On Human-Agent Collectives

N. R. Jennings, L. Moreau, D. Nicholson, S. Ramchurn, S. Roberts, T. Rodden, A. Rogers

Introduction

The computer has come a long way from its initial role as a scientific tool in the research lab. We live in a world where a host of computer systems, distributed throughout our physical and information environments, are increasingly implicated in our everyday actions. Computer technologies impact all aspects of our lives and our relationship with the digital has fundamentally altered as computers have moved out of the workplace and away from the desktop. Networked computers, tablets, phones and personal devices are now commonplace, as are an increasingly diverse set of digital devices built into the world around us. Data and information is generated at unprecedented speeds and volumes from an increasingly diverse range of sources and via ever more sensor types. It is then combined in unforeseen ways, limited only by human imagination. People's activities and collaborations are becoming ever more dependent upon and intertwined with this ubiquitous information substrate.

As these trends continue apace, it is becoming apparent that many endeavours involve the symbiotic interleaving of humans and computers. Moreover, the emergence of these close-knit partnerships is inducing profound change. The ability of computer systems to sense and respond to our on-going activities in the real-world is transforming our daily lives and shaping the emergence of a new digital society for the 21st century. More specifically, rather than issuing instructions to passive machines that wait until they are asked before doing anything, we are now starting to work in tandem with highly inter-connected computational components that act autonomously and intelligently (aka agents [Wooldridge and Jennings, 1995]). This shift is needed to cope with the volume, variety, and pace of the information and services that are available. It is simply infeasible to expect individuals to be aware of the full range of potentially relevant possibilities and be able to pull them together manually. Computers need to do more to proactively guide users' interactions based on their preferences and constraints. In so doing, greater attention needs to be given to the balance of control between people and machines. In many situations, humans are in charge and agents predominantly act in a supporting role, providing advice and suggesting options. In other cases, however, agents are in control and humans play the supporting role (e.g., automatic parking systems on cars and algorithmic trading on stock markets). Moreover, these relationships may change during the course of an activity (e.g. a human may be interrupted by a more pressing request and so take a less hands-on approach to the current task or an agent may encounter an unexpected situation and have to ask for human assistance for a task it was planning to complete autonomously).

We term this emerging class of systems **human-agent collectives** (HACs) to reflect the close partnership and the flexible social interactions between the humans and the computers. As well as exhibiting increased autonomy, such systems are inherently open and social. This openness means participants need to continually and flexibly establish and manage a range of social relationships. Thus, depending on the task at hand, different constellations of people, resources, and information need to come together, operate in a coordinated fashion, and then disband. The openness and presence of many distinct stakeholders, each with their own resources and objectives, means participation is motivated by a broad range

of incentives–*extrinsic* (e.g., money or tax-benefit), *social* or *image motivation* (e.g., public accreditation or leader-board position) or *intrinsic* (e.g., personal interest in a social cause, altruism, or hobby) [Ariely *et al.*, 2007]–rather than diktat. Moreover, once presented with such incentives, the stakeholders need to be evaluated and rewarded in ways that ensure they sustain behaviours beneficial to the system they form part of [Scekic et al., 2013].

Embryonic examples of future HAC systems where people routinely and synergistically interact and collaborate with autonomous software are starting to emerge. For example, as we travel, increasingly interconnected transport management systems cooperate to aid our journey. Systems such as Waze (http://www.waze.com/) blend citizen and (electronic-) sensor generated content to aid the user. Furthermore, software agents can proactively interact to arrange a place to stay and somewhere to eat in accordance with the traveller's preferences and current circumstances. However, despite relevant work on parts of the problem in the AI, HCI, CSCW and UbiComp communities, it is apparent that developing a comprehensive and principled science for HACs is a major research challenge, as is the process by which such systems can be designed and built, and the means by which HACs will be accepted and deployed in the wild.

What's Different about Human-Agent Collectives?

HAC systems exhibit a number of distinctive features that make it particularly challenging to engineer and predict their behaviour. Their open nature means control and information is widely dispersed among a large number of, potentially self-interested, people and agents with different aims and objectives. The various system elements exhibit a range of availabilities, some are persistent others are transient. The independent actors need to coordinate flexibly with people and agents that are themselves adapting their behaviours and actions to the prevailing circumstances to best achieve their goals. The real-world context means uncertainty, ambiguity, and bias are endemic and so the agents need to handle information of varying quality, trustworthiness and provenance. Thus, techniques are required to provide an auditable information trail from the point of capture (a sensor or a human participant), through the fusion and decision processes, to the point of action, and the agents will have to reason about the trust and reputation of their collaborators to take the best course of action. Finally, in many cases, it is important that the collective action of the volitionally participating actors results in acceptable social outcomes (such as fairness, efficiency or stability). When taken together, these features of HACs require us to:

- understand how to provide *flexible autonomy* that allows agents to sometimes take actions in a completely autonomous way without reference to humans, while at other times being guided by much closer human involvement.
- discover the means by which groups of agents and humans can exhibit *agile teaming* and come together on an ad hoc basis to achieve joint goals and then disband once the cooperative action has been successful.
- elaborate the principles of *incentive engineering* in which the actors' rewards are designed so that the actions the participants are encouraged to take generate socially desirable outcomes.
- design an *accountable information infrastructure* that allows the veracity and accuracy of seamlessly-blended human and agent decisions, sensor data, and crowd generated content to be confirmed and audited.

A number of research domains are beginning to explore fragments of this overarching vision. However, none of them is dealing with the totality, nor the associated system-level challenges. For example, interacting intelligent agents are becoming a common means of designing and building systems that have many autonomous stakeholders, each with their own aims and resources [Jennings, 2001]. To date, much of this work has focused on systems where all the agents are either software or hardware (e.g., robots or unmanned autonomous systems (UAS)). However, it is increasingly being recognised that it is both necessary and beneficial to involve humans, working as active information gatherers and information processors, in concert with autonomous software agents, within such systems [Tambe *et al.*, 2005; Kamar *et al.,* 2013]. For example, systems have been demonstrated where humans gather real-world information and pass it to an autonomous agent that performs some basic aggregation before presenting it online [Reddy *et al.,* 2007]. Such approaches are often termed participatory sensing [Paxton & Benford, 2009] or citizen-sensing [Krause *et al.*, 2008]. Likewise, a number of systems have been demonstrated in which autonomous agents pass information processing tasks to the human participants, then collect and aggregate the results [von Ahn *et al.,* 2008]. However, these broad strands of work typically assume the authority relations between humans and agents are fixed and that there exists a largely static set of skilled human participants who participate on a voluntary basis[1]. This contrasts with the HAC view of agents operating within a dynamic environment in which the flexible autonomy varies the human-agent authority relationships in a context-dependent manner and in which these actors individually make decisions based on their preferences and the properties of their owners.

In the areas of HCI and CSCW, research has increasingly turned to the crowd and how to exploit computer systems to harness and coordinate the endeavours of people [von Ahn and Dabbish*, 2004*]. Essentially, the task has been to manage the means through which people are instructed and to coordinate their responses in a manner that makes sense. This large-scale networked collaboration is typically achieved using software systems to coordinate and analyse these human endeavours. Moreover, software agents have emerged as a key technology for observing and reacting to human activities [Maes, 1994]. This approach has also gained popularity with mixed-initiative systems [Horvitz*,* 1999] and the development of context aware computing approaches within UbiComp [Abowd *et al., 2002*]. However, in most of this work, the software agents are a tool to aid in understanding and managing user interactions. Users are in the foreground and the agents are in the background. The challenge of HACs is moving from the presumption of a dominant relationship, to consider how users and agents co-exist on a common footing and are considered in a manner that allows flexible relationships to emerge.

The role of people within HACs also brings the incentives of the participants to the fore. Most current systems largely assume altruistic and benevolent behaviour, and have not dealt with the need to provide incentives to potentially self-interested participants, nor have they explicitly handled the inherent uncertainty of participatory content in a consistent manner (see [Rahwan *et al.*, 2013a] for examples of such behaviour and [Naroditskiy *et al.*,

[1] Amazon Mechanical Turk (AMT), and other similar systems, is an exception to this in that it allows software systems to automatically generate Human Intelligence Tasks and make payments to a large pool of human participants who complete them.

2012] for the design of incentive structures to combat it). Similarly, current approaches to accountable information infrastructures have focused on augmenting specific systems, such as databases [Buneman *et al.*, 2008] or computational workflows [Gil *et al.*, 2007], with the ability to track information provenance. Now, emerging efforts, such as the W3C PROV Recommendation [Moreau *et al., 2013*], are starting to allow for the tracking of provenance across multiple systems and to systems where confidentiality of data needs to be preserved [Kifor *et al.*, 2006]. However, no work deals with the simultaneous challenges of humans in the loop and long-term and online operation.

Human-Agent Collectives in Action

Consider the aftermath of a major natural disaster. A number of organisations are in the area, including first responders (FRs), humanitarian aid organisations and news reporters, as well as locals. A key aim for many of these actors is to assess the situation to determine the areas to focus on in the coming days and weeks. To assist in this task, a number of the organisations have UASs that can be used for aerial exploration and a number of locals have installed sensors to monitor the environment (e.g., two weeks after the Fukushima incident, locals had built and deployed over 500 Geiger counter sensors and were uploading their readings – see http://jncm.ecs.soton.ac.uk/), in addition to many locals using social media platforms such as Ushahidi or Google Crisis Response to record requests for help and complete maps of the stricken area. A representative system architecture for this HAC is shown in Figure 1 and an associated video is available at http://vimeo.com/76205207.

As can be seen, the information infrastructure contains a wide variety of content (e.g. maps of roads and key amenities, weather forecasts and social media reports from locals in the affected areas), from many sources. Some of these sources provide higher quality, more trustworthy information than others (e.g., international aid organisations versus locally built environmental sensor readings). To help account for and justify the decisions that are made, the provenance of information is stored wherever possible. Moreover the decisions made by both responders and autonomous agents (including UASs) are tracked to ensure that all members of the HAC are accountable for their actions and that the successes and failures of the rescue effort can be better understood when such data is reviewed at a later stage.

At the start of a day, the various actors (e.g., FRs or local volunteers) register their availability and relevant resources (e.g., UAS, ground transport vehicles or medical supplies) and indicate specific tasks they would like to perform (e.g., search the area near the school or determine if there is running water in a particular district). These tasks will be informed by their current assessment of the situation and may be influenced by particular requests from locals for assistance.

As a first step, some actor constructs a plan to achieve one or more of the tasks[2] (i.e. the HAC forms). This plan is likely to involve constructing teams of people, agents, and resources to work together on a variety of sub-tasks because many activities are likely to be beyond the capability of just one team member. As they join, the various responders may accept the plan as is and be ready to start enacting it. However they may wish to make minor

[2] It may be a human or a software agent that proposes the initial plan. Moreover, multiple actors may attempt to construct plans simultaneously, some of which may not come to fruition.

modifications (e.g., putting in way points en route to the chosen area to maximise the value of the information obtained or requesting help from volunteers for some parts of the plan). Some may even desire to make more major modifications (e.g., indicating that a particular subtask that has initially been excluded is more important than one of the suggested ones or that significant extra resources are needed for the plan to be successful). This plan co-creation iterates until agreement is reached, with various cycles of the humans taking charge and the agents re-planning to account for the responders' preferences.

Due to the nature of the problem and environment, the HAC's plan execution (operation) phase may not go smoothly. New higher priority tasks may appear, planned ones may turn out to be unnecessary, new actors and resources may become available, or committed ones may disappear (e.g., due to FRs being exhausted or UASs running out of power). All of these will involve the agents and the humans in an on-going monitoring and re-planning endeavour, potentially involving the disbanding of existing teams and the coming together of new ones with different combinations of capabilities that are a better fit for certain types of rescue missions. Moreover, the autonomy relationship between the humans and the agents may change during the course of plan execution. For example, a team of UASs may initially be instructed to gather imagery from a particular area in an entirely autonomous fashion and not to disturb the FRs until the whole task is complete. However, in performing this task the UASs might run into difficulty and request assistance with a complex operation (e.g., manoeuvring in a tight space to get a particular view of a building) or the UASs might discover something important which they believe is worth interrupting the responders for or worth requesting help from an online crowd to analyse the images collected. As unforeseen events happen and the complexity of the effort grows to involve hundreds of responders, volunteers, and UASs, the response coordinators may need to rely more on autonomous agents to compute plans and allocate resources and actions to individuals and UASs. Moreover, when the coordinators detect the agents are less able to judge human abilities (e.g., because fatigue reduces human performance) or the nature of a task (e.g., complex digging operations or surveillance), the coordinators may choose to step in to change the plan or input more information.

Over a number of days, responders may notice that locals are less willing to support the rescue effort and that those who do only work on tasks in their own areas. This means some areas don't receive sufficient attention. To help fill this gap, a number of additional incentive schemes could be put in place. First, volunteers could be rewarded if they sign up friends and family to assist with the rescue effort, such rewards could be financial (e.g., payment for hours worked or vouchers for fuel) or non-financial (e.g., promise of quicker return of their amenities). Second, individuals who complete tasks outside their local area and on high priority tasks could receive additional credit, such as community service awards or double time pay. Finally, to encourage accurate reporting of the importance and urgency of tasks, an incentive structure which increases the reputation of individuals who make assessments that accord well with those of the professional responders may be introduced and bonus payments made to recruiters who hire the top task performers.

Key Research Challenges

HACs present research challenges in, amongst other things, how we balance control between users and agents (flexible autonomy), dynamically (dis)assemble collectives (agile teaming), motivate actors (incentive engineering) and how we provide an information infrastructure to underpin these endeavours. While none of these are entirely new areas, the HAC system context introduces additional complexity and brings new elements to the fore.

Flexible Autonomy

HACs provoke fundamental questions about the relationship between people and digital systems that exhibit some form of autonomy. Specifically, the emergence of HACs highlights the growing extent to which computer systems can no longer be thought of as entirely subservient. We routinely obey navigation systems without question or follow computer-generated instructions delivered on our phones. As such autonomous systems increasingly instruct us, new forms of relationship are emerging. This shift not only raises issues about how we might design for interaction within these systems, it also brings into focus larger social and ethical issues of responsibility and accountability.

HACs are fundamentally socio-technical systems. Relationships between users and autonomous software systems will be driven as much by user focused issues (such as responsibility, trust and social acceptability), as technical ones (such as planning or coordination algorithms). Consequently, we need to uncover the interactive principles of flexible autonomy that shape these systems. A critical issue here is the balance of control between the agents and the users involved in the collective. In particular, when do users need to be in control and when should software systems override them?

Given this, a key challenge is how to ensure a positive sense of control within HACs. Core to this issue is the sense of social accountability and responsibility inherent in our everyday activities. Our professional and personal actions are routinely available to others and we are held accountable for them. In fact, it could be argued that this shapes many of our broader societal relations and understandings. But how will we feel when we are sharing our world with computational elements that exert as much control over the environment as us? Will the relationship be an easy or a tense one and how might we manage that relationship? What checks and balances are needed to allow a fruitful relationship to mature between software agents and humans in HACs as the balance of control shifts between them?

Tackling these questions requires us to think about how software agents might reveal their work to users. At present, such software often operates “behind the scenes” with limited visibility to users. They might make a recommendation on our behalf or schedule activities with the result of their endeavors presented to users. However, little if anything is conveyed of the rationale leading to this result. In contrast, computational agents within HACs will need to make their actions and rationale available so they can be socially accountable.

Revealing the role and actions of software agents to users will bring into focus a raft of questions that require us to consider the broader social and ethical issues, potentially prompting significant reflection on the legal and policy frameworks within which these systems operate. For example, given the collective nature of the endeavor, it is important to

determine who or what will ultimately be responsible for a particular outcome and what this might mean for the application of this approach. More routinely, to what extent will people allow the software agents in these collectives the trust and latitude they might give to other trained and qualified professionals? and will it be acceptable for software agents to make mistakes as they learn how to do a job as part of a collective?

A critical issue here is how to represent both human and agent endeavours in HACs, at multiple levels of scale and aggregation. As well promoting a sense of social accountability, the ability to recognise and understand the activities of others and to flexibly respond to these actions is necessary to enable cooperation and to coordinate actions as part of broader social endeavours. Thus the provision of mechanisms to make users aware of the actions of others is central to the design of many cooperative systems. In particular, the following are central issues: what mechanisms are needed to allow us to do this with users and autonomous software agents alike?, how might we sense human actions and recognise the various activities a user is involved in and how might these be conveyed to software agents?, and what are the most appropriate techniques for presenting agents' actions and on-going progress to users?

Agile Teaming

Humans and agents will form short-lived teams in HACs and coordinate their activities to achieve the various individual and joint goals present in the system before disbanding. This will be a continual process as new goals, opportunities and actors arrive. To date, research within the multi-agent systems community has generated a significant number of algorithms to form and coordinate teams; specifically those algorithms found within the areas of coalition formation and decentralised coordination [Rahwan *et al.*, 2009; Rogers *et al.,* 2011]. However, many of these approaches focus on interactions between software agents alone and do not consider the temporal aspects of agile teaming [Stone *et al.*, 2010].

In HAC settings, the above assumptions are challenged. Centralized control is simply not possible for large scale dynamic HACs. Moreover, approaches must be developed to consider not just what the optimal coalition looks like, but how the individual humans and agents, each with their own limited communication and computation resources, can negotiate with one another to form a coalition, without having explicit knowledge of the utilities and constraints of all the other actors within the system.

Addressing the decentralization issues is likely to involve local message passing approaches which draw on insights from the fields of probabilistic inference, graphical models and game theory. These allow coordination and coalition formation problems to be efficiently represented as a graph by exploiting the typically sparse interaction of the agents (i.e. not every agent has a direct interaction with every other). To date, however, these approaches have only addressed teams with tens of actors, while HACs will scale to hundreds or possibly thousands. Similarly, extant approaches have been developed with the explicit assumption that all of the actors engaged in the coordination have similar computational and communication resources; an assumption which will almost certainly not be valid within most HACs. Addressing the challenge of scaling up these approaches, such that they can deal with large numbers of actors with heterogeneous computational and communication resources, is likely to require principled approximations to be made (see [Rahwan *et al.*,

2013b] for promising work in this vein that uses well founded network flow optimization algorithms to address large-scale coalition formation problems).

Furthermore, previous approaches to forming teams and coordinating actors have typically assumed that complete and accurate knowledge regarding the utilities and constraints of the system is available to all. Whilst such assumptions may be valid within the small scale systems studied to date, they will not apply to larger HACs operating within dynamic environments where the sensing and communication capabilities of the actors are unknown. Previous work in this domain has addressed such uncertainty through the frameworks of Markov decision problems (MDPs) and partially observable MDPs. Now, while the Bayesian framework implicit in these approaches is well founded and principled, again it does not currently scale sufficiently to allow its use in large systems where time critical decisions must be made. As in the previous case, this requires novel computational and approximation approaches to be devised.

Most importantly, the novel approaches to HAC formation and operation must also address the needs of the humans within the system. Users will have to negotiate with software agents regarding the structure of the short-lived coalitions that they will collectively form, and then coordinate their activities within the resulting coalition. This represents a major departure for the mechanisms and techniques which have predominantly focused on computational entities with little regard for the ways in which users might form teams or consideration of their relationship to groupings such as teams. Consequently, the computational exploration must be balanced by a focus on the persuasion and engagement of human participants within collectives. For example, how might we understand and manage the conflict between the need to break up and reform teams with natural human preferences for stability and trust? How might users feel about the possibility of working across multiple teams simultaneously, and how might they feel about taking instructions from software agents? Initial work has begun to explore these questions within the context of mixed reality games [Moran *et al*., 2013] and social robotics [Fong *et al*., 2003]. However, to fully understand how these dynamics impact the requirements of the underlying team formation and coordination algorithms, we need to build, deploy and evaluate prototypical HACs in realistic settings.

Incentive Engineering

What will cause HACs to form and what will motivate them to work well together? How do we align the incentives of a set of actors, either individually or as a group, with the goals of the system designers to generate particular outcomes? Both of these endeavours are challenging whenever the behaviour of the actors is guided by individual and potentially conflicting motives [Dash *et al.*, 2003]. Now, while it is acknowledged that such actors may be influenced by incentives of many different types, most research to date has drawn upon micro-economics, focusing on monetary incentives and assuming the actors' utility functions are well defined and linear. Moreover, actors are typically considered to be *rational* in that they carry out complex computations to deduce their best action in equilibrium. Unfortunately, these assumptions often result in incentive mechanisms that are centralised and brittle in the context of open systems such as HACs.

HACs require us to reconsider many of the presumptions central to most current approaches to incentive engineering. They will involve actors that are boundedly rational [Kahnemann, 2003] and whose behaviour cannot always be controlled. For example, in disaster response settings, local volunteers (of different reliabilities) can be co-opted by their family members and friends (through their social network) and need to be coordinated to work alongside emergency responders from different agencies (with different capabilities). Moreover, humans and software agents may not always be receptive to monetary payments and may react better to social or intrinsic incentives. For example, in the DARPA Red Balloon Challenge, the incentive mechanism of the winning team was successful because it aligned individual financial incentives to find balloons with that of recruiting members from their social network and beyond [Rahwan *et al*., 2013a]. In contrast, essentially the same scheme attracted very few volunteers on the MyHeartMap Challenge (http://www.med.upenn.edu/ myheartmap/) which appealed mostly to altruistic motivations to save heart attack patients. Recently, Scekic *et al*. [2013] have surveyed and categorised such early examples of incentive mechanisms for social computing platforms. They note that the vast majority of current systems employ a simple contest between workers, from which a subjective assessment of the 'winner' is made. This winner is then typically rewarded financially for the work they have done. Far fewer systems make use of non-monetary incentives. Those that do, often focus on reputational issues. For example, they cite avvo.com which attracts large numbers of lawyers in the US to provide free response and advice to people visiting the website, and generates a reputation ranking of these lawyers based on the quality and timeliness of responses they provide. Now, the online ranking can clearly impact the chances of attracting customers to their private practices, but the effect is indirect, and admits a greater range of personal motivation, compared to a direct monetary payment. To date, however, comparatively little work has attempted to formally define this style of incentive mechanism. This is an important omission because we believe that bringing insights from behavioural economics and 'nudge' approaches to behaviour change [Kahnemann, 2003, Thaler and Cass, 2008], into the formal descriptions of mechanism design provided by game theory, is a promising point of departure to engineer the types of incentives that HACs require.

Now, even if the right incentives can be ascertained, the actors in the system may not perform to their best, either because they have limited capabilities or are inherently byzantine. For example, in systems such as AMT, where thousands of workers attempt micro-tasks for a few cents, strategic workers attempt to complete as many tasks as possible with minimum effort. Similarly, in citizen science projects such as Zooniverse, amateur scientists often select the tasks they most enjoy rather than those needed most for the project [Simpson *et al*., 2011]. Moreover, in long-lived interactions, human actors may suffer from fatigue and therefore their performance may degrade over time. However, naïvely filtering out the actors that cannot perform some tasks may mean their ability to perform other tasks properly is wasted. Hence, it is crucial to design mechanisms to ensure the incentives given to the actors to perform the tasks assigned to them are aligned with their capabilities and reliabilities. Against this background, initial work in crowdsourcing, and citizen science has, for example, demonstrated how to set the price paid for micro-tasks or how many tasks, of a particular type, each actor should be allocated to incentivise them to perform well [Tran-Thanh *et al.*, 2013]**.** Gamification approaches have also been successful in incentivising human participants to spend hours doing what would typically be viewed as

boring tasks [von Ahn and Dabbish, 2004]. Nevertheless, more work is needed to generalise these approaches and prove their efficacy in different application areas.

While the above challenges relate to how incentives are chosen and presented, the *computation* of these incentives in the context of HACs is also a major challenge. Indeed, the fact that HACs involve large numbers of actors means that computationally efficient algorithms need to be designed to enumerate and optimise the, possibly combinatorial, incentives to be given to a crowd, coalition or individual within a HAC. When the operation of such HACs unfolds over a long period of time, possibly involving many repeated interactions and negotiations, the schedule of incentives to be offered is an even greater computational challenge. In most cases, optimality will not be achievable and, hence, the goal should be to seek approximate solutions. Some relevant examples include algorithms to incentivise large numbers of electric vehicle owners to schedule charging their cars at different times to avoid overloading a local transformer [Robu *et al.*, 2013] and algorithms to approximate fair rewards for the participants of large teams [Michalak *et al.,* 2013].

Accountable Information Infrastructure

HACs will have a significant impact on the ways in which we think about the digital infrastructure that supports them. Specifically, we need to consider how the data underpinning can be shared. The provenance of this information is particularly critical. Here provenance describes which information data is derived from (what), the humans or agents responsible for it (who), and the methods involved in deriving it (how). In turn, the infrastructure processes provenance to assess information quality, to allow users to understand and audit the past behavior of HACs, and to help humans decide whether a HAC's decisions can be trusted.

The ways in which HACs operate requires us to reconsider some of the prevailing assumptions of provenance work. Provenance is generally thought of as being fine-grained, deterministic, and accurately and completely describing executions [Moreau, 2010]. This assumption is not valid in HACs, since human activities are both hard to capture and unreliable. Moreover, asynchronous communications may make provenance incomplete. Finally, the fine-grained nature of provenance makes it difficult for humans to understand. Addressing these challenges is crucial, and a variety of techniques are needed. For instance, probabilistic models built on provenance may help capture the uncertainty associated with what happened and abstraction techniques may allow common patterns to be collapsed, and thus, large graphs to be more manageable. These promising directions require the meaning of such provenance descriptions, and the kind of reasoning they enable, to be investigated. Given the potential size of HACs in terms of agents and humans, and also in terms of duration of execution, scalability of reasoning algorithms is also an important issue that requires further work.

The vision for an accountable information infrastructure is to help both humans and agents understand the decisions made and decide whether they can be trusted. Indeed, it is folklore that provenance can help derive trust and assess quality, but no principled approach, readily applicable to HACs, is currently available. In this context, the ability to learn from provenance is important as it has the potential to make provenance a rich source of information to establish trust, and also guide decision-making in HACs. In particular, given

that provenance information typically takes the form of a graph, some of the methods developed for graphs in general may be customizable, and potentially be executable efficiently. An example of such a solution is network metrics that summarize complex situations and behaviors in a convenient and compact way. Specifically, network metrics can be specialized to provenance graphs, helping characterize HACs' past behavior, in an application-agnostic manner [Ebden *et al.*, 2012]. Then, by applying machine learning techniques to provenance-oriented network metrics, we can label graphs and nodes to derive trust about agents or quality of data.

To date, existing infrastructure mechanisms tend to embody a "middleware" perspective, formalizing data models, developing algorithms, and engineering the integration of facilities such as provenance with applications. However, HACs need to understand and respond to the behavior of people and how this human activity is captured, processed and managed raises significant ethical and privacy concerns. Often at the core of these concerns is the manner in which people are separated from data collected about them. Specifically, in current infrastructures people are often unaware of the digital information they bleed, how this information is processed, and the consequential effects of the analytical inferences drawn from this data. Consequently, people are at an *ethical disadvantage* in managing their relationship with the infrastructure as they are largely unaware of the digital consequences of their actions and have no effective means of control or withdrawal. A HAC infrastructure will need to be accountable to people, allowing them to develop a richer and more bidirectional relationship with their data.

Developing an accountable infrastructure also responds to the call from privacy researchers such as Nissenbaum [2004] to understand and support the relationship between users and their data. Indeed, her Contextual Integrity theory frames privacy as a dialectic process between different social agents. Others have built upon this point, suggesting that a bidirectional relationship needs to be embedded into the design of services so they are recognised as inherently social [Steeves, 2009]. This suggests that users should have a significant element of awareness and control in the disclosure of their data to others [Palen and Dourish, 2003] and the use of this data by software agents. Establishing such bi-directional relationships also requires us to reframe our existing approaches to the governance and management of human data. Perhaps the most critical issues in this regard relate to seeking permission for the use of personal data within information systems. Current approaches adopt a transactional model where users are asked at a single moment to agree to an often quite complex set of terms of conditions. This transactional model is already being questioned in the world of bio-ethics, with Manson and O'Neill [2007] arguing for the need to consider consent as much broader than its current contractual conception. We suggest that HACs will similarly need to revisit the design principles of consent and redress the balance of agency towards the users [Luger, 2013].

References

G. D. Abowd, M. Ebling, G. Hung, H. Lei and H. W. Gellersen (2002) "Context-aware computing" *IEEE Pervasive Computing* **1** (3) 22-23.

L. von Ahn *et al*. (2008) "recaptcha: Human-based character recognition via web security measures" *Science* **321** (5895) 1465-1468.

L. von Ahn and L. Dabbish (2004) "Labeling images with a computer game" *Proc. SIGCHI Conf on Human Factors in Computing Systems, Vienna, Austria, 319-326*.

D. Ariely, A. Bracha and S. Meier (2007) “Doing good or doing well? Image motivation and monetary incentives in behaving prosocially” *American Economic Review* **99** (1) 544-55.

P. Buneman, J. Cheney and S. Vansummeren (2008) “On the expressiveness of implicit provenance in query and update languages” *ACM Transactions on Database Syst*ems **33** (4) 1-47.

R. K. Dash, D. C. Parkes and N. R. Jennings (2003) “Computational mechanism design: A call to arms” *IEEE Intelligent Systems* **18** (6) 40-47.

M. Ebden, T. D. Huynh, L. Moreau, S. D. Ramchurn and S. J. Roberts (2012) “Network analysis on provenance graphs from a crowdsourcing application” *Proc. 4th Int. Conf. on Provenance and Annotation of Data and Processes, Santa Barbara, USA, 168-182*.

T. Fong, I. Nourbaksha and K. Dautenhahn (2003) “A survey of socially interactive robots” *Robots and Autonomous Systems* **42** 143-166.

Y. Gil, E. Deelman, M. Ellisman, T. Fahringer, G. Fox, D. Gannon, C. Goble, M. Livny, L. Moreau and J. Myers (2007) “Examining the challenges of scientific workflows” *IEEE Computer* **40** (12) 26-34.

E. Horvitz (1999) “Principles of mixed-initiative user interfaces” *Proceedings of the SIGCHI conference on Human Factors in Computing Systems*, New York, USA, 159-166.

N. R. Jennings (2001) “An agent-based approach for building complex software systems” *Comms. of ACM* **44** (4) 35-41.

D. Kahneman (2003) "Maps of bounded rationality: Psychology for behavioral economics" *The American Economic Review* **93** (5) 1449-1475.

E. Kamar, Y. Gal and B. Grosz (2013) “Modeling information exchange opportunities for effective human-computer teamwork” *Artificial Intelligence Journal* **195** (1) 528-55.

T. Kifor *et al.* (2006) “Provenance in agent-mediated healthcare systems” *IEEE Intelligent Systems* **21** (6) 38-46.

A. Krause, E. Horvitz, A. Kansal, F. Zhao (2008) “Toward community sensing” *Proc. Int. Conf. on Information Processing in Sensor Networks, St Louis, USA,* 481-492.

E. Luger and T. Rodden (2013) “An informed view on consent for UbiComp” *Proc. Int. Joint Conf. on Pervasive and Ubiquitous Computing*, 529-538.

P. Maes (1994) “Agents that reduce work and information overload” *Comms. of ACM* **37** (7) 31-40.

N. C. Manson and O. O'Neill (2007) “Rethinking informed consent in bioethics” CUP.

T. Michalak , K. V. Aaditha, P. Szczepanski, B. Ravindran, and N. R. Jennings (2013) “Efficient computation of the Shapley value for game-theoretic network centrality” *Journal of AI Research* **46** 607-650.

S. Moran, N. Pantidi, K. Bachour, J. E. Fischer, M. Flintham and T. Rodden (2013) “*Team reactions to voiced agent instructions in a pervasive game” Proc. Int. Conf. on Intelligent User Interfaces*, Santa Monica, CA, 371-382.

L. Moreau (2010) “The foundations for provenance on the web” *Foundations and Trends in Web Science* **2** (2-3) 99-241.

L. Moreau, P. Missier, K. Belhajjame, R. B’Far, J. Cheney, S. Coppens, S. Cresswell, Y. Gil, P. Groth, G. Klyne, T. Lebo, J. McCusker, S. Miles, J. Myers, S. Sahoo and C. Tilmes (2013) “Prov-dm: The prov data model” W3C Recommendation REC-prov-dm-20130430, World Wide Web Consortium.

V. Naroditskiy, I. Rahwan, M. Cebrian and N. R. Jennings (2012) “Verification in referral-based crowdsourcing” *PLoS ONE* **7** (10) e45924.

H. Nissenbaum (2004) “Privacy as contextual integrity” *Washington Law Review* **79** (1) 119-158.

L. Palen and P. Dourish (2003) "Unpacking privacy for a networked world" *Proceedings of the SIGCHI conference on Human Factors in Computing Systems,* 129 – 136.

M. Paxton and S. Benford (2009) "Experiences of participatory sensing in the wild" *Proc. 11th Int. Conf. on Ubiquitous Computing,* Orlando, USA, 265-274.

T. Rahwan, S. D. Ramchurn, N. R. Jennings and A. Giovannucci (2009) "An anytime algorithm for optimal coalition structure generation" *Journal of Artificial Intelligence Research* **34** 521-567.

I. Rahwan et al. (2013a) "Global manhunt pushes the limits of social mobilization" *IEEE Computer* **46** (4) 68-75.

T. Rahwan, T-D Nguyen, T. Michalak, M. Polukarov, M. Croitoru, N. R. Jennings (2013b) "Coalitional games via network flows" *Proc. 23rd Int. Joint Conf. on Artificial Intelligence*, Beijing, China, 324-331.

S. Reddy, A. Parker, J. Hyman, J. Burke, D. Estrin and M. Hansen (2007) "Image browsing, processing, and clustering for participatory sensing" *Proc. 4th Workshop on Embedded Networked Sensors,* Cork, Ireland, 13-17.

V. Robu, E. H. Gerding, S. Stein, D. C. Parkes, A. Rogers, and N. R. Jennings (2013) "An online mechanism for multi-unit demand and its application to plug-in hybrid electric vehicle charging" *Journal of Artificial Intelligence Research.*

A. Rogers, A. Farinelli, R. Stranders and N. R. Jennings (2011) "Bounded approximate decentralised coordination via the max-sum algorithm" *Artificial Intelligence* 175 (2) 730-759.

O. Scekic, H.-L. Truong, and S. Dustdar (2013) "Incentives and rewarding in social computing" *Communications of the ACM* **56** (6) 72-82.

E. Simpson, S. J. Roberts, A. Smith and C. Lintott (2011) "Bayesian combination of multiple, imperfect classifiers" *Proc. 25th Conf. on Neural Information Processing Systems*, Granada, Spain.

V. Steeves (2009) "Reclaiming the social value of privacy" In *Lessons from the Identity Trail* (eds. I. Kerr, V. Steeves and C. Lucock) Oxford University Press, 193-208.

P. Stone, G. A. Kaminka, S. Kraus and J. S. Rosenschein (2010) "Ad hoc autonomous agent teams: Collaboration without pre-coordination" *Proc. 24th Conference on Artificial Intelligence*.

M. Tambe et al. (2005) "Conflicts in teamwork: Hybrids to the rescue" *Proc. 4th Int. Joint Conf. on Autonomous Agents and Multiagent Systems*, Utrecht, The Netherlands, 3-10.

R. H. Thaler and R. S. Cass (2008) "*Nudge: Improving decisions about health, wealth, and happiness*" Yale University Press.

L. Tran-Thanh, M. Venanzi, A. Rogers and N. R. Jennings (2013) "Efficient budget allocation with accuracy guarantees for crowdsourcing classification tasks" *Proc. 12th Int. Conf on Autonomous Agents and Multi-Agent Systems,* St Pauls, USA, 901-908.

M. J. Wooldridge and N. R. Jennings (1995) "Intelligent agents: theory and practice" *The Knowledge Engineering Review* **10** (2) 115-152.

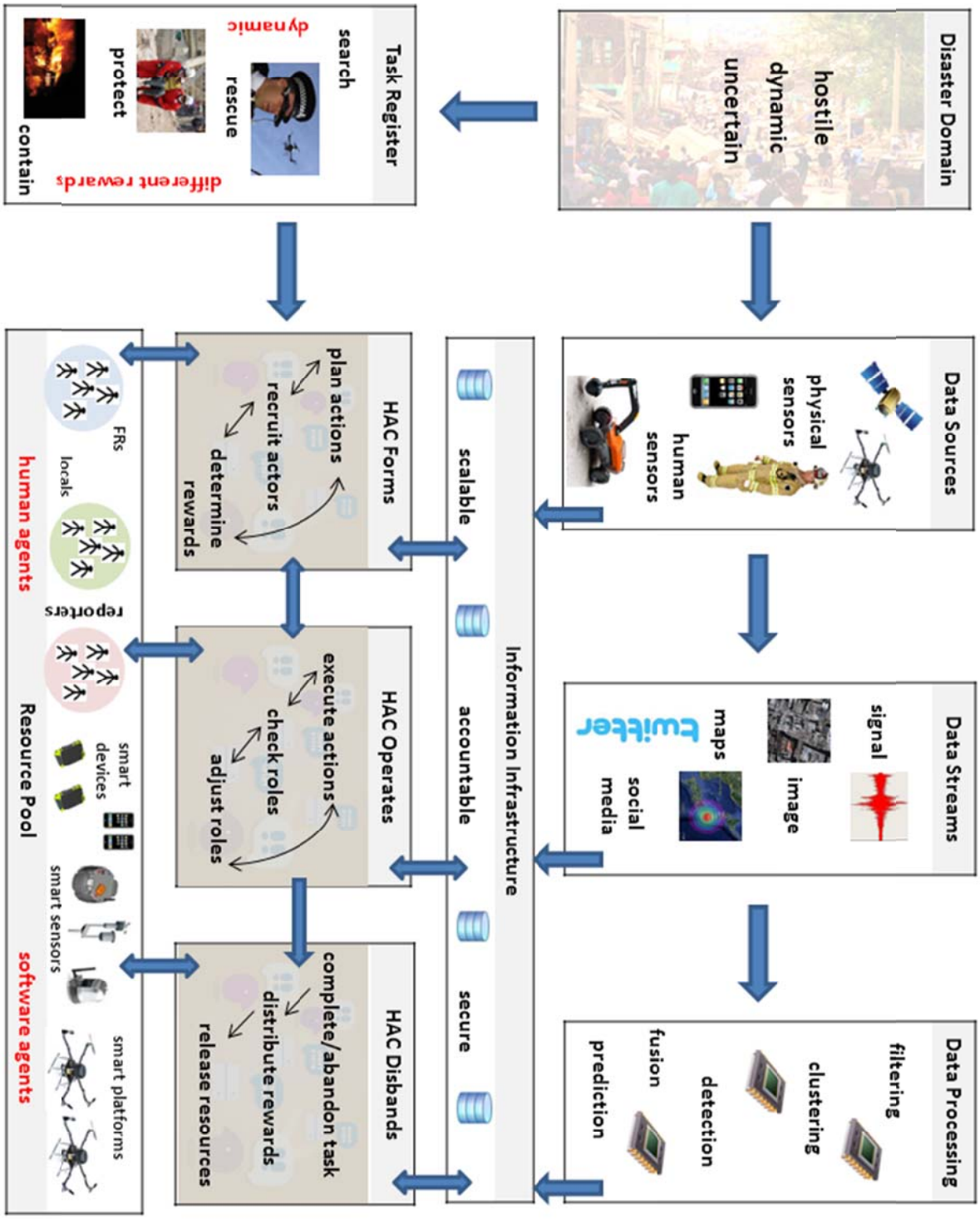

Figure 1: HAC System for Disaster Response.