Agency in complex information systems – Future research directions
Draft chapter / section for NESS Horizon 2020 Roadmap – v0.4 (15/12/2013)

Editors: David Hales, Rhett Gayle
See Annex 2 for list of contributors
Send input, comments, corrections etc. to dave@davidhales.com

Executive Summary

The concept of agency is central to our understanding of social systems. The way the concept is applied has implications for political economy, law, morality and engineering.

The engineering aspect is becoming more important because, increasingly, artificial (computational) agents form a significant part of 21st century social systems. For example, automated trading agents determine prices and volume on many global markets; semi-autonomous combat systems will make life or death decisions in future wars and high-level social policy will increasingly be informed by agent-based models of social phenomena.

We have consulted a number of thinkers in these areas, focused on the Information and Communication Technology (ICT) domain, asking the question: "what are the possible future application and research directions for agency in complex information systems". We have distilled the results of this consultation into a set of ambitious future research challenges that we report here.

Main findings

We identified the following timely and important future application domains in which agent research could have significant impacts:

- Financial and political stability
- Environmental sustainability
- Ethical and legal frameworks

We note each of these domains evidences *collective action problems involving multiple agencies situated in complex dynamic global networks*.

Based on this we propose eight novel future agent research areas encompassing aspects of:

- ICT engineering
- Policy design
- Legal and ethical frameworks

For each we briefly present the background and motivating problems in addition to specific research topics.

Structure of the report

In annex 1, we sketch the background to computational agent research focusing on the terminology used, major approaches and their limitations. This annex is valuable for those new to the area. In section 1, we outline some motivating high-level future application domains to which agent research may be applied. In section 2 we list a set of fundamental and applied research areas and associated topics that aim to

address the application domains. Annex 2 lists those who contributed to the report and describes the consultation process employed.

1. Motivating future application domains

Computational agents are already used in applications that are becoming increasing important for the day-to-day functioning of human social systems e.g. automated market trading, battlefield robotics and peer-to-peer information sharing. We call these *hard applications* because the behaviour of the agents directly affects and shapes the world. This involves designing and engineering computational artefacts that make use of the agency abstraction to perform a task autonomously.

Also agents are increasingly used in computer simulations in the form of agent-based models (ABMs) to help understand social phenomena such as markets, environmental impacts and riots. We call these *soft applications* because they help human actors to understand the world and this may in turn affect human social systems through policy or other actions. This involves using the agency abstraction to develop theories and models to aid the understanding of phenomena of interest[1].

The way in which agency is conceived, defined or programmed significantly influences fundamental aspects of social reality such as power relationships, institutional forms, collective decision making and more importantly possible foreseeable futures.

Here we briefly outline several high-level and challenging application domains that could be significantly addressed through both hard and soft agent research:

Financial and political stability increasingly dominate policy debate. Financial and technological innovations appear to have fundamentally changed the way economic and political systems operate. Specifically, traditional models of economic agency fail to capture crucial processes. Current controls by regulators appear to have limited influence in global, highly interconnected and technologically mediated networked systems. For example, the behaviour of high frequency trading algorithms (computational agents) on one exchange can have rapid knock-on consequences globally. Also, politically, new kinds of decentralised self-organising collective action such as public protests, revolutions or insider collusion are possible using social media platforms. At the same time centralised institutions find it increasingly difficult to control information and events leading to a crisis in democratic legitimacy and accountability.

Environmental sustainability and climate change has been proposed as the defining issue of our time. The difficulty of the issues raised often relates to the interdependence of outcomes. So for example, decisions made by one region can have dramatic knock-on consequences in another region. Resources, consumption and decision-making are unevenly distributed but solutions often require collective coordination. For example, common resources such as rivers, oceans and the atmosphere require protection against pollution or over exploitation. Yet traditional economic models and notions of rational action appear to preclude the development of adequate incentives. Also top-down planning and control has limited efficacy due to lack of adaptability to local conditions.

Ethical and legal frameworks struggle to operate in a globalised world. Global computer networks allow for capital and other forms of valuable information to be moved easily between different jurisdictions. Cloud-based services often remove any legal protections from the user since they are hosted in remote jurisdictions. Aggressive actions, both cyber and through physical robots (such as drones) may be initiated remotely, clandestinely and even autonomously[2] possibly outside of legal control. Intellectual property protection is increasingly untenable due to global open networks resulting in a techno-legal arms race that is costly, limits the free flow of information, and punishes innovation. In what way can ethical and legal codes be applied and adapted to such phenomena?

2. How agent research can address the application domains

Many of these challenges appear to arise from forms of *collective action problems situated in global and complex networks* where agents (both human and computational) interact in complex ways over dynamic large-scale networks producing outcomes that are not expected, planned or desirable.

Soft applications can be used to understand these issues by experimenting with new forms of agency and interaction structures that align desirable collective outcomes with individual behaviour. ABM allows for experimentation with radically different (and empirically informed) notions of agency and the structures and networks in which they are embedded and create. Results here could be used to inform the policy debate in each of the above areas.

Hard applications can be informed by these insights. The design of hard computational agents can be informed by more than purely engineering constraints or of-the-shelf notions of agency imported from earlier disciplines. For example ethical and legal frameworks can be applied and adapted towards a design paradigm for productive collective action in mutli-agent and peer-to-peer software systems.

We have identified the following **future research areas** related to computational agents in hard and soft applications:

- Collective-* for agent systems
- Agent-environment boundary
- Agent rationality
- Ethics, morality and law in agent design
- Policy design with agents
- Producing sharable agent knowledge
- Assistive and critical narratives with agents
- Agent adversarial scenarios

In the following subsections we briefly state the motivation and scope of each area and list some specific future possible research topics and questions associated with them.

2.1 Collective-* for agent systems

In order to model human social reality, design artificial computational agent societies and, increasingly, engineer and understand hybrid socio-technical systems, notions of collective behaviours and properties are essential.

Social realities often depend on collective properties that may not be easily reducible to the properties of individual agents because collective outcomes may not be the result of a simple additive or aggregate function. Often such collective properties relate to individual agent properties in interesting and counter-intuitive ways.

For instance, if people believe an epidemic is underway then that may change their behaviours. Here the issues of self-fulfilling and self-denying prophecies are relevant, and are related to recent work in dynamic epistemic logic on the paradoxes of public announcement - there are statements which when publicly announced become no longer true. Or consider the idea of a collective power to act, such as four people being able to move a table, when no individual has that power alone. These can be understood as collective-* issues in contrast with the existing area of self-* which focuses primarily on individual rules and tend to assume collective properties are reducible to them.

In order to model, understand and engineer these collective-* (collective-knowledge, collective-awareness, collective-rationality etc) properties within agent societies it is necessary to understand the relationship between collective-* and individual-* properties. This will facilitate design for collective-* properties that emerge bottom-up from individuals and, where necessary, to impose top-down control efficiently.

For example, what is the relationship between individual goals and collective goals in given contexts? In what sense do collective goals exist? How are conflicts between goals managed within collectives? This has obvious links to social choice theory and to political philosophy – areas that are becoming increasingly important within computer science. However, in order to understand how real collectives operate and hence can be engineered we need more fine-grained behavioural models than those used in mainstream economic theory.

Future research topics:

- *Designed emergence*. How do we balance the wisdom versus the madness of crowds for collective action problems?
- *Self-* to collective-**. In given contexts how do systems come to exhibit collective-* properties (e.g. collective-knowledge, collective-awareness, collective-rationality, collective-rights etc.)? What is the relationship between collective-* properties and self-* properties?
- *Fairness and Power in collective-* systems*. What collective-* properties can emerge bottom-up vs. being imposed top down? How can these be related to, and inform, political philosophy and social choice theory?
- *Management and control of collective-* systems*. How are collective-* systems affected through actions such as giving knowledge to one agent, making public broadcasts or inserting particular agent types.

2.2 Agent-environment boundary

For a given system (either artificial, real world, or hybrid) the boundary between an agent and its environment, while not completely arbitrary, can be drawn at different places for different purposes. This is perhaps obvious for software agents, where both agent and environment are made of code, but also applies to real world agents. Dennett explores this in discussing “free will” and personal responsibility[3]; Andy Clark with respect to how people incorporate tools and external representations into mental activity – the “extended mind” hypothesis[4].

Since these boundaries are flexible they may change over time – even from the point of view of the agent itself. Consider a robot building new actuators and sensors for itself or a transhumanist uploading “themselves” into some future computer.

This has implications for both hard (engineering) and soft (agent-based modelling) applications. It is important that an agent system designer realises that they have choices concerning where they place the agent-environment boundary also this may be dynamic and only partially under their control. In the context of agent-based modelling the position of agent boundaries should not be considered as given. For example institutions may be considered as agents since they take decisions and have goals but for a given social phenomenon there may be agency (for example, institutions, social class, waves of sentiment or other entirely novel organising constructs) that are not currently known.

In this latter case it may be possible to induce (or disprove hypotheses) from large dynamic data sets (big data) agencies that have hitherto been latent within our understanding of social reality. However humans often fail at this with a tendency to find agents where they do not appear to exist, such as fantastical conspiracies or supernatural entities. Consider how children ascribe agency to toys or imaginary friends. How can we ensure we are not victims of this? And more interestingly could we explore how such views emerge in human agents?

Future research topics include:

- *The extended agent mind*. How could the ideas of extended mind and dynamic agent-environment boundaries inform novel engineering design processes for agent systems including robotics?
- *Inducing agency from data.* Can “big data” be used to automatically induce the presence of previously unrecognised agents their goals and actions?
- *Setting the agent-environment boundary*. What constitutes a meaningful and useful agent-environment boundary for modelling different social phenomena – such as financial markets, riots, consumption patterns etc?

2.3 Agent rationality

Rationality is a wide concept that has different meanings in different disciplines. Generally it relates the determination of a “correct action” in a given situation. It therefore has descriptive, normative and predictive aspects.

Often rationality is defined relative to *a priori* goals (sometimes termed preferences) related to an agent. In this context an agent is said to be rational if it selects actions, given the information at hand, that would optimally attain those goals. This has been termed “narrow rationality”[5]. It can only be applied to situations in which agent boundaries and goals are known *a priori*. It is not possible, for example, to assess the rationality of goals *per se*.

Broad rationality, which is nearer to the common sense use of the term, relaxes these assumptions and applies to both goal formation and collective processes that may be located partially outside of the individual agent. This broader sense of rationality does not privilege the individual agent as the instigator of rational behaviour. Rather a *social process* is conceived. For example, the collective process of “science” might be seen as rational even if individual scientists are not[6].

Within economics and computational agent research rationality has often been associated with the narrow notion of individual agent utility optimisation[7]. More recently adaptive agents situated in dynamic networks have been explored which modify their behaviour based on past experience – for example through social imitation, individual learning or evolutionary processes.[8]

In human systems broad rationality appears evolutionarily prior to narrow rationality since tribe survival is a prerequisite for individual survival[9].

In connection with these issues there are four broad aspects of rationality that are underexplored in computational agent research: 1) social goals – where agents aim to improve some social aspect rather than individual aspect; 2) goal formation – where agents rationally form their goals; 3) collective or group rationality – where a) rational behaviour is not situated within individual agents but emerges from their social interactions in an historical process or b) team reasoning leading to collective behaviour without the need for social interaction or history; 4) understanding populations in which there are mixtures of agents using different kinds of rationality.

Future research topics:

- *Comparative computational rationality*. How can different notions of agent rationality be specified and compared?
- *Collective and individual rationality*. How can group (collective or social) rationality support individual rationality and vice versa?
- *Evolution of rationality*. Under what conditions do evolutionary processes produce rational outcomes (both individual and collective)?
- *Rational history*. What kinds of historical social process can be viewed as rational if at all?
- *Forming rational goals*. How can agent goals (or preferences) be rationally formed?

2.4 Ethics, morality and law in agent design

Morality refers to right and wrong or good and bad behaviour. In the context of human agency religious and philosophical traditions advance various ethical frameworks that specify what is and is not morally right.

Although related to both rationality and legal codes morality is often seen as a primary guide to behaviour in human social systems yet is comparatively underexplored in it's potential application to computational agents.

However moral codes *have* been used to program agent behaviour in complex information systems. For example, collective utility maximisation draws on a Utilitarian ethics[10]. Forms of reciprocity are compatible with the Golden Rule (or more generally the Categorical Imperative of Kant). The most widely deployed peer-to-peer system so far developed (the Bittorrent file-sharing protocol[11]) applies reciprocity *and* altruistic behaviour[12]. Hence ethical frameworks can be used as highly effective practical design principles for hard applications.

Both logical reason and evolutionary approaches have been applied in soft application areas including the development of deontological systems of reasoning and the evolution of cooperation[13] and altruism[14].

Where hard applications directly affect the wellbeing of humans – for example in financial markets (trading agents) and war (lethal autonomous robots) it may be useful to explicitly program and certify agents as following some ethical framework. This would allow those responsible for deploying them to make appropriate decisions.

It might be speculated that trust can be more easily established between agents (both computational and human) if they share similar ethical frameworks by offering greater predictability.

Legal codes (laws) regulate human social systems and rely on a notion of agency in order to function. Not only actions but also intentions are significant within Jurisprudence. This raises the question of how law might be applied to, or help to design, computational agents (and vice versa[15]). This is particularly significant given the recent developments in so-called Lethal Autonomous Robotics[16], financial trading algorithms and automated surveillance systems.

Future research topics:

- *Ethical "design patterns"*. Can ethical frameworks inform agent design for hard applications – that is to perform tasks that not able to be currently achieved?
- *Self-organised ethics.* Can agents formulate or evolve their own novel ethical frameworks suitable for given scenarios in which they are situated?
- *Trust through ethics*. Can ethical frameworks applied to agents and the moral acts they enable improve trust and / or predictability?
- *Computational morality*. Under what circumstances can one meaningfully ascribe moral agency to a computational artefact if at all? If not why not?
- *Computational Jurisprudence*. Can legal principles be adapted as practical design patterns for agent systems and vice versa? What forms could a "computational Jurisprudence" take?

2.5 Policy design with agents

Agent-based modelling (ABM) is increasingly used in policy applications. However, its full potential has not yet been realised because many policy areas are dominated by existing, yet less expressive, modelling approaches[17].

It was been argued that ABM should enter the policy mix by productively combining with other modelling approaches in a “robust decision-making approach”[18]. Such an approach allows for multiple, plausible, yet diverse scenarios to be generated and considered. In addition such approaches and models can include “theoretical plurality” providing multiple possible causal mechanisms for given phenomena.

Within such a context it will be important for ABM modellers to more clearly understand where the ABM approach is applicable and its relative strengths over other approaches – such as what scenarios can be generated *only* by ABM. For example, ABM approaches appear particularly suited to situations where information is dispersed among a large group of people and where there are heterogeneous populations – leading to cascades, threshold behaviour and tipping points etc.

Additionally ABM offers the potential for including policymakers and other stakeholders directly into the modelling process. Ideally policy modelling should be highly participatory and democratised. This can be achieved by, say, bringing models online and allowing anybody to “play” with them and suggest alternative ideas[19]. However, stakeholders may find it difficult to distinguish between model structure, parameterisations of the model and different scenarios.

Where ABM aim to directly affect public policy it is also necessary to understand how their use can reflexively relate to the phenomena being modelled[20].

It will be useful to build a body of evidence or examples where ABM has been used to productively inform policy. Such a portfolio would help with demonstrating and spreading good methods and practice to others. This is necessary because the idiosyncrasies in complex policy domains mean it is difficult to draw general policy conclusions. Rather one needs to demonstrate success (or otherwise) on a case-by-case basis.

Future research topics:

- *Building a portfolio of real world policy challenges* including modelling existing policies in addition to new and open policy challenges.
- *Comparative policy modelling*. Methods for combining and comparing policy orientated ABM at different levels of abstraction and also with different modelling techniques productively.
- *Participatory policy modelling* approaches involving construction of ABM through the direct involvement of stakeholders.
- *Wider democratisation of policy models* through novel online tools, games, and crowdsourcing.

2.6 Producing sharable agent knowledge

A powerful aspect of ABM is the relative ease with which people can *apparently* grasp a model[21]. The intuitive nature of the agency metaphor allows non-experts to construct narratives that make sense of the model behaviour. But there is a danger of over interpretation of the model resulting in ascribing explanatory and predictive power that it does not have.

It would appear that ABM's could be interpreted as forms of representation or narrative structure that compress time and virtual experience in a different way than traditional representations such as novels or movies – but similar to computer games. A single ABM can potentially generate an infinite set of plausible stories relating agent actions, interactions and outcomes. Also ABM often produces "never-ending" stores rather than "happily-ever-after" stories.

Different ABM can represent the same phenomena from different perspectives and at different levels of abstraction. Multiple model approaches are important for "wicked" and "superwicked" problems (where reflexivity is inherent and solutions cannot be routinised) such as many policy problems.

Additionally ABM often produce "emergent" properties which are not explicitly built-in to the model specification. In particular these are important when they result in macro to micro processes, where social structures shape agent behaviour, rather than only micro to macro effects – where agent behaviour shapes social structures. The former may be related to reflexivity.

How can we share knowledge from multiple models that is meaningful to different stakeholders (including the modellers themselves) such that they can critically evaluate and apply it? Do we require a new kind of rhetoric for understanding and communicating such knowledge?

Future research topics:

- *Modelling of Modelling*. Studying and modelling how expert ABM modellers create their models based on empirical analysis (e.g. video, interviews) of modellers at work. This could improve the ABM process itself and inform pedagogy.
- *Narrative Platforms*. Can "narrative platforms" be constructed that help non-expert stakeholders to critically evaluate, modify or construct ABM?
- *Legitimised Algorithms*. What legitimises an algorithmic representation of some social phenomenon?
- *Training for knowledge extraction*. What forms of training are necessary and possible for modelling and non-modelling experts in order to allow them to extract useful knowledge from ABM?

2.7 Assistive and critical narratives with agents

In both constructing and understanding computational agents the concept of narrative occupies a vital role. A narrative is a story that explains the behaviour of agents based on their worldview. It can therefore be both descriptive and prescriptive. Narratives may draw on values, goals and norms in addition to facts and constraints.

Narratives focus on the particular and specific in an historical way rather than the ahistorical, general and statistical. Yet they can inform understanding of the general. For example, a particular narrative may emphasise specific actions an agent takes in a given situation that leads to future consequences.

It is claimed by some that the role of agent-based models (ABM) is to produce – through simulation - consistent narratives that reflect those of the stakeholders being modelled[22]. Hence validation of such a model involves comparing output from the simulation model with the narratives presented by stakeholders. The ABM can then be a decision support tool for stakeholders based on their current view of social reality. We might call this *assistive* modelling because it assists stakeholders to see the implications of their existing view of social reality and helps to test intuitive dynamics. This does not mean that stakeholders may not be surprised or learn from such models – since they may not be aware of the implications of their assumptions.

An alternative view is that ABM can be used - through experimentation with many possible scenarios – to present alternative narratives that challenge, though perhaps explain, the existing views of stakeholders. We might call this *critical* modelling because it critically challenges stakeholders existing views. The critical approach may be useful for controversial applications that have political implications. Here use is made of counterfactual or alternative histories. It is less clear however, how these models should be validated.

ABM allows both assistive and critical narrative approaches – or a mixture of the two[23].

Future research topics:

- *Narrative specification*. How best can narratives be formally specified for use in agent programming?
- Agents as storytellers. How to program agents to productively explain their actions in terms of a narrative?
- *Narrative extraction*. How can narratives be extracted from agent-based models and empirical data? (so-called “thick data”).
- *Assistive and critical narratives*. When should assistive and / or critical approaches be used in agent-based models and how should they be validated?
- *Understanding path dependency*. How can narratives be used to reveal the importance and consequences of path dependency in a systems dynamics?
- *The individual within the system*. How can narratives be used to demonstrate / explain how individuals’ agency has effects across organisational scales?

2.8 Agent adversarial scenarios

Agent systems are often situated within, or used to study, adversarial scenarios. This means that the goals of different agents are often in opposition. This is clear in hard applications such as financial markets (where gain for one agent is a loss for another), battlefield robotics and cyber-warfare.

Historically these environments have been analysed using Game Theory (GT). GT has historical roots within Cold War politics and equilibrium mathematics and hence has limited application because communication and self-organising aspects are not considered; agent boundaries are strict and action selection follows a utility maximisation approach.

In the 21st century there are many adversarial scenarios involving more sophisticated notions of agency emerging from bottom-up coordination, distributed systems programming and various cultural and political phenomena such as 4th generation warfare (4GW)[24].

Adversarial scenarios can be highly productive in driving improvement in agent capabilities – this can be viewed in evolutionary or economic terms. In either case the idea is that competition drives innovation. On the other hand, it can lead to highly destructive or stalemate outcomes. Previous agent research has benefited from open competitions within adversarial scenarios such as collective robotics (through RoboCup tournaments) and financial markets (through auction tournaments).

Although there has been much work within soft applications on adversarial scenarios these are often highly abstract – drawing on simple games such as the Prisoner's Dilemma inherited from GT. More recently ideas from empirically grounded anthropology[25] have influenced agent modellers and designers[26].

Future research topics:

- *21st century adversarial scenarios*. How can the new emerging 21st century adversarial agent scenarios be detected, specified and modelled?
- *Agents as adversaries*. What kinds of agents are appropriate for given adversarial scenarios?
- *Reframing adversarial scenarios*. Can adversarial scenarios be reframed as cooperative scenarios through novel notions of agency?
- *Innovation through conflict*. How can adversarial scenarios be productively utilised to drive innovation processes? Can new tournaments be produced?
- *The cyber-social battlefield*. What kinds of agents and situational dimensions can productively characterise, and help to engineer, combined cyber and social conflict outcomes?

Annex 1. Brief background to agency in complex information systems

What is agency?

Agency refers to the nature of agents. Minimally an agent can be *viewed* as some entity (person or artefact) that performs actions to achieve some goal.

More generally an agent is situated is some environment, which is not part of the agent, from which the agent receives precepts (or inputs) and to which the agent performs actions (or outputs). An agent selects actions using some decision process in order to achieve their goal.

An agent may have several goals (and beliefs and other mental constructs) and may have complex decision processes or rules. Alternatively, in the minimal case, it may have a single goal and very simple decisions rules.

Agents have autonomy. Autonomy indicates that the actions taken by the agent are determined by the decision process associated with the agent and not by some entity external to the agent.

We view agency as an abstraction or metaphor that has value to the extent that it facilitates understanding, prediction or engineering of individual and social phenomena.

Different kinds of agents

The agency abstraction can usefully be applied in a number of contexts such as:

1. *Human* agent – we might consider this as the primary source of the agent metaphor. It is how we understand the social world and ourselves.
2. Computational *model* of an agent – computer programs that purport to model (often human but sometimes software) agents in given situations.
3. *Software* agent – a computational system that is understood and / or designed using the agent metaphor to perform some task.

We focus on the computational use of agency in 2-3) - although, of course, these take inspiration from, and have implications for, 1). Computational models of agency have been applied within the areas of computational social science, agent-based modelling (ABM) and complex systems. Their aim is to gain greater understanding of systems that contain many interacting agents. Examples include: understanding market failure, riots, land use change and innovation processes. Software agents have been applied within the areas of distributed artificial intelligence, multi-agent systems and, to a lesser extent, peer-to-peer systems. Their aim is to design and engineer systems that perform given tasks. Examples include: collective robotics, energy management, information sharing and distributed currency.

Computational agents are often described as either *Behavioural agents or Cognitive agents.* Behavioural agents do not store explicit goals and beliefs but rather follow behavioural rules. They act rather than deliberate and act. In cognitive agents goals and beliefs are *explicitly* represented within the agent and some form of deliberation or reasoning, based on a cognitive theory, informs action[27].

Traditionally computational agent research has clustered around the assumptions that guide the design of the decision rules (the program) that agents use to select appropriate actions over time. These assumptions are associated with approaches often derived from existing disciplines from the human sciences. Below are the major types found within the literature:

1. Utility optimising agent – from economics and game theory
2. Logical reasoning agent – from symbolic logic
3. Adaptive or learning agent – from psychology / machine learning
4. Evolutionary or social learning agent – from biology / anthropology
5. Probabilistic decision agent – from several disciplines

Often 1-2) are termed "rational" whereas 3-4) are termed "adaptive". However, these boundaries are not distinct or clearly defined. Also the actual behaviour of an agent is impossible to understand without reference to the environment in which it is embedded. Up to the present most work – particularly within the "rational" domains – assumes individual rather than collective goals although these frameworks do not preclude this. Approach 5) does not attempt to model any process of decision making but rather uses some dataset to induce probabilities of certain actions in certain contexts. This is generally associated with what is termed "individual based modelling" because agency is not explicitly modelled but individuals are.

Agent Organising principles and mechanisms

The way that populations of agents organise through interactions has been understood and designed based on a number of principles and mechanisms such as:

- Equilibrium (from economics, game theory)
- Evolutionary Stable Strategy (from evolutionary game theory)
- Self-organisation, self-adaptation, self-healing etc. (often called self-* and from biology, complex systems, artificial life)
- Contracts, joint plans and commitments (from distributed artificial intelligence, multi-agent systems)

Equilibrium approaches allow for the analysis of large systems without explicitly modelling individual agents because it is assumed that in given situations agents will eventually reach an equilibrium after which behaviour will not change. This relies on strong assumptions about possible interactions, outcomes and agent goals (often utility maximisation). It is widely employed within economics and game theory. Evolutionary game theory approaches can identify those agent actions (or strategies) that are stable against mutants (different strategies) under the assumption of differential reproduction based on fitness (often equated with utility). Contracts, joint commitments and other socio-symbolic mechanisms have been studied within distributed artificial intelligence and multi-agent systems. Self-Organising approaches (termed Self-*) apply a wide range of techniques originating in biology, physics, complex systems and artificial life, that propose, often, simple agent rules, that through interactions collectively self-organise interesting population level properties. However Self-* also encompasses top-down methods for addressing some types of application.

Strengths and Limitations

There are a number of limitations with existing approaches to computational agency that limit their application in many contexts, though each has its strengths. Equilibrium approaches suffer from strong assumptions concerning rationality, common knowledge and interaction structures that limit their applicability to many realistic applications, specifically where dynamics and heterogeneity are prevalent. However, where these assumptions do hold they allow for deductive mathematical proofs of systems of interacting agents. Economics and game theory research can be seen as attempting to advance this “tractability boundary” into areas where the assumptions do not hold (see figure 1).

Approaches from multi-agent systems and distributed artificial intelligence often require highly sophisticated cognitive agents with language level communication abilities. This limits their use in deployed applications due to computational tractability and lack of scalability. Also, in a similar way to equilibrium approaches, there is a focus on individual agent utility and goals rather than social aspects. However, the formal logical foundations used in much of this work allow for formal proofs of certain agent characteristics and the development of agent-orientated programming languages that have been used in practical applications in limited domains.

Self-* approaches are more diverse and permissive. They often rely on simulation models rather than analytical proof. This allows for applications in large-scale systems where dynamics predominate, uncertainty is high and use can be made of emergent phenomena. However, this limits predictability and the development of common methodologies. Furthermore this approach often necessitates very simple behavioural agents with no explicit goals or reasoning ability that limits generality so every application area requires new kinds of agent. It could also be argued that applying the agency concept to self-* systems stretches the abstraction too far. Many working in self-* would not use the term agent to describe their systems.

Summary

In this annex we have briefly outlined some of the background to computational agent research including the limitations of current approaches. It is not possible to do full justice to such a wide and diverse area in such a brief note. However, the aim has been to introduce the main concepts that are required to understand the substantive sections of this document. A much more detailed background can be found in previous work[28].

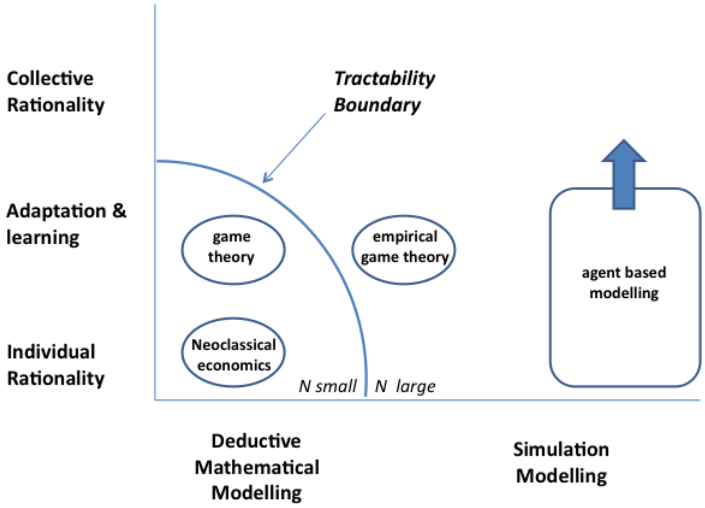

Figure 1 – Relating approaches along rationality and modelling method dimensions.

Notes

[1] Of course both hard and soft approaches are combined when ABM are used to model hybrid systems – where a significant part of the social phenomena incorporates hard applications. For example, financial market models may incorporate models of automated trading agents in addition to models of human traders. Simulations of peer-to-peer (P2P) systems often incorporate a "user model" capturing ways a person using the software may act.

[2] See for example the recent UNHCR report on Lethal Autonomous Robots (LARS) – Heyns (2013).

[3] See Dennett (2003).

[4] See Clark (2008).

[5] See Axelrod (2004) for a good discussion of narrow and broad rationality in the context of agent-based modelling.

[6] Other evolutionary processes can also be viewed this way. If agents are viewed as adaptive systems (learning how to behave from their interaction with the environment) then it is possible to apply the notion of "ecological rationality" which does not look for rationality as an internal property of an agent separable from it's environment (Bullock & Todd 1999).

[7] This may be due to similar assumptions, which have become orthodox economic theory, being inherited. Or it could be related to the existing paradigm of optimisation – a major area within computer science. It can also be argued such an approach follows a "keep it simple stupid" rule – although this can be disputed since in many contexts calculating optimal actions requires complex calculations and large quantities of information.

[8] Historically agent work that does not follow a utility optimising approach has tended to be termed "bounded rationality" (Simon 1957). Yet this term implies that there are deviations from some standard rationality that result from bounded knowledge or capacity to calculate or act and hence that the behaviour actually manifest is not rational in some way. In addition the focus is still on individual agents and goals.

[9] This insight has informed models of group selection (Wilson & Sober 1994) and cultural group selection (Boyd & Richerson 1985).

[10] See Kalenka & Jennings (1999).

[11] See Cohen (2003) for description of the the BitTorrent system.

[12] The Bittorrent system is often promoted with the tagline "Give and Ye Shall Receive".

[13] See Axelrod (1985) for seminal work using computer simulations to study evolution of cooperation.

[14] See Ridley (1996). Here a sociobiological theory of the evolution of altruism is presented with reference to in-group / out-group behaviour and selection.

[15] It has been argued that computer code is comparable to legal code in that it filters and shapes agent behaviour (Lessig 1999). Specifically that computer code has the same power to shape human society as law.

[16] See Heyns (2013).

[17] See for example: SIMIAN project workshop (2011). http://www.simian.ac.uk/about-simian/latest-news/55-or-society-sig-meet

[18] See Lampert et al (2002; 2003).

[19] See for example (Barreteau et al 2003) and geographical information systems (GIS) research (Evans et al 2004).

[20] See detailed discussion by McBurney (2012).

[21] See Axelrod (2004) for a discussion of this.

[22] See Moss (2008) for a discussion on this descriptive based approach to agent modelling.

[23] Previous work in operations research could aid this endeavour (Sterman 2000). Previous work in Geology and land use has focused on how ABM and narratives relate and may be developed (Millington et al 2012, Gotts & Polhill 2009).

[24] See Lind et al (1989) for seminal work in the area of 4GW.

[25] For example the work of Ostrom (1990) on common pool resource governance.

[26] For example see Pitt et al. (2012) on common pool resources. See Iruba model of the guerrilla war process (Doran 2005). Axelrod gives an overview of how ABM relates to other approaches within conflict scenarios (Axelrod 2004). Also there has been work developing and applying conflict resolution methods *within* cognitive agent processes (Broersen et al 2001).

[27] For the purpose of computational implementation cognitive agents are often approximated by behavioural agents. This is because in a suitably constrained environment it is possible to produce behavioural agents that perform more-or-less the same actions that cognitive agents would select and behavioural agents are generally more computationally tractable.

[28] See Luck, M. (2005) that considered the state of art and future prospects for agent-based computing back in 2005.

References

Axelrod, A. The Evolution of Cooperation. Basic Books, 1985.

Axelrod, R. (2004). Comparing Modeling Methodologies. Project Report, University of Michigan. http://www-personal.umich.edu/~axe/research/Comparing_Modeling_Methodologies.pdf

Barreteau, O. et al (2003). Our Companion Modelling Approach. Journal of Artificial Societies and Social Simulation vol. 6, no. 1 http://jasss.soc.surrey.ac.uk/6/2/1.html

Boyd, R., & Richerson, P. J. (1985). Culture and the evolutionary process. Chicago: University of Chicago Press

Broersen, J., Dastani, M., Hulstijn, J., Huang, Z. and Leendert van der Torre. (2001) The BOID architecture: conflicts between beliefs, obligations, intentions and desires. In Proceedings of the fifth international conference on Autonomous agents (AGENTS '01). ACM, New York, NY, USA, 9-16.

Bullock, S. and Todd, P. M. (1999) Made to measure: Ecological rationality in structured environments. Minds and Machines, 9, (4), 497-541. http://eprints.soton.ac.uk/261436/2/mm99.pdf

Clark, A. (2008). Supersizing the Mind: Embodiment, Action, and Cognitive Extension: Oxford University Press.

Cohen, B. Incentives build robustness in BitTorrent. In Proc. of IPTPS, 2003.

Davidson, P. (2001) Multi Agent Based Simulation: Beyond Social Simulation. Lecture Notes in Computer Science Volume 1979, pp 97-107. Springer.

Dennett, D. (1996), The Intentional Stance (6th printing), Cambridge, Massachusetts: The MIT Press.

Dennett, D. (2003) Freedom Evolves. Allen Lane The Penguin Press.

Doran, J. (2005) Iruba: An Agent-Based Model of the Guerrilla War Process. In “Representing Social Reality”, pre-proceedings of the Third Conference of the European Social Simulation Association (ESSA), Koblenz, Sept 5–9, 2005, ed Klaus G Troitzsch, Folbach. pp 198-205.

Evans, A., Kingston, R. and Carver, S. (2004) Democratic input into the nuclear waste disposal problem: the influence of geographical data on decision making examined through a web-based GIS. Journal of Geographical Systems, 6(2), 117-132.

Grimm, V. et al (2006). A standard protocol for describing individual-based and agent-based models. Ecological Modelling, Volume 198, Issues 1–2, 15 September 2006, Pages 115–126

Heyns, C. (2013) UNHCR Report of the Special Rapporteur on extrajudicial, summary or arbitrary executions, UNCHR, A/HRC/23/47.

John, D. (2000) Business Dynamics: Systems thinking and modeling for a complex world. McGraw Hill

Lempert, R. (2002) A new decision sciences for complex systems PNAS 2002 99 (Suppl 3) 7309-7313; doi:10.1073/pnas.082081699

Lempert, R.J., Popper, S. and Bankes, S. (2003) Shaping the Next One Hundred Years: New Methods for Quantitative, Long-Term Policy Analysis. Santa Monica, CA: RAND Corporation, 2003. http://www.rand.org/pubs/monograph_reports/MR1626.

Lessig, L. (1999) Code and Other Laws of Cyberspace. Basic Books.

Luck, M., McBurney, P., Shehory, O. and Willmott, S. (2005). Agent Technology: Computing as Interaction (A Roadmap for Agent Based Computing), AgentLink, 2005. ISBN 085432 845 9. http://www.agentlink.org/roadmap/al3rm.pdf

McBurney, P (2012): What are models for? Pages 175-188, in: M. Cossentino, K. Tuyls and G. Weiss (Editors): Post-Proceedings of the Ninth European Workshop on Multi-Agent Systems (EUMAS 2011). Lecture Notes in Computer Science, volume 7541. Berlin, Germany: Springer. http://www.dcs.kcl.ac.uk/staff/mcburney/downloads/pubs/2012/pm-2012-04.pdf

Millington, J. O’Sullivan, D. Perry, G. (2012) Model histories: Narrative explanation in generative simulation modelling. Geoforum 43, Pages 1025-1034.

Moss, S. (2008). Alternative Approaches to the Empirical Validation of Agent-Based Models Journal of Artificial Societies and Social Simulation vol. 11, no. 15 http://jasss.soc.surrey.ac.uk/11/1/5.html

Nicholas M. Gotts, J. Gary Polhill (2009) Narrative Scenarios, Mediating Formalisms, and the Agent-Based Simulation of Land Use Change. In Epistemological Aspects of Computer Simulation in the Social Sciences, Lecture Notes in Computer Science Volume 5466, pp 99-116. Springer.

Ostrom, E. (1990) Governing the Commons - The evolution of institutions for collective action. Cambridge University Press.

Pitt, J., Schaumeier, J., Artikis, A. (2012) Axiomatization of Socio-Economic Principles for Self-Organizing Institutions: Concepts, Experiments and Challenges. Trans. Autonomous Agent Systems (TASS) 7(4): 39 (2012).

Ridley, M. (1996) The Origins of Virtue. Penguin.

Simon, Herbert A. Models of Man (New York: Wiley, 1957)

William S. Lind, John F. Schmitt, Joseph W. Sutton, and Gary I. Wilson (1989) The Changing Face of War: Into the Fourth Generation. Marine Corps Gazette. October 1989, Pages 22-26

Wilson, D. S., & Sober E. (1994). Reintroducing group selection to the human behavioural sciences. Behavioral and Brain Sciences, 17, 585-564

Annex 2. Contributors and consultation process

The major ideas that comprise this report were based on a full day invite only workshop organised[1] at Imperial College, London, UK on Thursday 30th May 2013 and subsequent inputs from attendees. Several drafts of the report were circulated to attendees and others who expressed an interest for comments and additional input. Prior to the workshop potential topics for discussion were proposed and panels of interested attendees formed through an iterative e-mail consultation. This culminated in a final (entirely panel and discussion based[2]) programme sent to attendees as included below. The programme also included short bios of all attendees so they could become aquatinted with each other's areas of interest prior to the event.

Overall the workshop stimulated intense discussion and debate from which only a small subset could be presented in this report.

List of workshop attendees and contributors:

Name	Organisation	Email
Bromley, Jane	The Open University	j.m.bromley@open.ac.uk
Bullock, Seth	Southampton University	sgb@ecs.soton.ac.uk
Busquets, Didac	Imperial College	didac.busquets@imperial.ac.uk
Chli, Maria	Aston University	m.chli@aston.ac.uk
Doran, Jim	Essex University	doraj@essex.ac.uk
Elsenbroich, Corinna	Surrey University	c.elsenbroich@surrey.ac.uk
Fasli, Maria	Essex University	mfasli@essex.ac.uk
Fisher, Greg	Synthesis Think Tank	greg.fisher@synthesisips.net
Gal, Orit	Regents University	oritgal9@gmail.com
Gayle, Rhett	Colorado University	rhett.gayle@colorado.edu
Gotts, Nick	Independent Researcher	ngotts@gn.apc.org
Gruijic, Jelena	Imperial College	j.grujic@imperial.ac.uk
Hales, David	The Open University	dave@davidhales.com
Hermoso, Ramon	Essex University	rhermoso@essex.ac.uk
Johnson, Jeff	The Open University	j.h.johnson@open.ac.uk
McBurney, Peter	Kings College	peter.mcburney@kcl.ac.uk
Merali, Yasmin	Warwick University	yasmin.Merali@wbs.ac.uk
Millington, James	Kings College	james.millington@kcl.ac.uk
Neville, Brendan	Essex University	bneville@essex.ac.uk
Padget, Julian	Bath University	jap@cs.bath.ac.uk
Pitt, Jeremy	Imperial College	j.pitt@imperial.ac.uk
Riveret, Regis	Imperial College	regis.riveret@imperial.ac.uk
Rossiter, Stuart	Southampton University	s.rossiter@soton.ac.uk
Shardlow, Nigel	Sandtable Ltd.	nigel@sandtable.com

Others who received and / or commented on subsequent report drafts:

Edmonds Bruce	Manchester Met. University	bruce@edmonds.name
Gilbert, Nigel	Surrey University	n.gilbert@surrey.ac.uk
Jensen, Henrik	Imperial College	h.jensen@imperial.ac.uk
Musial-Gabrys, Katarzyna	Kings College	katarzyna.musial@kcl.ac.uk
Ormerod, Paul	Voltera Consulting	pormerod@volterra.co.uk
Polhill, Gary	Hutton Institute	Gary.Polhill@hutton.ac.uk
Serras, Joan	University College	j.serras@ucl.ac.uk
Wooldridge, Michael	Oxford University	mjw@cs.ox.ac.uk

[1] The workshop was organised by Jeff Johnson, Jeremy Pitt and David Hales.

[2] Powerpoint and extended paper presentations were not required in the workshop.

Workshop programme as sent to attendees:

[insert workshop programme here!]