Creativity through Autonomy and Interaction

Mark d'Inverno and Michael Luck
dinverno@gold.ac.uk
Department of Computing
Goldsmiths, University of London
London SE14 6NW

michael.luck@kcl.ac.uk
Department of Informatics
King's College London
London WC2R 2LS

January 20, 2012

Abstract.
In this paper we have sought to bring together several strands of our work, on motivation, autonomous agents and interaction between agents, to show how creativity can have a central place within what might be considered rather straightforward aspects of the design of modern computing systems. We review our previous work on the SMART agent framework, and re-interpret it in light of considerations of creativity arising from autonomy, motivation and contributing to the process of autonomous interaction. Here, behaviour is not prescribed but is determined in relation to motivation, leading to different, potentially creative outcomes for different individuals, especially during the process of interaction. Moreover, considering interaction as discovery imbues it with the same creative aspect as in scientific discovery, in which it can be argued that creativity plays a significant role in theory formation and revision. In fact, these are two sides of the same coin: in our view, the creativity in discovery arises from the motivation and autonomy of the individual involved.

Keywords: Autonomy, Motivation, Interaction and Creativity

1. Introduction

Much recent work on the design and development of multi-agent systems, together with all that they entail, has provided a means of enabling multiple different computational entities to work together effectively and efficiently. Fundamental to this work is the notion of interaction, first since agreements between the agents involved is typically established through interaction by means of communication to determine what to do and how to do it, and second since the enactment of joint work is typically established by explicit cooperation and coordination.

If agents are designed to be benevolent, or are simply invoked like functions or objects, then this interaction is in many ways straightforward. However, such systems are limited; the real value of multi-agent systems lies in the interaction of entities that are autonomous and can decide for themselves whether to engage with others, and to refuse to cooperate when it is not their own individual interest to do so. Moreover, such autonomous agents are particularly important in real world environments that admit an inherent uncertainty that must be considered if we are to cope with more than just toy problems. Here, agents must be autonomous, since they cannot know in advance the exact effects of their own actions or those of others. Agents must therefore be designed with a flexibility that enables them to cope with this uncertainty by evaluating it and responding to it in adequate ways.

This paper is concerned with the design of autonomous agents for interaction: what it means for an agent to be autonomous and what that entails for an appropriate model of interaction between autonomous agents. We argue that autonomy arises through the ability of an agent to generate its own goals, in support of a set of non-derivative motivations that cause the agent to act and reason. In this light, motivation also drives autonomous interaction, with agents engaging in communication processes that have uncertain outcomes (as they must in any real world context). Given this view, we develop a model for autonomous interaction that can be viewed as a process of discovery (like scientific discovery), making predictions about the outcomes of actions and communications, and adjusting an internal model of the world when predicted outcomes are not realised.

One interesting consequence of this view of both autonomy and of interaction, is that we can understand the involvement of *creativity* in two respects. First, behaviour is not prescribed but determined in relation to unique individual motivations, leading to different outcomes for different individuals, with the potential for creativity that entails. Second, considering interaction as discovery imbues it with the same creative aspect as in scientific discovery, in which it can be argued that creativity plays a significant role in theory formation and revision. In fact, these are two sides of the same coin: in our view, the creativity in discovery arises from motivation and autonomy of the individual involved.

The paper is an effort to bring together different aspects of previous work, reported in (d'Inverno and Luck, 1996; d'Inverno and Luck, 2004; Luck and d'Inverno, 1995; Luck and d'Inverno, 1998) in showing how autonomy and creativity are fundamental to agent behaviour and interaction, and seeks to re-interpret the work in that light. The paper continues, in the next section, with a discussion of creativity and its relation to interaction, and in Section 3, with a more detailed consideration of motivation and how it gives rise to autonomy. In Section 4, we review the SMART agent framework before showing how this supports goal generation and adoption in Section 5, and autonomous interaction in Section 6. The paper concludes in Section 7.

2. Creativity and Interaction

2.1. Creativity

In her seminal book (Boden, 2004), Boden essentially argues that creativity is the ability to produce things that are *new*, *surprising* and *valuable*. These include poems, compositions and other artworks or artefacts, such as paintings, sculpture or architecture, but also apply to scientific theories in the context of the progression of science and scientific discovery, as well as daily activities such as new jokes or the penning of a satirical cartoon.

In terms of newness, there are two notions of creativity: psychological or *P-creativity*, which is a process that provides a *new*, *surprising* and *valuable* idea to the person who created it; and historical or *H-creativity*, which is a process that provides an idea that has not been known to be reported at all.

In this respect, newness is in the eye of the beholder; what is new to one individual in this way may not be so to another. Similarly, surprise may arise simply from the unfamiliar (Macedo and Cardoso, 2002), it may fit into a style of thinking that existed previously but to which an observer had not made the link, or it may be a high-level concept that relates to the seemingly impossible. The idea of value is the most difficult to pin down, and can be understood through many parameters such as aesthetic, medical, financial and societal, for example, the value of which no two observers may agree. (It is perhaps the most personal element of creativity, which is why some authors have tried to avoid it and develop models of creativity that do not involve such concepts (Dorin and Korb, 2012).)

Boden's view of creativity thus offers plenty of scope for an entirely personal view of what is creative. We agree that creativity is an almost intangible but personal concept, and one that arises from the differences between individuals, and the ways in which they decide what to do and how to do it. Moreover, the idea that recognition of creativity is in the eye of beholder, with differences between individuals, suggests a very strong relationship to notions of autonomy by which individuals determine their own courses of decision and action.

It is also interesting, and relevant, to note that aside from artistic notions of creativity, perhaps the most prominent claim for creativity arises in scientific discovery, when new concepts that differ from prevailing views, seemingly appear through particular intangible characteristics of the scientist or scientific process. Indeed, the traditional view of scientific discovery holds that there is a clean and simple division between the contexts of discovery and justification. The context of discovery is concerned with the creation of hypotheses and theories, while that of justification is concerned with the testing of those theories and their subsequent refutation or continued use (at least temporarily). Discovery is deemed irrational and outside the scope of theories of scientific discovery, while the logical procedures of justification are capable of rational investigation (and by extension, automation). There are arguments against the rationality of justification, but these are limited and narrow, and shall not be considered here. However, the context of discovery is particularly problematic because it lies outside rigorous logical procedures, and is often explained by reference to insight, intuition, creativity, and a host of sociological and psychological factors. It has consequently been referred to as the *Aha! reaction* (previously referred to as the *Aha! experience* by psychologists of the Gestalt school, to label experiences in which an individual utters "Aha!" during a moment of revelation; see, for example, (Weisberg, 2006), for more on the *Aha! reaction*).

In considering agent interaction as a process of discovery, as we do in this paper, we thus see creativity in two aspects: first in the individual views of agents in driving their own behaviour through motivated autonomy, and second in the process itself that is analogous to the theory formation or revision of scientific discovery. In the next section, we consider autonomous interaction in more detail, and outline some key problems that motivate our particular approach, before proceeding to a consideration of how autonomy may be achieved in machines and, subsequently, its role in our model of interaction between autonomous entities.

2.2. AUTONOMOUS INTERACTION

In multi-agent systems, the interactions between agents are the basis for usefully exploiting the capabilities of others. However, such a pragmatic approach has not been the concern of many researchers who instead have sometimes focussed on small areas of interaction and communication, and in particular on specialised forms of intention recognition and interpretation. Indeed, in many existing models of interaction, agents are not autonomous.

Problem-solving can be considered to be the task of finding actions that achieve current goals. Traditionally, goals have been presented to systems without regard to the problem-solving agent, so that the process is divorced from the reality of an agent in the world. This is inadequate for models of autonomy that require an understanding of how such goals are generated and adopted, as we have considered previously (d'Inverno and Luck, 1996).

In traditional models in which the goals of one agent are adopted by another, goals are broadcast by one agent, and adopted by other agents according to their own relevant competence (Smith, 1980). This assumes that agents are already designed with common or non-conflicting goals that facilitate the possibility of helping each other satisfy additional goals. Negotiation as to how these additional goals are satisfied typically takes the form of mere goal-node allocation. Thus an agent simply has to communicate its goal to another agent for cooperation in the form of joint planning to ensue.

The concept of benevolence — that agents will cooperate with other agents whenever and wherever possible — has no place in modelling *autonomous* agents (Castelfranchi, 1990; Galliers, 1990). Cooperation will occur between two parties only when it is considered advantageous to each party to do so. Autonomous agents are thus *rational* agents (that are selfish in the sense of seeking to maximise their utility), and a goal (whether traditionally viewed as *selfish* or *altruistic*) will always be adopted so as to satisfy a *selfish* motivation. This view of goal adoption and allocation still dominates in one form or another, and can be seen in other models of interaction including communication based on Speech Act Theory (SAT), for example.

Indeed, SAT (Austin, 1962; Searle, 1969), underlies much existing work in AI (Campbell and d'Inverno, 1990), typically because as Appelt points out, speech acts are categorisable and can be modelled as action operators in a planning environment (Appelt, 1985). However, this work admits a serious flaw. Although the preconditions of these operators are formulated in terms of the understanding of the planning agent, the post-conditions or effects of these operators do not update the understanding of the planning agent, but of the agent at whom the action is directed (Allen, 1979). Yet no agent can ever actually know with any certainty anything about the effects of an action, whether communicative or otherwise. It is only through an understanding of the target agent and through observing the future behaviour of that agent, that one can discover the actual effects of the interaction. This uncertainty is inherent in communication between autonomous agents and must be a feature of any model of interaction that hopes to reflect this reality. While SAT underpins many models of communication, there remain many open issues, including this particular concern, that continues to engage the research community (Chopra et al., 2011).

Moreover, much work has modelled communicative actions in terms of mutual beliefs about the operator and its known effects (Perrault, 1990). This proposes to show not only how certain mental states lead to speech actions, but how speech actions affect mental states. We have argued previously (d'Inverno and Luck, 1996) that any account of autonomous interaction should only model the effects of an action upon the mental state of the agent initiating the interaction (or another single agent). Agents must make their own decisions about how and why they should take part in an interaction, and be aware of the consequences of doing so (d'Inverno et al., 2011; López y López et al., 2006).

In summary, there are several important claims here: first, an agent cannot be truly autonomous if its goals are provided by external sources; second, an agent will only adopt a goal and thus engage in an interaction if it is to its advantage to do so; third, the effects of an interaction cannot be guaranteed; fourth, the intentions of others cannot always be recognised; fifth, an agent can only know about itself with certainty (though it may also know some things about others). Note that the first claim requires goals to be generated from within. It is this internal goal generation that demands an explicit model of the motivations of the agent. The second claim requires a notion of advantage that can only be determined in relation to the motivations of the agent. The third and fourth claims demand that the uncertain nature of autonomous interaction be explicitly addressed. We argue that viewing autonomous interaction as motivated discovery provides us with a means for doing this. Finally, the fifth claim imposes constraints on the problem we are considering, and provides a strong justification for our concern with constructing a model of autonomous interaction from the perspective of an individual agent.

In what follows, we address the five issues detailed above, first by examining the mechanisms of autonomy and motivation that underpin the SMART agent framework, which is described next, and used as the basis for our model of interaction.

3. Autonomy and Motivation

3.1. Motivation

According to Halliday, the word *motivation* does not refer to a specific set of readily identified processes (Halliday, 1983), but is instead considered in terms of drive and incentive. Here, drives are related to physiological states such as the deprivation of food, hormones, etc., while incentives refer to external stimuli that affect motivation, such as the presence of food as an incentive to eat.

Some work from the field of psychology offers insight into the kinds of considerations we are concerned with, and in particular in their role in reasoning and action. Kunda, for example, informally defines motivation to be, "any wish, desire, or preference that concerns the outcome of a given reasoning task," and suggests that motivation affects reasoning in a variety of ways including the accessing, constructing and evaluating of beliefs and evidence, and decision-making (Kunda, 1990). Such arguments are supported by a large body of experimental research, but no attempt is made to address the issue of how motivations may be represented or applied in a computational context.

Computational work has also recognised the role of motivations, from very early work 30 years ago, to much more recent efforts (Kasmarik et al., 2005). For example, in some foundational work, Simon takes motivation to be "that which controls attention at any given time," and explores the relation of motivation to information-processing behaviour, but from a cognitive perspective (Simon, 1979). Sloman and Croucher elaborate on Simon's work, showing how motivations are relevant to emotions and the development of a computational theory of mind (Sloman, 1987; Sloman and Croucher, 1981).

More specifically, in proposing to develop a *computational architecture of a mind*, they make explicit mention of the need for a “store of ‘springs of action’ (motives)” (Sloman and Croucher, 1981). In the same paper, they try to explicate his notion of a motive as being a representation used in deciding what to do, including desires, wishes, tastes, preferences and ideals. The key feature of a motive, according to Sloman and Croucher, is not in the representation itself, but in its role in processing. Importantly, they distinguish between motives on the one hand, and *mere subgoals* on the other. “Sometimes,” they claim, “a mere subgoal comes to be valued as an end,” because of a loss of *reason* information. In this view, what are known as *first-order motives* directly specify goals, while *second-order motives* generate new motives or resolve conflicts between competing motives — they are termed motive generators and motive comparators. According to Sloman and Croucher, a motive produced by a motive generator may thus have the status of a desire. This relatively early work presents a broad picture of a two-tiered control of behaviour: motives occupy the top level, providing the drive or urge to produce the lower level goals that specify the behaviour itself. In subsequent work, the terminology changes to distinguish between *nonderivative* motivators or goals and *derivative* motivators or goals, rather than between motivators and goals themselves. Nevertheless, the notion of derivative and nonderivative mental attitudes makes one point clear: that there are two levels of attitude, one that is in some sense innate, and that gives rise to the other, which is produced as a result of the first.

Consider the example of crossing a river. The goal is to get to the other side of the river, but the way in which that goal will be achieved depends on the motivations that generated the goal. In normal circumstances, one would look for a bridge or a boat to get across. Though this may involve more effort than swimming across immediately, it is preferable because it is more comfortable. If there are urgent reasons for crossing the river, however, such as being pursued by a wild animal, then it might be better to jump into the river and swim across instead despite the discomfort this may cause. In both cases, the motivations are different and their strengths are different, but the goal remains the same. Motivations act as a control strategy for achieving the goal, directing reasoning, and providing it with the flexibility and strength that is often lacking.

3.2. Classes of Motivation

Much of the psychological literature stresses the distinction between two kinds of motivated reasoning phenomena (see (Kunda, 1990) for a review). These are reasoning in which the motivation is to arrive at an accurate conclusion, and reasoning in which the motivation is to arrive at a particular directed conclusion (towards some particular end rather than simply for knowledge).

Kunda (1990) suggests that both kinds of motivation affect reasoning by influencing the choice of beliefs and strategies applied to a given problem, but that they differ in the following respect: accuracy goals lead to the use of those beliefs and strategies that are considered most appropriate in getting the correct result, while directional goals lead to the use of those that are most likely to give the desired though perhaps inaccurate result. According to Kunda, accuracy goals thus demand greater (cognitive) effort on reasoning, more careful attendance to relevant information, and its deeper processing with more complex reasoning strategies. Directional goals impose constraints on "search and belief construction" that lead to support for the desired conclusion.

Similar distinctions have also been noted by others. In artificial intelligence, Ram and Leake (1991) describe two classes of goals motivating explanation at a lower level: knowledge goals that reflect an internal need for information, and goals based on accomplishing tasks in the external world. In psychology, Barsalou has distinguished between explicit problem-solving goals and implicit orientation goals for maintaining a world model (Leake and Ram, 1993). In education, Ng has distinguished task completion goals (such as completing an assignment) from instructional goals (what the assignment is intended to teach,) and knowledge-building goals, which relate to a student's own purposes and agenda for learning (Leake and Ram, 1993). Yet another formulation of this distinction is characterised as exploration (knowledge and accuracy goals) versus exploitation (directional, task-based goals) in a number of domains. All these are mirrored in the division of motivations below into knowledge motivations and action motivations.

3.2.1. *Motivations for Knowledge*

The motivation relating to the discovery of knowledge can be found everywhere, even in very limited models of simple creatures, either explicitly, or by a different name such as curiosity (eg., (Maes, 1991; Schmidhuber, 1991)). Any motivation that leads to the exploration of environment to discover more can be regarded as a motivation for knowledge. The desire for knowledge is relatively constant — even when action is taken to achieve some unrelated goal (to satisfy an unrelated motivation), it provides information that may be used to update a repository of knowledge. Consider, for example, eating a green banana because of hunger. Eating the banana not only satisfies the hunger motivation, but it also provides the knowledge that green bananas are not sweet. Such knowledge is always of interest and we are always motivated to acquire new knowledge even if it results from other actions.

3.2.2. *Motivations for Action*

Other motivations can be said to come under the broad heading of motivations for action. In this case, the motivations lead to the execution of certain actions and consequently to the manipulation of the environment in order to achieve goals. Traditional planning systems, for example, are motivated for action in that they generate plans for effecting changes in the world. These motivations are thus action motivations, and include motivations such as hunger, laziness and pleasure that lead to the taking of particular actions (or exhibition of behaviour). Action motivations vary in strength depending on circumstances; their strength may increase to a point at which they demand satisfaction, and also decrease once they have been satisfied. In the example of crossing the river when being chased by a wild animal, the strength of the fear motivation, say, caused the immediate action of swimming across the river. After having satisfied the motivation by fleeing across the river, the relative safety might lead to the strength of the fear motivation decreasing substantially.

An example illustrating this difference between motivations for knowledge and those for action is Crick and Watson's discovery of the double helix of DNA. In attempting to become the first to discover the structure of DNA, they used *quick and dirty* rather than the most reliable methods. Their first attempt at a model was a fiasco, according to Crick (1988), partly because of "ignorance" on his part, and "misunderstanding" on Watson's. By contrast, work by Wilkes and Franklin was progressing slowly as they concentrated on using their experimental data as fully as possible, and avoided resorting to guessing the structure by trying various models. Crick states that Franklin's experimental work was first class and could not be bettered, while Watson simply wanted to get at the answer as quickly as possible by sound methods or flashy ones. While the actual motivations of the individual researchers cannot be known, their apparent motivations can be *characterised* as motivations for knowledge, which demand accuracy and reliability, and motivations for action, which demand whatever behaviour will lead to the desired result.

Of course, the same action could mitigate multiple motivations, including both those for action and those for knowledge at the same time. Clearly, however, the degree of mitigation would be different for different motivations, leading to an ability to distinguish between actions on that basis, and on the basis of the importance or relevance of the motivations in a given context, as we see below.

3.3. AUTONOMY AND MOTIVATED BEHAVIOUR

The notion of autonomy has associated with it many variations of meaning. According to Steels, autonomous systems must be automatic systems and, in addition, they must have the capacity to form and adapt their behaviour while operating in their environment. Traditional artificial intelligence systems, such as planners, for example, and many robots, are automatic but not autonomous — they are not independent of the control of their designers (Steels, 1995). *Autonomous* systems, by contrast, are independent and exercise self-control. To do this, we have argued previously (Luck and d'Inverno, 1998), they must be motivated.

A given stimulus does not always evoke the same response. If the external situation is constant, differences in response must be ascribed to changes in the internal state of the responding agent. These differences are due to the motivations of the agent. In other words, the same action may be beneficial to differing degrees and with respect to different motivations at different times.

In seeking to provide a computational model for motivation, in support of the functionality and purpose described above, we can consider an agent as possessing a fixed range of identifiable motivations of varying strength. These motivations can be regarded as being innate, and certain behaviours may be associated with one or more motivations. For example, the behaviour of feeding is associated with the motivation of obtaining food, or hunger. In most cases, the execution of such a behaviour reduces the strength of the associated motivations, so that in the case of feeding, the motivation to obtain food is reduced. These behaviours are known as *consummatory* behaviours; other behaviours that are not associated with any particular motivation, but which make the conditions of a consummatory behaviour come true, are known as *appetitive* behaviours. For example, a go-to-food behaviour might make the conditions (that there is food within reach) of the feeding behaviour become true.

We can thus define autonomous agents to be agents with a higher-level control provided internally by motivations. Thus we can specify motivations of curiosity, safety, fear, hunger, and so on. In a simple agent design, we might then associate the motivation of curiosity with the goal of exploring an environment which, in turn, is associated with the actions required to achieve such results. Motivations will also vary over time according to the internal state of an agent. For example, if an agent spends a long time without food, then the hunger motivation will increase. When the agent feeds, the hunger motivation will decrease.

Each motivation thus has a strength associated with it, either variable depending on external and internal factors, or fixed at some constant value. In a very simple model (Luck and d'Inverno, 1998), motivation can thus be represented by a triple, $\langle m, v, b \rangle$ known as an *m-triple* where m is the kind of motivation, v is a real number, the strength (or intensity (Sloman, 1987)) value associated with that motivation, and b is a boolean variable taking the value *True* when the strength value, v , is fixed, and *False* when it is variable.

An autonomous agent can be regarded as embodying a set of n motivations, M, which comprises the m-triples, $\langle m_1, v_1, b_1 \rangle \ldots \langle m_n, v_n, b_n \rangle$. Thus the set of motivations, M, is a function of the kind of agent being considered, while the state of each motivation in this set at a particular point in time is a function of an instance of a particular kind of agent and its environment together. In order to act on motivations, a threshold value for strength may be necessary, which must be exceeded to force action. Alternatively, the highest strength value may be used to determine the motivation currently in control.

4. The SMART Agent Framework

We begin by reviewing previous work (d'Inverno and Luck, 1997; Luck and d'Inverno, 1995), but addressing only those aspects needed to provide an account of autonomous agents and autonomous interaction. We use the Z notation (Spivey, 1992) to formalise these notions and, though we assume some familiarity, the meaning should be clear.[1] In short, we propose a four-tiered hierarchy that forms the basis of our formal SMART framework comprising *entities*, *objects*, *agents* and *autonomous agents* (Luck and d'Inverno, 1995). (SMART is an acronym for Structured and Modular Agents and Relationship Types.) The basic idea underlying this hierarchy is that all components of the world are entities. Of these entities, some are objects, of which some, in turn, are agents and of these, some are autonomous agents.

As specified below, an *entity* comprises a set of *motivations*, a set of *goals*, a set of *actions*, and a set of *attributes* such that the attributes are non-empty. Entities can be used to group together attributes into a whole without any *functionality*. They serve as a useful abstraction mechanism by which they are regarded as distinct from the remainder of the environment, to organise perception.

While goals are sets of attributes that describe some state of affairs in the environment, motivations cannot be described in terms of the environment, and instead are high-level non-derivative drivers of agent activity.

$$
\begin{array}{l}
[\mathit{Attribute}, \mathit{Action}, \mathit{Motivation}] \\
\mathit{Goal} == \mathbb{P}\,\mathit{Attribute}
\end{array}
$$

$$
\begin{array}{|l}
\multicolumn{1}{l}{\mathit{Entity}} \\
\hline
\mathit{attributes} : \mathbb{P}\,\mathit{Attribute} \\
\mathit{capabilities} : \mathbb{P}\,\mathit{Action} \\
\mathit{goals} : \mathbb{P}\,\mathit{Goal} \\
\mathit{motivations} : \mathbb{P}\,\mathit{Motivation} \\
\hline
\mathit{attributes} \neq \{\,\} \\
\hline
\end{array}
$$

[1] The interested reader may consult any of several Z text books, such as (Bowen, 1996; Hayes, 1993; Spivey, 1988; Woodcock and Davies, 1996) for further introduction to the notation.

An *object* is then an entity with abilities that can affect the environment in which it is situated.

Object

Entity

$capabilities \neq \{\ \}$

Now, an *agent* is just an object, either that is useful to another agent in terms of satisfying that agent's goals, or that exhibits independent purposeful behaviour. In other words, an agent is an object with an associated set of goals, but one object may give rise to different instantiations of agents with different goals.

The idea behind this is that even non-computational entities can be understood as agents in the sense that the relationship between an object serving some purpose for someone provides useful information. For example, in (Luck and d'Inverno, 1995), this was illustrated by reference to a cup containing some coffee. The point is that although there is nothing computational about this example, knowledge of the relationship between the cup and the person who is drinking from it allows a waiter to understand that they should not remove the cup since it is in use. In this respect, there is a social relationship between the cup and the drinker, with the cup being an agent of the drinker.

Agent

Object

$goals \neq \{\ \}$

This notion of agency relies upon the existence of other agents to provide the goals that are adopted to instantiate an agent. In order to escape an infinite regress of goal adoption, however, we can define *autonomous agents*, which are just agents that generate their own goals from motivations.

AutonomousAgent

Agent

$motivations \neq \{\ \}$

4.1. Perception and Modelling of Agents

An environment *Env* is defined to be a simple set of attributes that describes all of the features of the world. The environment thus represents all possible percepts in a uniform way. It is convenient also to define a *View* to be the *perception* of an *Env* by an agent using the same type.

$$\begin{array}{l} \mathit{Env} == \mathbb{P}\,\mathit{Attribute} \\ \mathit{View} == \mathbb{P}\,\mathit{Attribute} \end{array}$$

Before considering agents that can interact with others (or social agents), we must first consider *situated* agents. In this respect, an agent in an environment can perceive certain attributes which we refer to as the *possible available percepts* (*posspercepts*) subject to its capabilities and current state but, due to limited resources, may not be able to perceive all attributes. As a result, the action that an agent chooses to undertake is based on a subset of percepts, the *actual percepts* (*actualpercepts*), determined in relation to its current goals and motivations.

Thus, in the schema below for agent perception, *perceivingactions* is a subset of agent capabilities. Two functions specify what the agent perceives: *canperceive* is applied to the current environment and the agent's capabilities to give potential percepts and *willperceive* is applied to its motivations, goals and the current environment to give those attributes that are actually perceived.

$$\begin{array}{|l} \hline \mathit{AutonomousAgentPercepts} \\ \mathit{AutonomousAgent} \\ \mathit{perceivingactions} : \mathbb{P}\,\mathit{Action} \\ \mathit{canperceive} : \mathit{Env} \to \mathbb{P}\,\mathit{Action} \nrightarrow \mathit{View} \\ \mathit{willperceive} : \mathbb{P}\,\mathit{Motivation} \to \mathbb{P}\,\mathit{Goal} \to \mathit{Env} \to \mathit{View} \\ \hline \end{array}$$

An *interaction* with the environment occurs as a result of performing actions in it. We provide the highest possible functional description of the effects on the environment by introducing the *effectinteraction* function in the axiom definition below. This is applied to the current environment, and the actions taken within it, to return a new environment that results from the actions being performed.

$$\begin{array}{|l} \mathit{effectinteraction} : \mathit{Env} \to \mathbb{P}\,\mathit{Action} \nrightarrow \mathit{Env} \end{array}$$

To specify the actions of an autonomous agent, the next schema includes the *AutonomousAgent* schema and an action-selection function that is determined in relation to the motivations, goals, perceived environment and environment of the agent. The function gives the set of actions the agent selects in order to achieve some goal.

$$\begin{array}{l}
\underline{\quad \mathit{AutonomousAgentAct} \quad} \\
\mathit{AutonomousAgent} \\
\mathit{autoactions} : \mathbb{P}\,\mathit{Motivation} \to \mathbb{P}\,\mathit{Goal} \to \mathit{View} \to \mathit{Env} \to \mathbb{P}\,\mathit{Action}
\end{array}$$

We also need to define the state of an agent by including the previous two schemas, but now situating it within an environment *env*. This presents the agent with a set of potential observable percepts that are a function of both the environment and its set of perceiving capabilities. However, as discussed above, and due to perceptual limitations, the actual percepts will be a subset of those possible, determined in relation to the current goals and motivations.

$$\begin{array}{l}
\underline{\quad \mathit{AutonomousAgentState} \quad} \\
\mathit{AutonomousAgentPercepts} \\
\mathit{AutonomousAgentAct} \\
\mathit{env} : \mathit{Env} \\
\mathit{posspercepts}, \mathit{actualpercepts} : \mathit{View} \\
\mathit{willdo} : \mathbb{P}\,\mathit{Action} \\
\hline
\mathit{actualpercepts} \subseteq \mathit{posspercepts} \\
\mathit{posspercepts} = \mathit{canperceive}\ \mathit{env}\ \mathit{perceivingactions} \\
\mathit{actualpercepts} = \mathit{willperceive}\ \mathit{motivations}\ \mathit{goals}\ \mathit{posspercepts}
\end{array}$$

All this defines autonomous agents, but we require more if they are to engage in interaction or communication episodes. Specifically, for effective interaction and communication, an agent must be able to group the attributes that make up the environment into entity-describing models so that it can identify the other individuals in the world. Of course, some such models may not contain *fully* detailed information of other agents, and may instead omit information about some aspects. For example, a model of another autonomous agent might include information only about their actions but not their goals.

$$\begin{array}{l}
\mathit{ObjectModel} == \mathit{Object} \\
\mathit{AgentModel} == \mathit{Agent} \\
\mathit{AutoAgentModel} == \mathit{AutonomousAgent}
\end{array}$$

Given this, we can define a social agent to be an agent that is aware of other agents, and their role and function, through these models. The schema below includes models of other autonomous agents available to a social agent as *modelautoagents*.

$$\begin{array}{|l}
\hline \textit{SocialAgent} \\
\textit{AutonomousAgentState} \\
\textit{modelobjects} : \mathbb{P}\,\textit{ObjectModel} \\
\textit{modelagents} : \mathbb{P}\,\textit{AgentModel} \\
\textit{modelautoagents} : \mathbb{P}\,\textit{AutoAgentModel} \\
\hline
\end{array}$$

As a result, and in summary, a multi-agent system consists of a set of entities, some of which are objects. Of these objects some are agents and of these agents some are autonomous. Those autonomous agents that can, in addition, model other agents are *social agents* and these are the focus of this paper: agents that are motivated, autonomous and can model others.

$$\begin{array}{|l}
\hline \textit{MultiAgentSystem} \\
\textit{entities} : \mathbb{P}\,\textit{Entity} \\
\textit{objects} : \mathbb{P}\,\textit{Object} \\
\textit{agents} : \mathbb{P}\,\textit{Agent} \\
\textit{autonomousagents} : \mathbb{P}\,\textit{AutonomousAgent} \\
\textit{socialagents} : \mathbb{P}\,\textit{SocialAgent} \\
\hline
\textit{autonomousagents} \subseteq \textit{agents} \subseteq \textit{objects} \subseteq \textit{entities} \\
\hline
\end{array}$$

4.2. Goal Generation and Adoption

The framework described above involves the generation of goals from motivations in an autonomous agent, and the adoption of goals by, and in order to instantiate, other agents. In this section, we build on earlier initial work in outlining goal generation and adoption (Luck and d'Inverno, 1996; Luck and d'Inverno, 1998). In this context, an autonomous agent will try to find a way to mitigate motivations, either by selecting an action to achieve an existing goal as above for simple agents, or by retrieving a goal from a repository of known goals. Thus, our model requires a repository of known goals that capture knowledge of limited and well-defined aspects of the world. These goals describe particular states or sub-states of the world, with each autonomous agent having its own such repository.[2]

[2] Of course, since this repository of goals determines how an agent can satisfy its motivations, the discovery of new ways of doing so (through new goals that enrich the goal library) is also interesting. For example, by applying analogical reasoning new goals may be identified that satisfy motivations and extend the repository. However, while an important issue, this is is outside the scope of this paper, and we leave its consideration to future work.

As we have described previously (Luck and d'Inverno, 1996), in order to retrieve goals to mitigate motivations, an autonomous agent must have some way of assessing the effects of competing or alternative goals. Clearly, the goals that make the greatest positive contribution to its motivations should be selected unless a greater motivational effect can be achieved by dropping some subset of its goals. The motivational effect of generating or dropping goals not only depends on the motivations, but also on the goals of the agent. For example, an autonomous agent should not generate a goal that it already possesses or that is incompatible with the achievement or satisfaction of its existing goals.

Formally, the ability of autonomous agents to generate their own goals is specified in the schema, *AssessGoals*, which describes how autonomous agents monitor their motivations for goal generation. First, the *AutonomousAgentState* schema is included, and a new variable, *goallibrary*, is declared to represent the repository of available known goals. Then, there are two functions to evaluate the benefit of generating and dropping current goals. The *motiveffectgenerate* function returns a numeric value representing the motivational effect of satisfying a new set of additional goals with a set of motivations, current goals and current perceptions. Similarly, the *motiveffectdestroy* function returns a numeric value representing the motivational effect of removing some subset of existing goals with the same set of motivations, goals and perceptions. The predicate part specifies that the current goals must be in the goal library. For ease of expression, a function is also defined that is related to *motiveffectgenerate*, called *satisfygenerate*, which returns the motivational effect of an autonomous agent satisfying an additional set of goals. The function *satisfydestroy* is analogously related to *motiveffectdestroy*.

$$
\begin{array}{|l}
\multicolumn{1}{l}{\underline{\quad \mathit{AssessGoals} \quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad}} \\
\mathit{AutonomousAgentState} \\
\mathit{goallibrary} : \mathbb{P}_1\, \mathit{Goal} \\
\mathit{motiveffectgenerate} : \mathbb{P}\, \mathit{Motivation} \to \mathbb{P}\, \mathit{Goal} \to \mathit{View} \to \mathbb{P}\, \mathit{Goal} \to \mathbb{Z} \\
\mathit{motiveffectdestroy} : \mathbb{P}\, \mathit{Motivation} \to \mathbb{P}\, \mathit{Goal} \to \mathit{View} \to \mathbb{P}\, \mathit{Goal} \to \mathbb{Z} \\
\mathit{satisfygenerate}, \mathit{satisfydestroy} : \mathbb{P}\, \mathit{Goal} \to \mathbb{Z} \\
\hline
\mathit{goals} \subseteq \mathit{goallibrary} \\
\forall\, \mathit{gs} : \mathbb{P}\, \mathit{goallibrary} \bullet \mathit{satisfygenerate}\ \mathit{gs} = \\
\quad\quad \mathit{motiveffectgenerate}\ \mathit{motivations}\ \mathit{goals}\ \mathit{actualpercepts}\ \mathit{gs} \land \\
\quad\quad\quad\quad\quad\quad \mathit{satisfydestroy}\ \mathit{gs} = \\
\quad\quad \mathit{motiveffectdestroy}\ \mathit{motivations}\ \mathit{goals}\ \mathit{actualpercepts}\ \mathit{gs} \\
\hline
\end{array}
$$

The *GenerateGoals* operation schema formally describes the generation of a new set of goals, which changes the state of the agent. The remaining part of the schema states that there is a set of goals in the goal library that has a greater motivational effect than any other set of goals, and the current goals of the agent are updated to include the new goals.

$$
\begin{array}{|l}
\hline
\mathit{GenerateGoals} \\
\Delta \mathit{AutonomousAgentState} \\
\mathit{AssessGoals} \\
\hline
\exists\, gs : \mathbb{P}\, \mathit{Goal} \mid gs \subseteq \mathit{goallibrary} \bullet \\
\qquad (\forall\, os : \mathbb{P}\, \mathit{Goal} \mid os \in (\mathbb{P}\, \mathit{goallibrary}) \bullet \\
\qquad\qquad (\mathit{satisfygenerate}\ gs \geq \mathit{satisfygenerate}\ os) \wedge \\
\qquad\qquad \mathit{goals}' = \mathit{goals} \cup gs) \\
\hline
\end{array}
$$

Once generated by an autonomous agent, goals persist until, for whatever reason, they are explicitly destroyed by that autonomous agent. The destruction of goals is defined analogously to the generation of goals, but we do not specify it here.

This view of goal generation and adoption can be understood as follows. I might use a bottle to store water for a trip I am taking. Here, the bottle is instantiated as an agent by me, an autonomous agent, satisfying my goal. However, I might also share the water, making the bottle of water available as a community resource, and allowing others to drink from it. The bottle thus becomes an agent for all of us, satisfying all of our goals. Alternatively, I could instead hand the bottle to someone else, giving up my ownership of it. In this situation, I release the bottle from its agency so that it can be an agent for the other person, who re-instantiates it as such. Interestingly, however, I cannot force the other person, an autonomous agent, to do so, since my interaction with them is uncertain. I can suggest that they take the bottle but they may not do so. The point here is that when there is interaction with other autonomous agents, we cannot know what will result, so we require a process that reflects this uncertainty: we require a process of *autonomous interaction*.

It is worth noting here that this applies as much to assessing and generating social goals, using models of other agents, as to individual goals. For example, if one agent has a model of another as being particularly competent at performing a certain action this may lead to the generation of the goal to persuade them to perform that action on behalf of the first agent. Similarly, goals for collaboration may be generated as a result of information from models of others. For example, if my model of you suggests that you are an excellent pianist, and I have the goal to sing, then I may generate the goal to persuade you to play the piano so that I can sing. In all this can be seen aspects of the creativity process. The better we are at understanding the actions, goals and motivations of others, the greater the potential to engage, collaborate, empathise, provoke and challenge others in all kinds of creative partnerships. Modelling others well is a great skill; the more accurately we can do so, the better we become at generating and assessing goals that involve us working together.

5. Autonomous Interaction

Once the goals defining the purpose of the interaction are generated, in the manner described above, an agent can continue in its attempt to achieve those goals. Sometimes these may be specific goals such as wanting to know the time, or asking a favour, but they can also be more nebulous and include simply being friendly at a party, or just chatting in a pub. However, in all of these cases, the Stanislavskian theory of interaction applies: that behind every utterance (verbal or otherwise) there is a goal of trying to change the mental state of the listener, and that the utterance serves as a strategy for trying to achieve this (Merlin, 2006). This notion also underlies our model, and at its heart is the idea that agents can only know whether they have achieved that goal by considering the evidence that is available to them after an interaction episode.

As we have discussed in the introduction, many traditional models of interaction have assumed an ideal world in which unfounded assumptions have given rise to inadequate characterisations of interaction amongst autonomous agents. If we consider autonomous interaction to be a goal driven process of uncertain outcome (which it must be), then we can characterise it in a more general way as a process of discovery in terms of the effects of actions (d’Inverno and Luck, 1996). This allows us to deal effectively with the inherent uncertainty in interaction. In the following discussion, we will begin to introduce the language of discovery to make the relationships clear.

In order to make sense of our environment and to function effectively in it, we continually anticipate the effects of our actions and utterances: we make predictions (or expectations) about what will happen next. The action-selection function, *autoactions*, of the *AutonomousAgentAct* schema encompasses the deliberation of the agent in this respect.

The action that is selected is intended to satisfy the goals of the agent through its resulting effects and consequent changes to the environment. In the case of an interaction episode involving two agents, the initiating agent selects an action that is intended to cause the desired response in the responding agent. The uncertainty inherent in such interaction means that the effects cannot be known in advance, but can only be discovered after the event has taken place, or action performed. We describe this by specifying the *predicted* effects of actions selected in the *AutonomousAgentAct* schema by applying the *socialeffectinteract* function to the current view of the environment and those actions. The agent thus predicts that these actions will change the environment to achieve the desired results. Remember that the environment includes all of the entities in it, so that a change to an agent in the environment will in turn cause a change to the environment itself. We also introduce a variable, *oldpercepts*, to store an agent's actual percepts prior to an operation.

$$\begin{array}{|l}
\multicolumn{1}{l}{\mathit{SocialAgentPredict}} \\
\hline
\mathit{SocialAgent} \\
\mathit{socialeffectinteract} : \mathit{View} \rightarrow \mathbb{P}\,\mathit{Action} \nrightarrow \mathit{View} \\
\mathit{oldpercepts}, \mathit{prediction} : \mathit{View} \\
\hline
\mathit{prediction} = \mathit{socialeffectinteract}\ \mathit{actualpercepts}\ \mathit{willdo} \\
\mathit{prediction} \cap (\bigcup \mathit{goals}) \neq \{\} \\
\hline
\end{array}$$

Such predictions encapsulate the notions of theory formation in the context of scientific discovery, in which creativity has a strong role to play. It is exactly here that there is scope for those aspects that have been considered by many to be outside the scientific process, for it is here that theories are created in order that they can subsequently be tested and revised as appropriate. The interaction process is no different. An agent's understanding of its environment (and other agents within it) gives rise to expectations of behaviour in others, leading to a determination of what action to take to elicit a desired response. Note that the model says nothing of how prediction is actually achieved other than that it is dependent on the environment, perceptual abilities, goals, motivations and models of others as discussed above; it may indeed be done through some complex process involving inspiration or creativity, or it may simply be a more prosaic operation.[3] Either way, this is where it would be.

Now, in order to achieve the desired result, the relevant actions must be performed. Effectively, this acts as an experiment, testing whether the predictions generated are consistent with the resulting effects. In this sense, experimentation is central to this model, for such interaction with the environment is the only way in which an agent's understanding of its capabilities and its environment can be assessed to bring to light inadequacies, inconsistencies and errors. When an action is performed (or utterance spoken), the evidence of the impact of that action is considered and evaluated by others. In particular, the action affects the models of other agents, and these are modified to provide new models (derived from the previous ones) through the *updateagentmodels* function. These models of agents are critical in determining if the action was successful.

[3] One possible technique that might be adopted in this context is reinforcement learning, but this is outside the scope of this paper, so we do not consider it further.

$$\begin{array}{|l}
\multicolumn{1}{l}{\textit{SocialAgentInteract}} \\ \hline
\textit{SocialAgentPredict} \\
\textit{updateagentmodels} : \mathbb{P}\,\textit{AutoAgentModel} \rightarrow \textit{View} \rightarrow \mathbb{P}\,\textit{AutoAgentModel} \\ \hline
\end{array}$$

The action also has an effect on the environment, which changes accordingly, and a similar effect on the acting agent itself, whose percepts also change. For example, in the case of an action that issues a request to another agent to tell the current time, the resulting model will either encode the fact that the agent is telling the time, or not. By inspecting this model and its attributes, the requesting agent can determine if its action has been successful. Note that the new value of *oldpercepts* takes the previous value of *actualpercepts* for later use.

$$\begin{array}{|l}
\multicolumn{1}{l}{\textit{SocialEnv}} \\ \hline
\Delta\textit{SocialAgentPredict} \\
\textit{SocialAgentInteract} \\ \hline
\textit{env}' = \textit{socialeffectinteract env}' \textit{ perceivingactions} \\
\textit{posspercepts}' = \textit{canperceive env}' \textit{ perceivingactions} \\
\textit{actualpercepts}' = \textit{willperceive motivations goals posspercepts}' \\
\textit{willdo}' = \textit{autoactions motivations goals actualpercepts}' \textit{ env}' \\
\textit{modelautoagents}' = \\
\qquad \textit{updateagentmodels modelautoagents actualpercepts} \\
\textit{oldpercepts}' = \textit{actualpercepts} \\
\textit{prediction}' = \textit{prediction} \\ \hline
\end{array}$$

Evaluating the results of actions appears simple. At the most basic level, it involves the comparison of predictions with observations. Thus, if the intended effects of one's actions and the actual effects match, then the actions have achieved the desired result and the episode is successful. If they are anomalous, then it reveals an erroneous understanding of the environment and the agents within it, or an inadequate capability for perception of the results. The important point here is that there is no guarantee of success, and failure can be due to any number of reasons.

This analysis assumes that the evidence is perfect, however, which may not always be appropriate. In any real environment this is not so, and error can be introduced into evidence in a variety of ways, reducing the quality of the observed evidence accordingly. Not only may there be inaccuracy due to the inherent uncertainty in both performing actions and perception of the results (experimentation and observation respectively), but also, if the actions taken by the agent are communicative actions intended to elicit a response from another autonomous agent, then there may be inaccuracy due to malicious intent on the part of the responding agent by providing misleading information, for example (Marsh, 1994). Thus the response may itself be the vessel for the error. For example, when requesting someone to tell the time in a noisy room, one may not be certain of whether the request was heard or even understood. Evaluating the evidence allows an appropriate response, which may or may not included modifying models of others.

In addition to assessing the fit of observations with predictions, therefore, the quality of the observations themselves must also be assessed in order to ascertain whether they are acceptable to be used in the comparison at all. Simple tolerance levels for assessing the acceptability of perceived evidence are inadequate, for they do not consider the need for the interaction episode, and the importance of achieving the desired result. The quality demanded of the observations can thus only be assessed in relation to the motivations of the agent. These provide a measure of the importance of the situation, and take into account the implications of success and failure. In medical domains, for example, where agents are highly motivated, even a small degree of error in interaction concerning relevant patient details may be unacceptable if it would lead to the loss of a patient's life, while neighbourly discussion of the weather with low motivations and little importance may allow a far greater error tolerance.

The schemas below describe evaluation with two functions. First, *accept* is a predicate that holds between the capabilities of the agent, its perceived environment before and after the actions were performed, and its agent models, when there is sufficient evidence to accept it. The capabilities of the agent capture the uncertainty information that arises from the agent itself, while the perceived environment and agent models include details of difficulties arising through the environment, or other agents. The predicate *consider* compares predictions and observations once evidence is accepted. The predicate holds when the predictions and observations are considered to be consistent. Note that the potentially difficult question of when observations match predictions is bound up in the predicate itself, which may be interpreted either as a simple equality test or as something more sophisticated.

The second schema also states at the beginning that though the agent changes (Δ*SocialAgent*) as a result of this evaluation, the state of the agent remains the same (Ξ*AutonomousAgentState*). Finally, if the evidence is accepted, and the observations are not consistent with the predictions, then the agent models must be revised in an appropriate way as specified by the *revisemodels* function.

Note that creativity can manifest itself strongly in both evaluation and revision. First, evaluation is intimately tied to perception as indicated above, where perception can be regarded as having a very strong bearing on creativity since this perception may vary from individual to individual. Second, revision of models can be achieved in infinitely many ways. Determining which revision to apply is also determined by perception, and can give rise to different outcomes, depending on the individual involved. Importantly, both of these aspects determine how these models can change.

$$\begin{array}{|l}
\hline
\mathit{SocialAgentEvaluate} \\
\hline
\mathit{SocialAgent} \\
\mathit{accept}\_ : \mathbb{P}(\mathbb{P}\,\mathit{Action} \times \mathit{View} \times \mathit{View} \times \mathbb{P}\,\mathit{AutonomousAgent}) \\
\mathit{consider}\_ : \mathbb{P}(\mathit{View} \times \mathit{View}) \\
\hline
\end{array}$$

$$\begin{array}{|l}
\hline
\mathit{SocialAgentDecides} \\
\hline
\Delta \mathit{SocialAgent} \\
\Xi \mathit{AutonomousAgentState} \\
\mathit{SocialAgentPredict} \\
\mathit{SocialAgentEvaluate} \\
\mathit{revisemodels} : \mathit{View} \rightarrow \mathbb{P}\,\mathit{AutoAgentModel} \rightarrow \mathbb{P}\,\mathit{AutoAgentModel} \\
\hline
\mathit{accept}(\mathit{capabilities}, \mathit{actualpercepts}, \mathit{oldpercepts}, \mathit{modelautoagents}) \land \\
\mathit{consider}(\mathit{prediction}, \mathit{actualpercepts}) \Rightarrow \\
\qquad \mathit{modelautoagents}' = \mathit{revisemodels}\ \mathit{actualpercepts}\ \mathit{modelautoagents} \\
\hline
\end{array}$$

6. Discussion

6.1. Creativity through Software

The emerging field of computational creativity is no longer a nascent area, but is now becoming increasingly well established (e.g., (Colton et al., 2009; Cardoso et al., 2009)). Colton (Colton, 2012) describes the field as being concerned with building software that is independently creative so as either to act in partnership with people or to be an "autonomous artist, musician, writer, designer, engineer or scientist." Interestingly, he suggests that software can act as more than a simple tool for creative activity, a view also held by many others in the computational creativity field (Cohen et al., 2012; Jones et al., 2012; McCormack and d'Inverno, 2012a), where creativity can be inspired by interacting with computational systems as autonomous agents(Blackwell et al., 2012; McCormack and d'Inverno, 2012b).

Yet simple tools are also relevant for creativity; Clark (Clark, 2008), for example, discusses how the application of pen on paper in writing means that neither can be considered as passive artefact but as fundamental machinery responsible for “the shape of the flow of thoughts and ideas.” Though not directly referring to such objects as *adopting the goals* of a human as part of the creative process, the ideas here are clearly analogous to goal adoption by objects as we have described in this paper. With software, which can be responsive in ways not possible with passive artefacts, new possibilities arise for interacting in the creative process.

Indeed, Jones et al. (Jones et al., 2012) argue that partnerships with autonomous agents, for which the outcome of an interaction cannot be known, is different from the use of traditional tools in two important aspects. First, these can extend our practice through new capabilities or trajectories that would not otherwise have been possible. Second, by mirroring our own creative behaviours, they enable us to reflect on our behaviour in order to disrupt existing habits. By moving one step further, to autonomous software, we arrive at a new kind of partnership in which the tool itself can be viewed as creative. This suggests a new kind of relationship with (at least partially) equal partners rather than the more typical relationship involving unidirectional usage of a tool.

Consider the the work of Cohen and his system for drawing and painting, AARON (McCorduck, 1991), which represents one of the biggest successes of creative partnerships with a computational tool, both in terms of critical appraisal in the art world and some acceptance by collectors and galleries. Cohen does not consider AARON as autonomous, but discusses the way that it has been developed to paint in his style, as well as how his style has been affected by feedback from the software (Cohen et al., 2012). Such a creative partnership can arise when the interaction is much closer to autonomous interaction as described in this paper and moves beyond what is possible with non-computational systems. (Indeed, when discussing AARON, Cohen often ascribes the system human characteristics.)

AARON is not unique. Colton is developing the Painting Fool, a system that he hopes will one day produce visual art that can be displayed alongside human art, and given equal value (Colton, 2012). As work progresses, he increases the artistic autonomy of the system, acting as the evaluator of outputs to improve the system, and often expressing surprise and delight at the system producing novelty in his eyes. More challenging for Colton, however, is to get an external audience to see its creative value. Similarly, Pachet has spent a decade or more developing autonomous and semi-autonomous systems that capture the musical performance of a user and then play a modified version back using a range of learning and statistical techniques (Pachet, 2002; Pachet, 2010; Pachet, 2012). Importantly, these are both instances of autonomous interaction as described here, since it is not known which aspects of the music (in Pachet’s case) will be developed and in what way, nor how it will interact with a human over the course of any performance.

One consequence of this is that it can lead to a heightened desire on the part of the user to listen more carefully, and open awareness to the system in order to build up meaningful musical conversations, in ways that might not be done otherwise, even (or especially) with human interactions. The unexpected results of such systems may thus be used to push and develop creativity in people. Of course, much of this depends on how a human user chooses to model any computational system. By choosing a motivated agent model with its own musical goals and perceptions, and treating interaction as a process of discovery as described earlier, a creative partnership is much more likely than if treating it merely as a passive artefact.

There are new opportunities for creativity to be challenged, provoked, augmented or extended by more autonomous interaction with increasingly autonomous computational systems. These systems can provide the critical element surprise in interaction and even "super-surprise" as suggested by the pioneering artist Frieder Nake (Cohen et al., 2012). Moreover, they provide the potential for beginners as well as practicing artists to engage in creative practice.

6.2. Creativity through Autonomy

In our view, and as we have argued in this paper and elsewhere, the autonomy of agents arises through their being guided by internal motivations in determining their perception and action. This gives rise to different behaviour from different individuals, even in the same circumstances, which is seemingly a fundamental characteristic of creativity. Indeed, if everyone behaved in the same way, then consideration of creativity becomes meaningless, since it is indistinguishable from anything else. In this respect, autonomy can be seen to be a driver for creativity; it may not specify exactly what arises as a result, but in our view it is a necessary pre-requisite for creativity to occur.

Importantly, there are two aspects here. Behaviour, as discussed throughout this paper, is the manifestation of creativity, since it is only through the actions of agents that creativity is realised. However, the SMART framework also reveals a second, no less critical aspect in that autonomy affects not just what agents do, but also what they perceive. Since an agent's perceptions (as well as its behaviour) are determined in relation its motivations, we open up the possibility for different individuals to be exposed to different perceptual inputs, leading to different outputs as a result. For example, the vivid colours and style of impressionist paintings suggest a different way of seeing the world that one might argue is determined by the observers' motivations. Without dwelling on the detail, we believe that our simple model captures this notion of motivated perception in a very abstract way, but one that shows the capacity of perception to influence behaviour of all kinds, potentially leading to the kind of outputs that we would happily deem creative. Clearly, the detail is critical to a full understanding of the creative process, but the key point is that it is captured and can realised (or instantiated) in appropriate ways.

6.3. Creativity through Discovery

The history of science is littered with examples of creativity in devising new theories, models and experiments. While science is commonly taken to be a methodical and systematic process, there is still room for inspiration and intuition; indeed, some argue that this is perhaps the single key ingredient for fundamentally new scientific discoveries. Despite not providing a computational model for this creative aspect, we recognise its role in the process. Importantly, we see discovery as a process of interaction with the environment, carrying out experiments to test some theory, and continuing to provide new theories in response to the results. But this is not confined to science; it is what happens every day in all manner of situations when dealing with the uncertain environment that is the real world. We continually make predictions about our environment and adjust our behaviour in light of events. In this view, it makes sense to consider autonomous interaction as discovery as we have done in this paper, bringing to bear the same processes, but perhaps for ends that are important only for the individual rather than society as a whole.

More specifically in relation to interaction as discovery, we have necessarily presented a simple model; there are many other aspects for which we could have given far more detail. For example, we have not considered how to revise an agent's models in case its actions do not bring about the desired result, yet this is in itself an important but complex process, and one in which creativity must be present in some form. Indeed, much research into the problems of discovery comes under the heading of theory revision, or in the case of interaction as discussed here, model revision. Given an existing theory, new evidence will either be consistent with that theory or it will be anomalous. If it is consistent, then there is no cause for further reasoning since the theory is adequate. If the observations are anomalous, however, then the theory is refuted and must either be discarded or revised so that the anomaly is removed and the theory is once more consistent with observations. In the revision stage can be seen that part of the discovery cycle that is actually responsible for the construction of new theories, for the essence of the creativity involved. Yet perhaps we can argue that this creativity arises as a result of interaction, whether it be with the environment as is typical in science, or with other individuals (human or agent) in other areas.

6.4. Conclusions

Many critics of artificial intelligence argue that the real power of the technology lies in the hands of the programmers who manipulate and tweak their designs to suit circumstances. They argue that extensive programmer (or user) intervention invalidates much of the benefit to be gained from this work by demanding modifications for each novel situation. One of the ways in which this can be countered is by making intelligent systems autonomous, by giving them the power to control themselves. A significant contribution of this paper has been to show that autonomy can be achieved through the modelling and use of the motivations of a reasoning agent. In contrast to other work on motivations, this paper aims to show how motivations affect reasoning and action in the world, rather than how action and reasoning in the world affect motivations. Moreover, the reasoning and action addressed are not the simple behavioural-response kind, but of higher level reasoning strategies.

Of course, the model proposed in this paper is clearly at a high-level, and provides only a basic indication of how motivations may themselves be organised and related, and how they give rise to different kinds of behaviours. However, it provides well-defined framework within which a richer analysis of the relationships between motivations, by which some motivations may inhibit others, or some motivations may generalise others, could provide much stronger insights into impact on creativity. The model we have presented opens up all kinds of potential avenue for further research.

In summary, in this paper we have sought to bring together several strands of work, on motivation (Luck and d'Inverno, 1998), autonomous agents (d'Inverno and Luck, 2004; Luck and d'Inverno, 1996) and interaction (d'Inverno and Luck, 1996) on the one hand, and on discovery on the other. In doing so, we have sought to show how the notions of creativity that Boden has so eloquently espoused (Boden, 2004; Boden, 2010) have a central place in relation to the design of modern computing systems of intelligent agents. However, it is only in the detail of analysis that these aspects are visible. Just as it is sometimes easy to dismiss the creativity of the street artist, it is also easy to dismiss the creativity in everyday interaction. To do so would be a mistake.

References

Allen, J.: 1979, 'A plan-based approach to speech act recognition'. Technical Report 131, Department of Computer Science, University of Toronto.

Appelt, D. E.: 1985, *Planning English Sentences*. Cambridge University Press.

Austin, J. L.: 1962, *How to do Things with Words*. Oxford University Press.

Blackwell, T., O. Bown, and M. Young: 2012, 'Live Algorithms: towards autonomous computer improvisers'. In: J. McCormack and M. d'Inverno (eds.): *Computers and Creativity*. Berlin: Springer.

Boden, M.: 2004, *The Creative Mind: Myths and Mechanisms*. London: Routledge.

Boden, M.: 2010, *Creativity and Art. Three Roads to Success*. Oxford University Press.

Bowen, J. P.: 1996, *Formal Specification and Documentation using Z: A Case Study Approach*. International Thomson Computer Press.

Campbell, J. A. and M. d'Inverno: 1990, 'Knowledge interchange protocols'. In: Y. Demazeau and J.-P. Mueller (eds.): *Decentralized AI: Proceedings of the First European Workshop on Modelling Autonomous Agents in a Multi-Agent World*. pp. 63–80, Elsevier.

Cardoso, A., T. Veale, and G. A. Wiggins: 2009, 'Converging on the Divergent: The History (and Future) of the International Joint Workshops in Computational Creativity'. *AI Magazine* **30**(3), 15–22.

Castelfranchi, C.: 1990, 'Social power'. In: Y. Demazeau and J.-P. Mueller (eds.): *Decentralized AI: Proceedings of the First European Workshop on Modelling Autonomous Agents in a Multi-Agent World*. pp. 49–62, Elsevier.

Chopra, A. K., A. Artikis, J. Bentahar, M. Colombetti, F. Dignum, N. Fornara, A. J. I. Jones, M. P. Singh, and P. Yolum: to appear, 2011, 'Research Directions in Agent Communication'. *ACM Transactions on Intelligent Systems and Technology*.

Clark, A.: 2008, *Supersizing the Mind*. New York, NY: Oxford University Press.

Cohen, H., F. Nake, D. C. Brown, P. Brown, P. Galanter, J. McCormack, and M. d'Inverno: 2012, 'Evaluation of Creative Aesthetics'. In: J. McCormack and M. d'Inverno (eds.): *Computers and Creativity*. Berlin: Springer.

Colton, S.: 2012, 'The Painting Fool: Stories from Building an Automated Painter'. In: J. McCormack and M. d'Inverno (eds.): *Computers and Creativity*. Berlin: Springer.

Colton, S., R. L. de Mántaras, and O. Stock: 2009, 'Computational Creativity: Coming of Age'. *AI Magazine* **30**(3), 11–14.

Crick, F.: 1988, *What Mad Pursuit: A Personal View of Scientific Discovery*. New York, NY: Basic Books.

d'Inverno, M. and M. Luck: 1996, 'Understanding Autonomous Interaction'. In: W. Wahlster (ed.): *ECAI '96: Proceedings of the 12th European Conference on Artificial Intelligence*. pp. 529–533.

d'Inverno, M. and M. Luck: 1997, 'Development and Application of a Formal Agent Framework'. In: M. G. Hinchey and L. Shaoying (eds.): *ICFEM'97: Proceedings of the First IEEE International Conference on Formal Engineering Methods*. pp. 222–231, IEEE Computer Society.

d'Inverno, M. and M. Luck: 2004, *Understanding Agent Systems*. Springer, second edition.

d'Inverno, M., M. Luck, P. Noriega, J. A. Rodríguez-Aguilar, and C. Sierra: 2011, 'Weaving a Fabric of Socially Aware Agents'. In: D. Kinny, J. Y. jen Hsu, G. Governatori, and A. K. Ghose (eds.): *Agents in Principle, Agents in Practice, 14th International Conference, PRIMA 2011*, Vol. 7047 of *Lecture Notes in Computer Science*. pp. 263–274, Springer.

Dorin, A. and K. B. Korb: 2012, 'Creativity Refined: Bypassing the Gatekeepers of Appropriateness andValue'. In: J. McCormack and M. d'Inverno (eds.): *Computers and Creativity*. Berlin: Springer.

Galliers, J. R.: 1990, 'The positive role of conflicts in cooperative multi-agent systems'. In: Y. Demazeau and J.-P. Mueller (eds.): *Decentralized AI: Proceedings of the First European Workshop on Modelling Autonomous Agents in a Multi-Agent World*. Elsevier.

Halliday, T.: 1983, 'Motivation'. In: T. R. Halliday and P. J. B. Slater (eds.): *Causes and Effects*. Blackwell Scientific.

Hayes, I. J. (ed.): 1993, *Specification Case Studies*. Hemel Hempstead: Prentice Hall, second edition.

Jones, D., A. R. Brown, and M. d'Inverno: 2012, 'The Extended Composer'. In: J. McCormack and M. d'Inverno (eds.): *Computers and Creativity*. Berlin: Springer.

Kasmarik, K., W. T. B. Uther, and M. L. Maher: 2005, 'Motivated Agents'. In: L. P. Kaelbling and A. Saffiotti (eds.): *IJCAI*. pp. 1505–1506, Professional Book Center.

Kunda, Z.: 1990, 'The case for motivated reasoning'. *Psychological Bulletin* **108**(3), 480–498.

Leake, D. and A. Ram: 1993, 'Goal-drive learning: Fundamental issues and symposium report'. Technical Report 85, Cognitive Science Program, Indiana University,, Bloomington, Indiana.

López y López, F., M. Luck, and M. d'Inverno: 2006, 'A normative framework for agent-based systems'. *Computational & Mathematical Organization Theory* **12**, 227–250.

Luck, M. and M. d'Inverno: 1995, 'A Formal Framework for Agency and Autonomy'. In: *Proceedings of the First International Conference on Multi-Agent Systems*. pp. 254–260, AAAI Press / MIT Press.

Luck, M. and M. d'Inverno: 1996, 'Engagement and Cooperation in Motivated Agent Modelling'. In: C. Zhang and D. Lukose (eds.): *Distributed Artificial Intelligence Architecture and Modelling: Proceedings of the First Australian Workshop on Distributed Artificial Intelligence, Lecture Notes in Artificial Intelligence*, Vol. 1087. pp. 70–84, Springer.

Luck, M. and M. d'Inverno: 1998, 'Motivated Behaviour for Goal Adoption'. In: C. Zhang and D. Lukose (eds.): *Multi-Agent Systems: Theories, Languages and Applications — Proceedings of the Fourth Australian Workshop on Distributed Artificial Intelligence*, Vol. 1544 of *Lecture Notes in Artificial Intelligence*. pp. 58–73, Springer.

Macedo, L. and Cardoso: 2002, 'Assessing creativity: the importance of unexpected novelty'. In: *Proceedings of the ECAI Workshop on Creative Systems*.

Maes, P.: 1991, 'A Bottom-Up Mechanism for Behaviour Selection in an Artificial Creature'. In: J. A. Meyer and S. Wilson (eds.): *Proceedings of the First International Conference on Simulation of Adaptive Behaviour: From Animals to Animats*. pp. 238–246, MIT Press/Bradford Books.

Marsh, S.: 1994, 'Trust in distributed artificial intelligence'. In: C. Castelfranchi and E. Werner (eds.): *Artificial Social Systems: Fourth European Workshop on Modelling Autonomous Agents in a Multi-Agent World*, Vol. 830 of *Lecture Notes in Artificial Intelligence*. pp. 94–114, Springer.

McCorduck, P.: 1991, *AARON's Code: Meta-Art, Artificial Intelligence, and the Work of Harold Cohen*. W.H. Freeman and Company.

McCormack, J. and M. d'Inverno: 2012a, 'Computers and Creativity: the Road Ahead'. In: J. McCormack and M. d'Inverno (eds.): *Computers and Creativity*. Berlin: Springer.

McCormack, J. and M. d'Inverno: 2012b, 'Why does Computing matter to Creativity?'. In: J. McCormack and M. d'Inverno (eds.): *Computers and Creativity*. Berlin: Springer.

Merlin, B.: 2006, *The Complete Stanislavsky Toolkit*. London: Nick Hern.

Pachet, F.: 2002, 'Playing with Virtual Musicians: the Continuator in Practice'. *IEEE Multimedia* **9**(3), 77–82.

Pachet, F.: 2010, *The Continuator Strikes Back: a Controllable Bebop Improvisation Generator*. The International Conference on Computational Creativity, Lisbon, Portugal: University of Coimbra, pp.292.

Pachet, F.: 2012, 'Musical Virtuosity and Creativity'. In: J. McCormack and M. d'Inverno (eds.): *Computers and Creativity*. Berlin: Springer.

Perrault, C. R.: 1990, 'An application of default logic to speech act theory'. In: P. R. Cohen, J. Morgan, and M. E. Pollack (eds.): *Intentions in Communication*. MIT Press, pp. 161–186.

Ram, A. and D. Leake: 1991, 'Evaluation of explanatory hypotheses'. In: *Proceedings of the Thirteenth Annual Conference of the Cognitive Science Society*. pp. 867–871.

Schmidhuber, J.: 1991, 'A Possibility for Implementing Curiosity and Boredom in Model-Building Neural Controllers'. In: J. A. Meyer and S. Wilson (eds.): *Proceedings of the First International Conference on Simulation of Adaptive Behaviour: From Animals to Animats*. pp. 465–474, MIT Press/Bradford Books.

Searle, J. R.: 1969, *Speech Acts*. Cambridge University Press.

Simon, H. A.: 1979, 'Motivational and emotional controls of cognition'. In: *Models of Thought*. Yale University Press, pp. 29–38.

Sloman, A.: 1987, 'Motives, mechanisms, and emotions'. *Cognition and Emotion* **1**(3), 217–233.

Sloman, A. and M. Croucher: 1981, 'Why robots will have emotions'. In: *Proceedings of the Seventh International Joint Conference on Artificial Intelligence*. pp. 197–202.

Smith, R. G.: 1980, 'The contract net protocol: High-level communication and control in a distributed problem solver'. *IEEE Transactions on Computers* **29**(12), 1104–1113.

Spivey, J. M.: 1988, *Understanding Z: A Specification Language and its Formal Semantics*. Cambridge University Press.

Spivey, J. M.: 1992, *The Z Notation: A Reference Manual*. Hemel Hempstead: Prentice Hall, 2nd edition.

Steels, L.: 1995, 'When are robots intelligent autonomous agents?'. *Journal of Robotics and Autonomous Systems* **15**(3), 3–9.

Weisberg, R. W.: 2006, *Creativity: Understanding Innovation in Problem Solving, Science, Invention and the Arts*. Wiley.

Woodcock, J. and J. Davies: 1996, *Using Z: Specificiation, Refinement and Proof*. Hemel Hempstead: Prentice Hall.