A manifesto for conscientious design of hybrid online social systems

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

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

Abstract. Online Social Systems such as community forums, social media, e-commerce and gaming are having an increasingly significant impact on our lives. They affect the way we accomplish all sorts of collective activities, the way we relate to others, and the way we construct are own self-image. These systems often have both human and artificial agency creating what we call online hybrid social systems. However, when systems are designed and constructed, the psychological and sociological impact of such systems on individuals and communities is not always worked out in advance. We see this as a significant challenge for which coordination, organisations, institutions and norms are core resources and we would like to make a call to arms researchers in these topics to subscribe a conscientious approach to that challenge.
In this paper we identify a class of design issues that need attention when designing hybrid online social systems and propose to address those problems using *conscientious design* which is underpinned by ethical and social values. We present an austere framework to articulate those notions and illustrate these ideas with an example. We outline five lines of research that we see worth pursuing.

1 Introduction

We are witnessing major social changes caused by the massive adoption of online social systems that involve human users alongside artificial software entities. These hybrid online social systems promise to satisfy and augment our social needs and the rise of such systems and their use are nothing short of spectacular. Because of the speed of their uptake their has been limited research that looks at the relationship between system design and potential long-term psychological, sociological, cultural or political effects.

Examples of the undesirable consequences of such systems (with varying degrees of autonomous agency participation) include:

– the increasing importance of social media expressions and reactions in building and maintaining identity,

– the possibility of determining personal data from facial recognition applications such as *FindFace*,
– the possibility of determining personal information via automatic scrubbing of online dating services such as *OKCupid*,
– the everchanging algorithm for presenting messages on *Facebook*, outside of the control of the user

The social impact of these applications is magnified by the accessibility of mobile devices, ubiquitous computing and powerful software paradigms that enable innovations in AI to be readily integrated. Despite this, design takes place in an *ad-hoc* and opaque way so that the social consequences of online actions are unknown. The effect of online actions in the real social world is often not understood, we often do not know whether actions are private or public, we cannot be sure of the way in which the actions of others is presented to us, and nor do we know how information about our activity is being used.

As the AI community plays a key role as inventors and builders of the scientific and technological enablers of this phenomenon, we have a moral responsibility to address these issues that requires a sustained, long term commitment from our community. We believe that what is needed is a collective interdisciplinary endeavour across design, sociology, formal methods, interface design, psychology, cultural theory, ethics, and politics to develop a clearer understanding of how we approch and design online social systems. Together we can play an active role in the design of systems where users' understanding of actions, relationships and data is fair and clear. The challenge is great, but then so is the responsibility. Those of us working in the theory, design and implementation of agent-based systems now have a fantastic opportunity to apply our methods and tools in ways which could have impact far beyond that we might have imagined even a few years ago.

This paper then is *a call to arms* for such an initiative, specifically to the COIN community, in the spirit of the "Research Priorities for Robust and Beneficial Artificial Intelligence: an Open Letter". We articulate our proposal around the notion of *conscientious design* as a threefold commitment to a design that is *responsible*, *thorough* and *mindful*.[5].

Conscientious design starts by developing an awareness of the concerns manifest in the current landscape, and understanding how multi-agent techniques can be applied as an effective means to operationalise systems to ameliorate such concerns, and bring it to bear upon our everyday scientific and technological activity. For this we need to (further) develop theories and models of norms, roles, relationships, languages, architectures, governance and institutions for such systems, and do so in a way that naturally lends itself to interdisciplinary research. We need to be *empiricists* (in applying our techniques to modelling current systems), *theorists* (in building implementable models of hybrid social systems), and *designers* (in designing systems); open to working in a strong *interdisciplinary* way across arts, humanities and social sciences. We may also need to break away from our natural comfort zones describing idealised scenarios for

[5] `http://futureoflife.org/static/data/documents/research_priorities.pdf`

agents but we can do so when we recognise just how potentially significant the impact of our research can be.

In this paper we postulate the need to address this challenge, propose a focus of attention —Hybrid Online Social Systems (HOSS)— and give a rough outline of what we see as the main research questions. The paper is structured as follows: In Sec. 2 we point to some background references so as to motivate our election of problematic aspects of HOSS and our proposal of conscientious design, addressed in Sec. 3. In Sec. 4 we propose the core ideas —based on the WIT framework [15]— to make conscientious design operational and in Sec. 5 we illustrate these ideas with an example. All these elements are then put together as a research programme towards conscientious design and implementation of HOSS.

2 Background

2.1 The problem

The range of behaviours that we can carry out online make available all kinds of activity that was not possible even a few years ago. It can affect how we see ourselves, how we choose to communicate, how we value notions of privacy and intimacy, and how we see our value in the world. We are building new metaphors of ourselves while we are in contact with everyone and everybody [9]. The issue that is overlooked by many users is that almost anything that can happen in the real social world —i.e. the one which existed before online systems— can potentially happen in any online one, and worse. We are facing a "Collingridge dilemma": We do not yet know how to take advantage of the opportunities of this technology and avoid its unwanted consequences but we are justifiably concerned that by the time we understand its side-effects it may be too late to control them [6].

2.2 An approach to the solution

We concern ourselves with those systems where there is artificial agency; either because there are software socio-cognitive agents that have some autonomy or because the system infrastructure incorporates agency (such as by actively producing outcomes that are not the ones users expect, or because third parties may interact with that system without the system or its users being aware or intending it to happen). For these "hybrid online social systems", or HOSS, we identify the generic type of features we find problematic and propose a "conscientious" design approach in response.

Our proposal is in tune with the *Onlife Manifesto* [9] and thus aims to respond to the sensitivities and challenges captured in that document. For instance, a *new understanding* of values, new uses of norms and the new guises that their enforcement should take; attention to how values like trust, fairness, solidarity are understood; give users control over the way their own values may become incorporated in the tools they create or adopt. Our proposal can be framed as a part of the "value alignment problem".[6]

[6] Stuart Russell:"... *The right response [to AI's threat] seems to be to change the goals of the field itself; instead of pure intelligence,* we need to build intelligence that is provably aligned with human values...". `https://www.fhi.ox.ac.uk/edge-article/`

Our proposal is akin to the Value-sensitive design (VSD) research framework [10] and similar approaches like *Values in Design* [13] and *disclosive computer ethics* [3]. The main concern in VSD is how values are immersed (mostly unconsciously) in technological artifacts, and postulate that what is usually missing during the design and development phases is a critical reflection upon this unconscious inscription of values. We advocate a *conscientious* approach to put in practice that critical reflection.

VSD offers three "investigation" schemata for inscribing values into the design of systems (i) *conceptual-philosophical* whose aim is to identify relevant values, and relevant direct and indirect stakeholders (not only users), (ii) *empirical* the use of qualitative and quantitative research methods from the humanities and social sciences, to study how people understand and apply values, and (iii) *technical* to determine the role that values play in technologies and how to implement those values identified in the two previous schemata into the systems that are being designed.

We propose a narrower but complementary strategy. We propose to focus attention in those values that are associated with three broad areas of concern that we believe are encompassed by conscientiousness: *thoroughness* (the sound implementation of what the system is intended to do), *mindfulness* (those aspects that affect the individual users, and stakeholders) and *responsibility* (the values that affect others). We postulate an approach to software engineering that is directed towards a particular class of systems (HOSS). It is an approach close to VSD because it rests on a particular categorisation of values but we go further because we understand that those values are instrumented by means of institutional (normative) prescriptions that have an empirical and conceptual grounding, and then implemented through technological artifacts that have a formal grounding. Consequently, while from a teleological point of view we see our approach closer to the ideas of value-sensitive-design, from a technological and methodological point of view, the domain and the proposal are clearly within the COIN agenda.

2.3 The Role of COIN

We believe there is a critical need for a science and discipline of conscientious design for online hybrid social systems which contain human and computational entities. Some of the questions that present themselves to our community are given below.

- How can the agent/AI community collectively recognise this opportunity and spring into action to take part in the development of a science of hybrid online social systems (HOSS) that can lead to their principled design?
- How can we build models, tools, methods and abstractions that come from our own specialities across agent design, interaction protocols, organisations, norms, institutions and governance to underpin the principled design of software incorporating human and artificial agents?
- How can we encourage and support a greater degree of responsibility in the design of online environments in exactly the same way as an urban planner would feel when designing a new locale?

This is not an easy task as the domain is such a diverse and complex one. This is necessarily an early foray into setting up the challenges of charting this space and defining some of the challenges we face in order to do so and doing so in way in which we

can build bridges to other communities. Naturally, we want any undertaking to be wide ranging, to be inclusive so that people from all fields of the agent and AI communities can take part, and where groups from other parties can join with a clear sense of what we mean by a science of online social systems. Studies from other disciplines often lead to important critiques of technological development, what *our community can uniquely provide is a scientific framework* for system design that can both critique current systems but also enable a collective design of future conscientious systems. We will all lose out if there cannot be a collective and interdisciplinary approach to understanding how to design such systems. We need a common technological and scientific framework and language to argue for how we should design the next generation of such systems.

3 Choice of problems and approach: conscientious design of HOSS

The first challenge we propose to address is to develop a precise characterisation of HOSS. As suggested in [5], this can be approached in two directions. First a bottom-up task that consists of studying existing HOSS to identify their essential features and typologies. For each typology we suspect there will be particular ways in which desired properties may be achieved. The task would be to elucidate how values like transparency, accountability, neutrality, and properties like hidden agency and such are achieved in the actual systems and look for those design and implementation resources that tell the degree to which those properties exist. Secondly, top-down research would aim to approximate agent-based abstract definitions of ideal classes of HOSS and gradually make them precise in order to *analytically* characterise the features and properties of the HOSS we design and build.

Far the moment we will speak of HOSS in not-formal terms from the top-down perspective. Loosely speaking, HOSS are IT enabled systems that support collective activities which involve individuals —human or artificial— that reason about social aspects and which can act within a stable shared social space.[7]

This is a tentative "analytic" definition of HOSS (from [15]):

Notion 1 *A* Hybrid online social ssytem (HOSS) *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,*

[7] Such systems have been labelled "socio-technical" [20], *socio-cognitive technical systems* [4], *intelligent socio-technical systems* [12] and we called them *socio-cognitive technical systems* in [15].

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.*[8] *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.*

3.1 Our focus of attention: Hidden agency

The main problems with HOSS are what for a lack of a better term we'll call "unawareness problems" such as *hidden agency*, *insufficient stakeholder empowerment*, and *lack of social empathy*.

Perhaps more than anything, we need to draw out the extent to which these systems have or may acquire *hidden agency*. We mean, those side-effects or functionalities of the system that are exploitable by its owner or others without the user being fully aware of them, even if they were unintended by the designer of the system. In the language of multi-agent systems from 25 years ago, there is an assumption that the agency of online systems is benevolent [11] but if the hidden agency was revealed to users it would often be entirely unwelcome and unwanted. And in the same language, we may see hidden agency as hidden limits to the autonomy of the user.

An example of hidden agency is the recent case of mining on *OKCupid* where a group of researchers not only mined the data of the online dating service but even put

[8] For example if, in a TV game show, participants may form coalitions to express a collective vote in favour of one option, there are two ways of addressing that functionality. First, define a *collective vote* and provide a proper interface for casting it without regulating explicitly how the agreement comes about. Alternatively, *regulate the process of coalition formation and collective voting* and implement it, thus making the system mediate all interactions that lead to a collective vote.

the data collection of 70,000 users online on the Open Science Framework for anyone to use. Although real names were not included, the data of personal and intimate character could easily be linked to find the real identity behind the user names. Even more so, if it would be connected via the profile pictures (which the researchers left out of the database due to space reasons, not ethical considerations) to other social media when using software such as Facefind (http://www.findbyface.com/) and Findface (http://www.findface.ru) Although *OKCupid* managed to have the data removed on copyright violations, in what way the users had an opinion on or say in this is very unclear (a case of insufficient stakeholder empowerment).

A case of lack of social empathy is how the use of *Facebook* for memorial pages may have distressing effects [17]. Large turn-ups at funerals offer comfort and support to those who have lost a loved one. The same effect also applies to online shows of mourning such as the deluge of messages posted when a famous person dies. They show up in the trending topics bar on *Facebook*, spreading the news fast. Even for less famous persons, *Facebook* is playing a role in the mourning process. *Facebook* pages are kept alive, messages are sent to the deceased and memorial pages are put online. But not all is good. Just as a low turn-up at a funeral will cast doubt on the legitimacy of ones sorrow so is the failure of attention in *Facebook* creating doubts. Moreover, the turn-up at a funeral is a private observation limited in time and space whereas *Facebook* measures and shows it all. The number of visitors can be compared to the number of likes or other *emojis* and the number of comments, for all to see.

3.2 What we mean by conscientious design

We will go beyond value-sensitive design towards conscientious design and development. As we mentioned in Sec. 2, we propose to look into a particular set of values —involving technical, individual and social domains— that are linked to the description, specification, implementation and evolution of HOSS. Thus conscientious design and developent of HOSS responds to three properties:

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.

It is here the agent metaphor for system design provides a clear opportunity for providing models that can be understood by academics, users and designers of HOSS. For the commercial-driven applications we might think of designing conscientiousness sensors, small apps that show warning flags when the online application in use collides with the values of the user. But in the remainder of the paper we will look at applications developed in a conscientious way and illustrate the points we wish to make by revisiting applications developed by or close to us.

4 An abstract understanding of HOSS

In order to design HOSS using a conscientious approach we need to come up with a clear characterisation of these systems. Eventually, we should be able to articulate a set of features that discriminate the online social systems that we are interested in — the ones with "unawareness problems" we mentioned — from other online social systems. In our research programme we propose to take a twofold approach for this task: an empirical, bottom-up line that starts from paradigmatic examples and a top-down line that provides an abstract characterisation. We already took a first step along this second line with the WIT framework proposal that we summarise here.[9]

We start from the observation that HOSS are systems where one needs to *govern* the interaction of agents that are situated in a physical or artificial world by means of technological artifacts. The key notion is "governance" because in order to avoid hidden agency and other unawareness problems we need to control on one hand, the frontier between the system itself and the rest of the world and, on the other, the activity of complex individuals that are at the root of HOSS. In order to elucidate how such governance is achieved we proposed the following tripartite view of HOSS (Fig. 1):

View 1: An *institutional* system, $\mathcal{I}$, that prescribes the system behaviour.
View 2: The *technological artifacts*, $\mathcal{T}$, that implement a system that enables users to accomplish collective actions in the real world ($\mathcal{W}$), according to the rules set out in $\mathcal{I}$.
View 3: The system as it exists in the *world*, $\mathcal{W}$, as the agents (both human and software) see it and with the events and facts that are relevant to it.

In other words, $\mathcal{W}$ may be understood as the "organisation" that is supported by an "online system" $\mathcal{T}$ that implements the "institutional conventions" $\mathcal{I}$.

Notice that we are referring to one single system but it is useful to regard it from these three perspectives because each has its own concerns. Notice also, these three perspectives need to be *cohesive* or "coherent" in a very particular way: at any given time t, there is a *state of the system* s_t that is exactly the same for all agents that are in the system, and when an agent interacts with the system (in $\mathcal{W}$), that state of the system changes into a new state s'_t, which is again common to all agents, if and when the agent's action is processed by the system (in $\mathcal{T}$) according to the specifications of the system (in $\mathcal{I}$).

[9] See [15] for a more leisurely discussion of the WIT proposal.

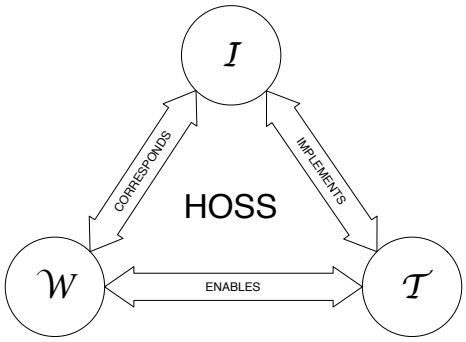

Fig. 1. The WIT trinity: The ideal system, $\mathcal{I}$; the technological artifacts that implement it, $\mathcal{T}$, and the actual world where the system is used, $\mathcal{W}$.

In order to make this cohesion operational, we define three binary relations between the views. As sketched in Fig. 1, the institutional world *corresponds* with the real world by some sort of a "counts-as" relationship [19] —and a mapping between entities in $\mathcal{W}$ and entities in $\mathcal{I}$— by which *relevant* (brute) facts and (brute) actions in $\mathcal{W}$ correspond to institutional facts and actions in $\mathcal{I}$ (and brute facts or actions have effects only when they satisfy the institutional conventions and the other way around). Secondly, $\mathcal{I}$ specifies the behaviour of the system and is *implemented* in $\mathcal{T}$. Finally, $\mathcal{T}$ *enables* the system in $\mathcal{W}$ by controlling all inputs that produce changes of the state and all outputs that reveal those changes.

It should be obvious that HOSS are not static objects. Usually, each HOSS has a lifecycle where the process of evolution is not all that simple [5].

4.1 A WIT understanding of conscientious design

Conscientious design adds meaning to the WIT description by throwing light upon certain requirements that the three binary relations should satisfy. Thus, in the first phase of the cycle, the main concern is to make the design value-aware from the very beginning, in line with the recommendations of value-sensitive-design. That is, analyse systematically the *thoroughness*, *mindfulness* and *responsibility* qualifications of the system, so those ethical, social and utilitarian values that are significant for the stakeholders are made explicit. This examination would then pursue a proper operationalisation of the intended values so that they may be properly translated into institutional conventions. Note that it is in this phase where mindfulness and responsibility analysis of requirements are more present, while thoroughness is the focus of the next stage.

As suggested in [15], the operationalisation of those values together with the usual software engineering elements (functionalities, protocols, data requirements, etc.) should be properly modelled (in $\mathcal{I}$) and then turned into a specification that is implemented in $\mathcal{T}$. The passage from the elicitation of requirements to the modelling of the system is facilitated by the availability of *metamodels* [1] that provide the *affordances* to represent correctly those requirements. Ideally, such representation should satisfy three criteria:

they should be *expressive*, they should be formally *sound* and it should become *executable*. The metamodel should also provide *affordances* to model the evolution of the system. Note that when relying on a "metamodel", its expressiveness will bias the way conscientiousness is reflected in the eventual specification.

The running system requires components for validation of the functionalities of the system, for monitoring performance and the devices to control transfer of information into and out of the system. These validation and monitoring devices should be tuned to the conscientious design decisions and therefore reveal how appropriate is the implementation of the system with respect to conscientious values and where risks or potential failures may appear.

5 How to achieve conscientious compliance

The abstract WIT and conscientious design ideas take rather concrete forms when building new HOSS.

5.1 An example of conscientious design, the *uHelp app*

Picture a community of *monoparental* families that decide to provide mutual support in everyday activities: baby-sitting, picking up children from school, go shopping, substitute at work during an emergency, lending each other things like strollers or blenders. One may conceive an *app* that facilitates such coordination. But —sensitive to conscientious design— one wants to make sure that coordination is in accordance with the values of the community. In this case, for example, *solidarity*: everyone helps each other for free; *reciprocity*: no free riding; *involvement*: old people may want to help; *safety*: no one without proper credentials should be able to pick up a child; *privacy* (no revelation of personal data, of behaviour of members of the network); *trust*: you demand more trustworthiness in some tasks than others and trust is a binary relation that changes with experience.

You program the *app* so that it reflects those values faithfully and effectively. Moreover, you want the community to be aware of the degree of compliance/usefulness of the network, and that the community may change the specification to improve it or adapt to new preferences or values. Also you want the *app* to be unobtrusive, reliable, practical (light-weight, easy to download, easy to support, easy to update), and not contain hidden agency.

Abstracting away from the actual specification, the main conscientious-compliance features that the *app* should have are:

1. *From a practical perspective:* (i) Useful for the relevant coordination tasks, (ii) Faithful and responsive to the community's goals, preferences and values, (iii) Have the community in control of evolution (iv) No hidden agency.
2. *From an institutional perspective:* (i) shared ontology, (ii) common interaction model and interaction conventions (the *smartphone app*), (iii) govern a core coordination process: values, norms, governance (iv) controlled evolution: participatory, reliable, effective, (v) no unwanted behaviour.

3. *From a technical perspective:* (i) proper monitoring (key performing indicators, historical logs), (ii) automated updating (iii) robust and resilient *app*. (iv) Safe against intrusions and "zero information transfer" (only the intended information is admitted into the system and only intended information is revealed).

This type of application and the conscientious-design perspective have been under development in the IIIA for some time [16], and there is a working prototype, *uHelp*, that implements these ideas in a *smartphone app* and has already undergone field tests with actual users [14].

Where in WIT is conscientiousness

This example also serves to illustrate how conscientious design considerations may be reflected in the WIT cycle:

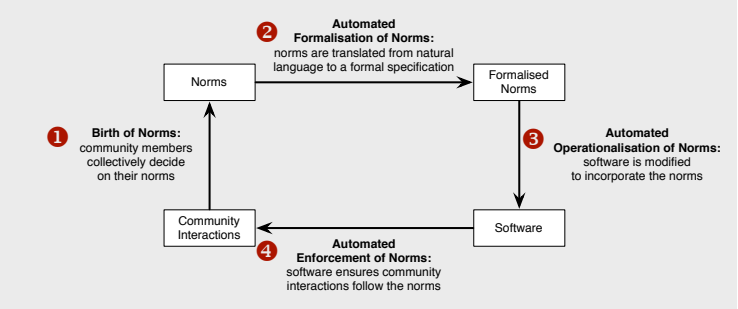

Fig. 2. Life-cycle of norms in the *uHelp app* from [16]

For specification: The *UHelp app* exists as a smartphone-based social network in $\mathcal{W}$. It involves two realms: The first one consists of the physical components of the system, which includes smartphones, addresses, schools, ID cards, blenders and strollers, as well as the organisation of parents that own the application and the group of technicians that support is everyday use and maintenance. The other is the activities that are coordinated with the *app* (picking children up, help with shopping) and the activities that are needed to use the *app* (running a server, uploading the *app* in *iTunes*). Thus in order to describe (in $\mathcal{I}$) how it should work, WIT would need an *expressive description language* that should include coordination conventions, values, norms, and so on. In other words, a description language that can handle *mindful* and *responsible* values. On the other hand, the specification should be such that users are comfortable with the conventions that govern the system and its evolution; and in this respect, the system needs to be *thorough*.

For formalisation: Description needs to be made precise: How are values associated with norms? Does the system support norm changes with some formal mechanism? Is simulation the appropriate tool for validation and monitoring? In our case, *UHelp* is intended to have a development *workbench* that uses electronic institutions coordination and governance affordances (an EI-like metamodel [8]) that is being extended

to handle values. Furthermore, the *UHelp* workbench shall contain also an argumentation environment for arguing about normative changes (to empower stakeholders) and a simulation module to test and anticipate (responsibly) potential changes of the system.

For implementation: One would like to rely on technological artifacts that make a *thorough* implementation of the specification of the system. Those artifacts may include devices like model checking, agent-mediated argumentation, agent-based modelling and simulation. In particular, the *uHelp* workbench shall be coupled with a platform that deals with the implementation of the functionalities of the value-based social network and also with the implementation and maintenance of the *app* itself.

What does it mean to be *conscientious* in the *uHelp app*?

This is a sketch of an answer for a *uHelp*-like HOSS.

Thorough: For specification purposes, a metamodel that *affords* proper representation, sound formalisation, correct implementation of: (i) Coordination and governance (activities, communication, social structure, data models, procedural norms, enforcement, etc.) (ii) Values, (ontology, norms, inference) (iii) Monitoring (KPI, use logs) (iii) Evolution (automated or participatory updating, validation).

Mindful: Proper elicitation and operationalisation of *values*, preferences and goals, sensible selection of functionalities; lucid assessment of performance; explicit *stakeholders entitlements and responsibilities*; sensible attention to *usability and culturally sensitive* issues; due attention to *privacy.* What *agency* is afforded by the system?

Responsible: (i) Clear and explicit commitments about *information transfer* in the system, uses of performance data, and about *management* of the system. (ii) Clear requirements and commitments of system *updating*: what may users do; what type of guarantees and requirements are part of the evolution process. (iii) Proper description of coordination behaviour (requirements and outcomes for intended behaviour of automated activities and support functionalities). (iv) Explicit description about *ownership* of the system, about relationship with *third-party software* and about *commercial* and other commitments with *third parties*.

5.2 Three roads to application:

Rather than Quixotic fighting of *Facebook* windmills and trying to make existing HOSS conscientious-compliant we identify three lines of attack: (i) Conscientiousness by design, like the *uHelp* example; (ii) methods and devices to test the extent to which an existing HOSS is conscientious-compliant. This includes means to determine analytically whether a given HOSS has problems like hidden agency, insufficient user empowerment, inadequate social empathy; and (iii) *plug-ins* that may provide some conscientious-compliant features to existing HOSS.

6 Towards a new Research Programme

In order to support conscientious design, we propose a research programme (based on [15]) around the following five topics (see Fig. 3):

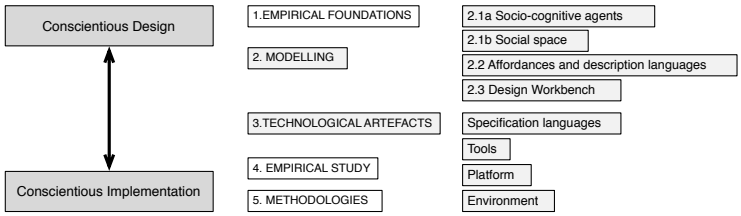

Fig. 3. The main challenges in the development of a framework for conscientious design of hybrid online social systems.

1. Empirical foundations: Conscientious design intends to build systems that support expected values and avoid unwanted features and outcomes. As we have been arguing in previous sections, we find that a systematic examination of actual socio-technical systems and of the values and unwanted outcomes involved need to be at the root of formal, technological and methodological developments in conscientious design. The outcomes should be, on one hand, a proper characterisation of HOSS and, on the other, a proper operationalisation of problematic manifestations in HOSS and the preventive and remedial features based on design conscientiousness.

2. Modelling: Conscientious design means: (i) that the creation of each HOSS be founded on a precise description of what the system is intended to be; (ii) that such description be faithfully implemented; and (iii) that the implementation actually works the way it is intended to work. In fact, it would be ideal if one could state with confidence the actual properties —scalability, accuracy, no unwanted side-effects, etc.— that the working HOSS has, because either we design the system with those properties in mind or because we are able to predicate them of an existing HOSS or an existing HOSS supplemented with *ad-hoc* plug-ins.

We propose to split the problem of conscientious modelling in three main parts: (2.1) Separate the design of a HOSS in two distinct concerns (the design of socio-cognitive agents and the design of a social space); (2.2) develop high-level description languages; and (2.3) develop a "design workbench" that provides concrete modelling components that translated the description of a HOSS into a specification.

2.1.(a) Socio-cognitive agents. First it is important to provide a conceptual analysis of the types of agents that may participate in a HOSS. The significant challenge is to create agent models that exhibit true socio-cognitive capabilities Next to it is the challenge of developing the technological means to implement them; hence the definition of agent architectures using a formal and precise set of agent specification languages with the corresponding deployment and testing tools.

2.1.(b) The social space. In addition one has to provide a sufficiently rich understanding of the social spaces which are constituted in HOSS. What are the relationships, what are the norms, how can it evolve, and a clarity about how this space is related to the external world. Any model would also need to consider how several HOSS may co-exist in a shared social space. Features that need to be included are openness, regulation, governance, local contexts of interaction, organisational and institutional structures.

2.2. Affordances and description languages. We need to identify the *affordances* that are needed, both, to achieve conscientious design in general, and also to support a *thorough* implementation of particular HOSS (as illustrated in Sec. 5). In other words, what are the concepts, analogies and expressions that a social scientist, an urban planner, a game designer or a sociologist may find more suitable to model agents and social space of a HOSS. In practice, a description language for modelling agents should afford the means for the agent to be aware of the state of the system, of its own state, and to hold expectations of what actions it and other participants can take at a given state. For modelling the social space, the language should be able to express those elements that *afford* participants the means to have a shared ontology, a common interaction model and communication standards coupled with some form of governance.

2.3. Design workbench. It would include the concrete versions of the affordances. That is, the "vocabulary" that the description languages will use in order to model an actual system. So, for instance, if the system will involve norms, then the workbench would have norms expressed with a particular structure together with concomitant paranormative components like normative inference, nor-enforcement mechanisms, etc. In the *uHelp* example, we need functional norms that have the shape of "permissions" and they are represented as production rules.

3. Technological artifacts: The challenge is to build technological artifacts that facilitate and ensure the conscientious deployment of HOSS. One way of addressing this is to have an artifact for each modular component of the design workbench the components that are needed to assemble those modules. Again, for *uHelp* there is a specification language *SIMPLE* [7], that is interpreted by the *uHelp app*. An ambitious approach towards thorough implementations is to have *full platforms* that allow a translation form a specification to technological platform that implements that specification. The [2] volume discusses this line, and several frameworks for meta-modelling and implementation are available [1]. Another way to achieve this formal soundness is to start with an existing platform —*BrainKeeper*, *Amazon Turk*, *Ushahidi*— provide its formal counterpart and use it to analyse applications of the platform.

4. Empirical study of HOSS: Complementing Topic 1, we find two further reasons to study working HOSS. One is to document compliance and failure of conscientious principles and recommendations, the other is to use the information that arises from their use as source data for socio-cognitive research.

5. Methodologies for conscientious design and deployment The challenge is to develop a precise conceptual framework to describe conscientious features and methodological guidelines that prescribe how to recognise and achieve the intended properties and behaviour in conscientious HOSS. We need to explore key values like fairness, trustworthiness, social empathy in principled terms (see [12,18]) so that we can speak properly of achieving engineering tasks like requirement elicitation or tooling conscientiously.

7 Peroration in four claims

First: The era of online social systems that on the surface seem to satisfy augmented social needs is *here to stay*. However, the rise of such systems has been so dramatic that *we simply do not know* what the effects will be either psychologically, sociologically, culturally or politically.

Second: Some online social systems that involve human and artificial agency (HOSS) exhibit behaviours like hidden agency, inadequate stakeholder empowerment and lack of social empathy that may be problematic and deserve to be prevented or contended with in a sound manner.

Third: The challenge we face is to develop precise notions and the associated methodological guidelines and tools to design HOSS systems in a *conscientious* way that is *thorough*, *mindful* and *responsible*.

Fourth: This paper is a *call to arms* for such an initiative. Those of us working in the theory, design and implementation of agent-based systems, work in a field where there is an unharvested opportunity to apply our methods and tools in ways which could have impact far beyond that we might have imagined. It may mean a changing of the focus of our community and having to break away from our comfort zones describing idealised scenarios for agents, and in doing so we would need to be extremely humble about what we might achieve. But we should try, as the potential for sustained lasting impact for social and cultural good is potentially large.

The responsibility is substantial but the opportunity is ours.

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. This research was partially supported by project MILESS (MINECO TIN2013-45039-P).

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. Giulia Andrighetto, Guido Governatori, Pablo Noriega, and Leendert W. N. van der Torre, editors. *Normative Multi-Agent Systems*, volume 4 of *Dagstuhl Follow-Ups*. Schloss Dagstuhl - Leibniz-Zentrum fuer Informatik, 2013.
3. P. Brey. Values in technology and disclosive computer ethics. In L. Floridi, editor, *The Cambridge Handbook of Information and Computer Ethics*, pages 41 – 58. Cambridge University Press, Cambridge, 2010.

4. Cristiano Castelfranchi. 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. http://www.slideshare.net/sleeplessgreenideas/castelfranchi-aamas13-v2?ref=httpMay 2013.
5. Rob Christiaanse, Aditya Ghose, Pablo Noriega, and Munindar P. Singh. Characterizing artificial socio-cognitive technical systems. In Andreas Herzig and Emiliano Lorini, editors, *Proceedings of the European Conference on Social Intelligence (ECSI-2014), Barcelona, Spain, November 3-5, 2014.*, volume 1283 of *CEUR Workshop Proceedings*, pages 336–346. CEUR-WS.org, 2014.
6. David Collingridge. *The Social Control of Technology*. St. Martin's Press, London, 1980.
7. Dave de Jonge and Carles Sierra. Simple: a language for the specification of protocols, similar to natural language. In Murat Sensoy Pablo Noriega, editor, *The XIX International Workshop on Coordination, Organizations, Institutions and Norms in Multiagent Systems*, Istanbul,Turkey, May 2015.
8. Mark d'Inverno, Michael Luck, Pablo Noriega, Juan A. Rodriguez-Aguilar, and Carles Sierra. Communicating open systems. *Artificial Intelligence*, 186(0):38 – 94, 2012.
9. L. Floridi, editor. *The Onlife Manifesto: Being Human in a Hyperconnected Era*. Springer International Publishing, Cham, 2015.
10. B. Friedman, editor. *Human Values and the Design of Computer Technology*. Cambridge University Press, Cambridge, 1997.
11. J. R. Galliers. The positive role of conflicts in cooperative multi-agent systems. In Y. Demazeau and J.-P. Mueller, editors, *Decentralized AI: Proceedings of the First European Workshop on Modelling Autonomous Agents in a Multi-Agent World*. Elsevier, 1990.
12. Andrew J. I. Jones, Alexander Artikis, and Jeremy Pitt. The design of intelligent socio-technical systems. *Artif. Intell. Rev.*, 39(1):5–20, 2013.
13. C. P. Knobel and G. C. Bowker. Values in design. *Commun. ACM*, 54(7):26–28, 2011.
14. Andrew Koster, Jordi Madrenas, Nardine Osman, Marco Schorlemmer, Jordi Sabater-Mir, Carles Sierra, Dave de Jonge, Angela Fabregues, Josep Puyol-Gruart, and Pere García. u-help: supporting helpful communities with information technology. In *Proceedings of the First International Conference on Agreement Technologies (AT 2012)*, volume 918, pages 378–392, Dubrovnik, Croatia, 15/10/2012 2012.
15. Pablo Noriega, Julian Padget, Harko Verhagen, and Mark d'Inverno. The challenge of artificial socio-cognitive systems. In A. Ghose, N. Oren, P. Telang, and J. Thangarajah, editors, *Coordination, Organizations, Institutions, and Norms in Agent Systems X*, Lecture Notes in Computer Science 9372, pages 164–181. Springer, 2015.
16. N. Osman and C. Sierra. A roadmap for self-evolving communities. In A. Herzig and E. Lorini, editors, *Proceedings of the European Conference on Social Intelligence (ECSI-2014), Barcelona, Spain, November 3-5, 2014*, volume 1283 of *CEUR Workshop Proceedings*, pages 305–316. CEUR-WS.org, 2014.
17. Whitney Phillips. Loling at tragedy: Facebook trolls, memorial pages and resistance to grief online. *First Monday*, 16(12), 2011.
18. J. Pitt, D. Busquets, and S. Macbeth. Distributive justice for self-organised common-pool resource management. *ACM Trans. Auton. Adapt. Syst.*, 9(3):14, 2014.
19. John R. Searle. What is an institution? *Journal of Institutional Economics*, 1(01):1–22, 2005.
20. Eric Trist. The evolution of socio-technical systems. *Occasional paper, Ontario Ministry of Labour*, 2, 1981.