated in [15] that a metamodel consists of *a collection of languages, data structures and operations that serve to represent* the agents and the social space of a given SCTS with an appropriate level of detail and accuracy. The model, hence, would be a *representation* of a phenomenon through a particular abstract or symbolic notation specific of the metamodel. A notation that, consequently, will be useful as long as it has the expressiveness needed to capture the relevant features of the phenomenon at the appropriate level of detail.

The closer the metamodel is to the phenomenon one wishes to understand, the easier it is to instantiate it. The more distant, the wider the variety of systems that would be fit to model, and the more detailed the effort of tailoring to the particular system. One way to make this proximity precise is to refer to the "affordances" of the metamodel [17]. In [15], we defined an *affordance* to be a property that enables coherent change of states in a SCTS. Namely,

Notion 5 *An* affordance *is a property of the SCTS (of individual agents or of social space) that supports effective interactions of agents within an SCTS.*

We postulated three essential affordances of every SCTS: *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, *Coordination*, so that the actions of individuals are conducive to the collective endeavour that brings them to participate in the SCTS and *Validity* that preserves the proper correspondences of the tripartite view.

Those affordances may be achieved through several means and will be reflected in the features that can be directly expressed by the languages and constructs of the metamodel. Those features include, for instance, the description of the entities ("ontology") that are involved in the representation of a state of the system, the primitive actions that agents may take, the way actions are taken by agents and their are reflected and perceived in the social space, the possibility of organising certain interactions in a subcontext of the whole social space, whether the conventions that regulate interactions are regimented or may be enforceable and through what means, etc.[9]

2: Platforms. The implementation of the model is facilitated when that implementation is associated with the metamodel. This is the purpose of developing a suite of software tools – a *platform* – that is powerful enough to capture all the distinctive features of the metamodel and consequently enables the designer to move smoothly from a precise instantiation of the model to the code that runs it. The ideal situation would be to have a specification language that is used to make the model precise and generates the corresponding executable code.

There are several proposals for metamodels for socio-cognitive technical systems and a few of them are accompanied by a corresponding platform; see [1] for thorough descriptions of some of the most developed and examples of their application.

[9] See Sec.4 for a more detailed list.

3: Methodological considerations. An important goal for coupling metamodel and platform is that one can get assurances about the correctness of the implementation and the completeness of the specification.

Notwithstanding that interplay, one is still confronted with the choice of platform and metamodel and making sure that correctness and completeness hold. There is an extended discussion of this matter, with respect to socio-cognitive technical systems in [13] and a complementary one in [15]. Sec. 4 deals with these matters.[10]

We do not yet have a metamodel for modelling directly second order emergence and we find no platform that is convenient enough to model *EP*2 specifically.

Ithe next section we take the first steps in that direction, following a top-down approach. Thus, rather than trying to adapt an available metamodel & platform framework like electronic institutions [9] – that is too general – we proceed towards a rather specific phenomenon (a scenario where reputation emerges among a closed group of individuals through the exchange of a given class of messages (Sec. 4.3)) and identify those features and affordances that are needed for a convenient representation, starting from affordances for *EP*2 and reputation in general. From these we aim to develop formalisms and specification languages that make those affordances operational. Similarly, we will start from an ad-hoc implementation of the affordances towards a platform that is closely linked to the resulting metamodels.

4 Affordances for modelling second order emergence

This section describes a three-stage top-down process to uncover tentative lists of individual and social space affordances, firstly at a generic level (section 4.1), second at the level of a class of particular phenomena, namely reputation (section 4.2), and thirdly in the case of a specific reputation model (section 4.3). We emphasise this is not the only such answer: its purpose is primarily to illustrate how one might go about affordance identification, rather than being definitive either about process or outcome.

4.1 Second order emergence

At the core of the old debate on micro foundations (individualism) versus macro properties (structuralism) of societal systems – also known as the micro-macro link problem – we find the notion of *emergence* and how the micro and macro levels interact. Specifically we have to differentiate between two different approaches to the *emergence* of social phenomena.

Following a *generativist paradigm* [10], we can approach the emergence of social phenomena as a process that goes from micro to macro, from the individuals and their local behaviour to the macro structures that *emerge* as a result of the local interactions. In this approach

[10] In [27] we elaborate on the convenience of separating design ($\mathcal{M}$) and implementation ($\mathcal{P}$) concerns and also the advantage of building a metamodel that facilitates design, and a corresponding platform that supports implementation. We also discuss the advantage of having a "design environment" to deal separately with the definition and management of simulations.

> "the only action takes place at the level of individual actors, and the 'system level' exists solely as emergent properties characterising the system of action as a whole." [5]

This is known as *first order emergence* and is the main approach followed in current state-of-the-art social simulations:

> "Given some macroscopic explanandum – a regularity to be explained – the *canonical agent-based experiment*[11] is as follows: Situate an initial population of autonomous heterogeneous agents in a relevant spatial environment; allow them to interact according to simple local rules, and thereby generate – or 'grow' – the macroscopic regularity from the bottom up."[10]

This, however, is only half the story. To what extent do macro-level properties exercise some kind of causal influence on the micro-level individuals' behaviour? [6]. In many cases, in a real human society, many of the macro structures that start to appear as a result of the individual's local behaviour have an effect on macro-level attributes (for example, the creation of ghettos may imply the increase of the crime rates and, as a consequence, devaluation of houses in that area[12]). The modification of those macro-level attributes, at the same time, has an effect in the individual's local behaviour modifying it (what is known as a *'downward causation'* [6]). This change in the individual's behaviour influences again how the macro structures emerge; how the emergence of the new macro structures modify the macro-attributes; and so on.

The scenario is even more complex if we consider that individuals may recognise that the phenomenon is emerging and, as a consequence, this phenomenon (and the emergence process itself) can be intentionally supported, maintained, changed or contested by the same agents. This is what is known as *second-order emergence*, a feature that characterises many important social phenomena. Examples of these phenomena range from social movements like the African-American civil rights movement, the Arab spring or the 15-M movement in Spain, to relevant social constructs like *reputation*, which is the basis of the exercise in Sections 4.2 and 4.3.

Affordances for second order emergence

What are the generic affordances that allow (are necessary for) *second order emergence*? As we have said above, the main characteristic of *second order emergence* is the capacity of the individuals at the micro level to detect that the social phenomenon

[11] When we talk about social simulation we have to talk invariably about *agent-based social simulation* (ABSS). The main characteristic of a social simulation is that the simulated *individuals* are not entities whose aggregated behaviour can be adequately described using mathematical equations. Every individual is unique and interacts with the other individuals and the environment in an autonomous way. This particularity is what makes the multiagent systems paradigm the predominant approach in social simulation nowadays. From now on, we will use the terms social simulation and agent-based social simulation interchangeably.

[12] We only make reference to Schelling's dynamics example for sake of reader familiarity, rather than to engage in debate about its appropriateness or correctness.

that will show up in the macro level is starting to emerge. This means that the individuals (or at least some of them) know about the existence of that phenomenon and, more importantly, know about the signals that identify its emergence in a given society. On the one hand, the *social space* makes more or less explicit these signals to the individuals. On the other hand, individuals need to have the capacity to perceive them and again, more importantly, of interpreting them as indicators of the emergence of the social phenomenon. Invariably, this goes through the capacity to anticipate what the other individuals will do in the future, in other words, the individual has to operate with a theory of mind. Theory of Mind is "the ability to understand others as intentional agents [8], and to interpret their minds in terms of intentional concepts such as beliefs and desires" [12]. Having a theory of mind has been recognised by several authors as a fundamental requirement of an architecture of the social mind [3,26].

The detection of the emergence of a social phenomenon is only the first stage of *second order emergence*. Once the individuals at the micro level become aware of the emergence process, they should have the capacity to influence it. This implies some kind of capacity for action embedded in the individual that at the same time is facilitated by the *social space*.

That said, a tentative list of generic affordances necessary for a *second order emergence* scenario can be summarised as follows:

Individual affordances

1. Cognitive capabilities to understand the emergent social phenomenon.
2. Theory of mind. Anticipate what others intend to do, how they will do it and what are they motivations.
3. Sensor capabilities to detect the signals that the *social space* makes available and that are associated with the emergence of the social phenomenon.
4. Cognitive capabilities to interpret the signals as indicators of the emergent process.
5. Actuator capabilities to influence the emergent process.

Social space affordances

1. A shared ontology of objects, agents, actions and events.
2. Some sort of social model to represent roles, groups, organisations and their relationships.
3. Some sort of governance or coordination support.
4. Perception channels adapted to the sensor capabilities of the individuals.
5. Actuation channels adapted to the actuator capabilities of the individuals.

4.2 Reputation

While the affordances we enumerated are generic for modelling second order emergence phenomena (*EP*2), if one wants to model a specific phenomenon, one may profit from the availability of affordances that are specific to the particular phenomenon. Thus we look into a well-known social construct: reputation.

Reputation can be defined as "what a *social entity says* about a target regarding his/her/its behaviour and characteristics". A social entity is "a group which is irreducible to the sum of its individual members, and so must be studied as a phenomenon

in its own right" [22]. The definition postulates that whoever is saying something about the target is not an individual, but a social entity. An individual is just a messenger of what is supposed to be the opinion of the social entity (in fact, the messenger does not even have to be a member of that social entity to spread a reputation). This is a key aspect because it allows reputation to be an efficient mechanism to spread social evaluations by reducing fear of retaliation [20].

The next important element in the definition above is the action of "saying". Reputation exists because an evaluation circulates. Without communication, reputation cannot exist. You can have the members of a community sharing a belief. This belief however is not a reputation until it starts to circulate. In fact, communication is so important for reputation that there is a specific type of communication specialised for building reputation values: *gossip*.

When messages start circulating and people realise that a reputation on a target is starting to form, many times they will start performing actions (in the form of new rumours, support messages, shame messages, etc.) that are intended to influence the formation of that reputation. Therefore, as in any second order emergent phenomenon, the perceived signals that a reputation is emerging influence the behaviour of the individuals, that at the same time influence how that reputation emerges.

Affordances for reputation

First of all, the individual needs to have a reputation model. Our goal is that this model has to go beyond the traditional computational models of reputation [21] that focus only on how reputation is evaluated. The individual has to be able to influence reputation, so it has to know how it spreads (how gossip works), how it is evaluated and what are the elements that lead to the emergence of reputation or its undermining. Notice that this level of knowledge about reputation requires a theory of mind (when will other individuals spread a reputation value?, who will be receptive to a specific reputation value?). It is also important that the individual knows about the utility of reputation: what is it good for? How can reputation favour/limit the achievement of my goals?

From the previous definition of reputation, it is clear that the nentity social group is essential for reputation and needs to be present at both levels, individual and *social space*. An individual needs to be able to detect social groups and determine the membership to those groups. At the same time, the *social space* can make more or less explicit this membership to the rest of members of the society. Linked to this capacity and as part of the reputation model, the individual has to be able to understand social relations and how they influence reputation and its spreading.

Finally, we reiterate that reputation depends on communication, so the individual has to be able to communicate with other individuals and the *social space* has to enable and support this communication.

Our proposed tentative list of affordances at this level of abstraction is the following:

Individual affordances

1. A [complete] model of reputation (including a "reputation oriented" theory of mind).

2. Notion of group. Capacity to detect groups. Understanding of social relations.
3. Capacity to communicate with other individuals (receive and send messages).

Social space affordances

1. Support for group formation and identification.
2. Communication channels.
3. Messages of different types.

4.3 Reputation scenario

After identifying the affordances for second order emergence in general (section 4.1) and those for the specific and illustrative second order phenomenon of reputation (section 4.2), the next level of concretisation in our exercise is to identify the affordances associated to a specific scenario related with the social phenomenon. A scenario is a particular environment (that can include a physical space, a set of possible actions, behavioural restrictions, etc.) where the social phenomenon is present and relevant. The scenario that we will use to illustrate this third step is an idealised environment to study the spread of rumours and the formation of reputation. Notice that this is one of many possible scenarios and that the affordances identified at this level are strongly related to the particularities of the scenario.

The individuals in our scenario are directed by motivations. Each individual has a set of basic needs that she tries to satisfy. The set of needs that are relevant for a specific agent determine its personality and the kinds of actions the individual is motivated to perform in the world. In our scenario, the kinds of actions that an individual can perform are actions that influence the reputation of others.

The world where the agents evolve is divided into what we call *social contexts*. A *social context* is a physical space where individuals perform a social activity. For example, your home is a *social context* where you interact with the individuals that belong to your family in domestic activities, the gym is a *social context* where you interact with people that, like you, enjoy practising sport. Each *social context* has different characteristics that facilitate or restrict social interaction.

In our scenario, at every turn individuals are randomly assigned to a *social context*. Once in a *social context*, an individual can approach or avoid other individuals present in that *social context*. We want to simulate the dynamics of individuals that have different motivations to approach or avoid other individuals in the same *social context*. These dynamics take into account the preferences of each pair of individuals. First, all the individuals express their intention to approach or avoid other members present in the *social context*. Second, given these intentions the system calculates the *communication groups* (groups of individuals that at some moment will be together to exchange messages) that will be formed in that *social context*, for a given pair of agents (A, B), using the following rules:

1. If one of the two agents has explicitly expressed its intention to avoid the other, the system will take care that they never meet. This simulates the situation when an individual wants explicitly to run away from another.

2. If A wants to approach B and (i) B also wants to approach A or (ii) B has not expressed any intention related with A, the system will place the agents in a common *communication group*. This simulates the situation when an individual wants to approach another individual and the latter either agrees on that approach or she is indifferent.
3. If neither A nor B have expressed any intention related to the other, the system will randomly decide to place them in a common *communication group* or not. This simulates the chance approach of one individual to another.

Notice that an individual can be in more than one *communication group* because group communication does not take place at the same time, but rather in sequence (see Algorithm 1). As an example, consider agents A, B and C. We have the following intentions (A approach B) (A approach C) (B avoid C). In this scenario, the system will generate two *communication groups*: [A,B] and [A,C] and will never generate a communication group with both B and C (see Algorithm 1 for the details about how we calculate the *communication groups*).

Individuals in a *communication group* can exchange messages (*rumours*) and can listen to the messages exchanged by the other individuals in that group. As a result of a received or heard *rumour*, an individual can react and send a *support* message (reinforcing the original *rumour*) or a *shame* message (expressing her disapproval of the original *rumour*). The message-reaction cycle is repeated until all the individuals in that *communication group* have had the opportunity to send a message, after which the group is dissolved. When all the groups are dissolved, the system asks again about the intentions of approaching or avoiding other individuals in that *social context* and this generates a new set of *communication groups*. This is repeated *n* times, after which the system starts a new turn. The sequence of a turn in the reputation scenario is illustrated in Algorithm 1.

Affordances for the reputation scenario

Our proposed list of affordances at this level of abstraction follows the guidelines established at the previous level (section 4.2) taking into account the specific scenario described above:

Individual affordances

1. Agent architecture directed by motivations with a "reputation oriented" theory of mind.
2. Capability to decide which individuals to avoid or to approach (according to the individual's internal motivations and the personality of individuals in the communication group).
3. Reasoning mechanisms to decide when to send a $\{$*rumour* $\|$ *support* $\|$ *shame*$\}$ message (according to the individual's internal motivations an the personality of individuals in the communication group).
4. Capability to send a $\{$*rumour* $\|$ *support* $\|$ *shame*$\}$ message.

Data: *SocialContexts*: Set of social contexts; *Agents*: Set of agents;

Each agent in *Agents* is assigned randomly to a social context in *SocialContexts*;
foreach *SC in SocialContexts* **do**
  repeat
    Each agent in *SC* evaluates which other agents in *SC* wants to *approach*, to *avoid* or is indifferent about it (*neutral*);
    The environment collects from each agent the list of intentions (*approach* | *neutral* | *avoid*) towards the rest of the agents in *SC*;
    The environment assigns a *distance* between each pair of agents in *SC* according to the following table:

intention(A $\rightarrow$ B)	intention(B $\rightarrow$ A)	distance(A,B)
approach	*approach* | *neutral*	0
approach | *neutral* | *avoid*	*avoid*	1
neutral	*neutral*	random(0,1)

    The environment creates the communication groups that will happen in that *SC* Taking the graph where the nodes are the agents in *SC* and the edges connect any pair of agents at distance 0, the *communication groups* are defined as the maximal cliques of that graph.
    foreach *CG in SC* **do**
      repeat
        The environment chooses randomly one agent from those that want to send a *rumour*;
        The selected agent sends the *rumour*;
        The other agents send reactions to that *rumour* till no one has anything to say;
      until *no agent wants to send a new* rumour;
    end
  until *n times*;
end

Algorithm 1: A turn in the reputation scenario.

Social space affordances

1. Creation of *social contexts*.
2. Creation of *communication groups*.
3. Make explicit to each agent in a *social context* which are the other members of the society in the same *social context*.
4. Make explicit to each agent in a *communication group* the other members of the society in the same *communication group*.
5. Enable movement of individuals according to the scenario rules.
6. Communication channel between agents that belong to the same communication group.
7. Enforce the communication protocol in a *communication group*.

5 Closing remarks

Our long-term aim is the creation of a conceptual model, leading to a family of computational frameworks, that can support the creation and exploration of complex socio-cognitive technical systems. In this paper we begin to address the questions arising from how to observe, understand and model the ways in which actors engage with social processes, for which they do not necessarily have existing conventions or norms to guide their behaviour, and which by their actions affect the emergence and properties of the nascent process. We put forward the $\mathcal{WIT}$ framework as a way to structure the dissection and understanding of three perspectives on the action space, coupled with a three step methodology for the refinement of the individual and social space affordances, taking reputation as the target social phenomenon for this particular exercise.

In earlier work [15,16], our focus was on the framework alone, as we sought to establish the characteristics of the perspectives of $\mathcal{W}$orld, $\mathcal{I}$nstitution and $\mathcal{T}$echnology and their inter-relationships, as set out in Notion 1. This in turn was informed by our experience in developing electronic institutions – from the earliest conceptual versions [11] to its current metamodels and implementation platforms and numerous applications [14,19] – which give us confidence that the $\mathcal{WIT}$ approach is sufficient to the task and also that we should aim for an *EP2* metamodel that is powerful (capture a large class of *EP2*), intuitive (so non-experts can use it to simulate *EP2*) and easy to use. The additional notions of the state of the social space (Notion 2), the views afforded by $\mathcal{WIT}$ (Notion 3) and coherence between those views (Notion 4), flesh out the framework in order to focus on how the framework may be applied to simulation in a straightforward manner.

We have sought to illustrate our exploration of *EP2* by taking the case of reputation. First, because it is an *EP2* that is well-known to social scientists and also one with which we already have experience. Second, because, as was the case with auctions, we believe that it contains archetypal *EP2* features. Thus, our expectation in choosing it, is that it can take us towards a conceptual framework that is generic enough to be applied for modelling of a wide class of *EP2*, as well as specific enough for particular *EP2*, and that it is practical for implementing *EP2* although it might be quite impractical for modelling other social coordination artefacts.

We illustrate our analysis of reputation through the three-step process outlined in Section 4, which takes a top-down route from generic (for *EP*2 in general) through class (for reputation in general) to instance (for a specific reputation model), to identify the individual and social space affordances that appear to be sufficient in the case as presented here, while also providing pointers for future work and other case studies.

Related work is rather widely spread across the social sciences and computer science, with more in the former than the latter, making a comprehensive discussion quite challenging. Given the preliminary nature of the vision presented, we instead therefore highlight those works that have inspired us, and whose threads we are attempting to draw together in this paper, which have been cited at various places throughout the discussion, such as [8,3,25,10], while regretting the omission through ignorance of doubtless many other works.

No discussion of institutional frameworks is complete without reference to the management of Common Pool Resources (CPRs) and the work of Ostrom [18], and although

the focus of the case studies is typically physical, natural resources, the principles for enduring institutions can, with some thought, be applied to the intangible, such as reputation, as discussed here. While Ostrom plots the emergence of CPR institutions through narrative, the focus is primarily on the institution that emerges, and rather less on the process of emergence itself, and the factors that facilitate or inhibit that emergence – although they do receive coverage. The consequent ADICO framework [7] also tends to focus on capturing the static end result, although there are elements that could be developed and applied to the earlier steps we discuss in Section 4, which we will explore as part of future work.

Shove et al. dissect the notion of social practice [24], putting forward an analysis in which actors utilise what they know, mean or intend to reconfigure available resources (including other actors) to achieve accepted outcomes. This offers an interesting parallel with pattern languages [2], which work in similar ways to achieve design goals. Both appear to leave aside the matter of how patterns emerge, but it seems quite plausible that the affordances at the generic and the class level might give rise to patterns or practices for the creation of patterns (or practices), which again will be considered as part of future work. The nature of the phenomena studied in [18] and [24] describe emergence of institutions in the realm of finite resource distributions, in essence a zero-sum game between the stakeholders. Our example of reputation is neither a zero-sum game nor a distribution of finite resources and as such may have different affordance characteristics. Thus, future research will even extend to different social situations.

Acknowledgements

This research has been partially supported by project MILESS (Ministerio de economía y competitividad - TIN2013-45039-P - financed by FEDER) and SCAR project (Ministerio de economía y competitividad - TIN2015-70819-ERC). We also thank the Generalitat de Catalunya (Grant: 2014 SGR 118).

References

1. Huib Aldewereld, Olivier Boissier, Virginia Dignum, Pablo Noriega, and Julian Padget. *Social Coordination Frameworks for Social Technical Systems*. Number 30 in Law, Governance and Technology Series. Springer International Publishing, 2016.
2. Christopher Alexander, Sara Ishikawa, Murray Silverstein, Joaquim Romaguera i Ramió, Max Jacobson, and Ingrid Fiksdahl-King. *A pattern language*. Gustavo Gili, 1977.
3. Cristiano Castelfranchi. Cognitive architecture and contents for social structures and interactions. In Ron Sun, editor, *Cognition and Multi-Agent Interaction*, pages 355–390. Cambridge University Press, 2006.
4. Cristiano Castelfranchi. *InMind and OutMind; Societal Order Cognition and Self-Organization: The role of MAS.* http://www.slideshare.net/sleeplessgreenideas/castelfranchi-aamas13-v2, May 2013. Invited talk for the IFAAMAS "Influential Paper Award". AAMAS 2013. Saint Paul, Minn. US.
5. J. S. Coleman. *Foundations of Social Theory*. Belknap Press, 1990.
6. Rosaria Conte, Giulia Andrighetto, Marco Campennì, and Mario Paolucci. Emergent and immergent effects in complex social systems. In *Proceedings of AAAI Symposium, Social and Organizational Aspects of Intelligence*, pages 8–11, 2007.

7. Sue ES Crawford and Elinor Ostrom. A grammar of institutions. *American Political Science Review*, 89(3):582–600, 1995.
8. Daniel Dennett. *The Intentional Stance*. MIT Press, 1989.
9. Mark d'Inverno, Michael Luck, Pablo Noriega, Juan A. Rodriguez-Aguilar, and Carles Sierra. Communicating open systems. *Artificial Intelligence*, 186(0):38 – 94, 2012.
10. Joshua M. Epstein. *Generative Social Science: Studies in Agent-Based Computational Modeling*. Princeton University Press, 2006.
11. Marc Esteva, Julian Padget, and Carles Sierra. Formalizing a language for institutions and norms. In Jean-Jules Meyer and Milinde Tambe, editors, *Intelligent Agents VIII*, volume 2333 of *Lecture Notes in Artificial Intelligence*, pages 348–366. Springer Verlag, 2001. ISBN 3-540-43858-0.
12. Maaike Harbers, Karel van den Bosch, and John-Jules Ch. Meyer. Modeling agents with a theory of mind: Theory-theory versus simulation theory. *Web Intelligence and Agent Systems*, 10(3):331–343, 2012.
13. Andrew J. I. Jones, Alexander Artikis, and Jeremy Pitt. The design of intelligent socio-technical systems. *Artif. Intell. Rev.*, 39(1):5–20, 2013.
14. Pablo Noriega and Dave de Jonge. Electronic institutions: The ei/eide framework. In *Social coordination frameworks for social technical systems*, pages 47–76. Springer, 2016.
15. Pablo Noriega, Julian Padget, Harko Verhagen, and Mark d'Inverno. Towards a framework for socio-cognitive technical systems. In *Coordination, Organizations, Institutions, and Norms in Agent Systems X*, volume 9372 of *Lecture Notes in Computer Science*, pages 164–181. Springer International Publishing, Berlin / Heidelberg, 2015.
16. Pablo Noriega, Harko Verhagen, Mark d'Inverno, and Julian Padget. A manifesto for conscientious design of hybrid online social systems. In S. Cranefield, S. Mahmoud, J. Padget, and A.P. Rocha, editors, *Coordination, Organizations, Institutions and Norms in Agent Systems XII*, Lecture Notes in Computer Science. Springer, In press.
17. Donald A. Norman. Affordance, conventions, and design. *interactions*, 6(3):38–43, May 1999.
18. Elinor Ostrom. *Governing the commons*. Cambridge university press, 1990.
19. Julian Padget, Emad ElDeen Elakehal, Tingting Li, and Marina De Vos. *InstAL: An Institutional Action Language*, pages 101–124. Springer International Publishing, Cham, 2016.
20. Isaac Pinyol, Mario Paolucci, Jordi Sabater-Mir, and Rosaria Conte. Beyond accuracy. reputation for partner selection with lies and retaliation. In *Multi-Agent-Based Simulation VIII*, volume 5003, pages 128–140. Springer, 2008.
21. Isaac Pinyol and Jordi Sabater-Mir. Computational trust and reputation models for open multi-agent systems: a review. *Artificial Intelligence Review*, 40(1):1–25, 2013.
22. DH. Ruben. The existence of social entities. *Philosophical Quarterly*, 32:295–310, 1982.
23. John R. Searle. What is an institution? *Journal of Institutional Economics*, 1(01):1–22, 2005.
24. Elizabeth Shove, Mika Pantzar, and Matt Watson. *The dynamics of social practice: Everyday life and how it changes*. Sage, 2012.
25. Flaminio Squazzoni. The micro-macro link in social simulation. *Sociologica*, (1/2008), 2008.
26. Ron Sun. Desiderata for cognitive architectures. *Philosophical psychology*, 17(3):341–373, 2004.
27. Harko Verhagen, Pablo Noriega, and Mark d'Inverno. Towards a design framework for controlled hybrid social games. In Harko Verhagen, Pablo Noriega, Tina Balke, and Marina de Vos, editors, *Social Coordination: Principles, Artefacts and Theories (SOCIAL.PATH). AISB 2013 Convention Proceedings*, pages 83–87, 2013.