Three Paradigms of Rational Agency

<Pre-publication version. Please Cite published version: R. Barnard and T. Simon, "Three Paradigms of Rational Agency," Journal of Models and Modeling, 1:2 (April 2003): 31-54.>

Robert Barnard and Tyler Simon
University of Mississippi

<rwbjr@olemiss.edu>
<tasimon@olemiss.edu>

I. Introduction

The development of a model is always at the service of a set of purposes or goals. These goals either help to guide us or compel us in making choices regarding the level of detail or complexity to be represented within the model. Economic models, for example, regularly ignore the particularities of actual individuals or governments or corporations in favor of representing the variety of economic agents in terms of sets of more-or-less easily represented behaviors governed by an idealized form of rationality. Weather forecasting models often deal with fairly large ‘packets’ of air uniformly assigning them characteristics such as temperature and humidity even though in actuality these conditions might vary widely. It follows rather straightforwardly from this observation that in order to make intelligent choices about what details to retain and what details to omit in a model we ought to have a more complete understanding of what we are trying to model than is given by or in the model alone.

In formal contexts, this intuition has been made explicit in the past by those who, following Tarski (1944), recognize that if we introduce a concept into a system of formal representation, such as a model, the concept as introduced has two complementary obligations, formal correctness, i.e. that it not introduce inconsistencies or other formal defect into the system, and material adequacy, i.e. that it actually capture the content of something close to the concept itself. It is the material adequacy issue which concerns us here. If there is a reason for including the ability to represent a concept such as agency or rationality within a formal system then that reason derives from the pre-theoretic significance of the notion, not from the notion as realized within the model.

In the case of agents and agency this means that we must look at what it is to be an agent or be *agentive*. How shall we answer the fundamental question of what it is to be an agent? Can we give necessary and sufficient conditions? Must we appeal to paradigm cases? Or, do we presume to operate on the basis of a naïve principle such as: 'I can't define it but I know it when I see it'? This project is complicated by the fact that the more precise we make a proposed definition of agency the closer it gets to being a model of agency itself—however informal. It is for this reason that we choose to investigate the *conceptual geography* of 'agency' rather than agents themselves in Section II of this paper.

In Section III, we turn to the issue of rationality. Agents in general can be rather mundane. The really interesting cases of agents tend to be agents capable of doing interesting things. A number of conceptually related terms have been applied to these interesting agents: intelligent, autonomous, rational, social, etc. Our aim in this section is to investigate the nature of rationality in relation to agency and in a general way to alternative strategies for modeling rational agents.

Finally in section IV, we lay out what we take to be three paradigms for thinking about rational agency in general and make some general suggestions about the sorts of model appropriate to each.

II. Conceptions of Agency

Our goal in this section is simple; we want to unpack and clarify the concept of what it is to be an agent. Several researchers have reflected, in specific terms, upon the inadequacy of our understanding of the notion of *agency*, in its various forms. For example, Wooldridge and Jennings (1995) have previously noted the ambiguity of the notion of agency while distinguishing between strong and weak agents in a manner that mirrors the strong/weak AI distinction. Luck and d'Inverno (2001) have proposed an agent framework within which agents differ from mere objects in virtue of having goals. Similarly Norman and Long (1995) have also placed a great deal of importance upon the capacity of so-called BDI agents (agents capable of having beliefs, desires, and forming intentions) to develop and pursue goals. At a more metaphysical level John Sowa (1998) asserts that the mark of an agent is its possession of an artificial *psyche*. And at the level of behavior, Russell and Norvig (1995) have claimed that agents must at a minimum have the ability to perceive and act. But who is closest to the mark? Aren't they all just waxing poetic about something that is more or less operationally defined within a formal system? There is a sense in which all are certainly capturing part of our pre-theoretical notion of agency, but how much more is there? David Moffat and Nico Frijda suggest the scope of problem:

> The word *agent* is a technical term without a generally agreed usage. It can refer to something as trivial as a persistent computer program, like the Unix *at* daemon; or to an independent system with enough knowledge and interactive capabilities to ensure its effectiveness in a wide range of somewhat unforeseen circumstances. The biological model for this latter kind of agent would be an animal of some kind, somewhere on the scale from insect to human. (1995)

What we wish to claim is that the concept of agent is in some senses much broader than even this passage suggests.

1. Broad Agents and Narrow Agents

The most central notion embedded in the concept of agency is that of action. To be an agent at a minimal level is to be an actor or a doer. This connection is revealed by reflecting upon related concepts or terms: e.g. the role of a re*agent* in a chemical re*action*. Likewise, one employs an agent to act on one's behalf in a variety of matters. The connection between agency and action should come as no surprise, rather it helps to define what we shall term the *Broad Conception of Agency* (BCA). Anything that counts as an agent under this conception we shall call a *broad agent*. Examples of broad agents might plausibly include anything that acts. To make this somewhat more specific, let us say that ***x acts when x plays any role in the explanation of some change or evolution of***

a system. It follows that while human beings, or software "agents" are broad agents, so are things like glaciers, forest fires, chemicals, molecules, and weather systems. These things can all "do." They can all produce a change or evolution in some system, be it environmental, virtual, chemical, biochemical, or atmospheric.

The attribution of agency to a forest fire, a hurricane, or a glacier may sound a bit odd at first, but it is consistent with the sorts of things we sometimes say about them. We often attribute psychological or personal characteristics such as motive or desire to these sorts of things. But, while we may say that a forest fire desires to consume everything around it, we do not really mean that it has a motivating desire or any conception of the aim embedded in the desire. Rather, we are attributing a kind of as-if intentionality to the fire; it acts in the same way some agent that actually desired to consume a forest would act.

The notion of agency with which we tend to be more familiar we shall call the *Narrow Conception of Agency* (NCA) and we shall refer to such agents as *Narrow Agents*. Examples of narrow agents might include any sort of human like agent, some sort of device with an autonomous control system, most animals, and any other thing that is capable of action that is *actually* goal directed (where the notion of goal still remains quite broad).

We can now see that while all narrow agents are also broad agents, the converse is not the case. However, the real distinction between mere agency and narrow agency resides in the notion of an agent's capacity to be oriented to a goal. This is not to say that all actions performed by narrow agents must necessarily also be goal oriented. Clearly the capacity for narrow agency does not require that all actions be narrowly agentive. Rather the concept of narrow agency recognizes the in-principle capability of such agents to perform goal directed actions. This distinction between aimless and goal directed actions is not always easy to make. You will recall that even broad agents were characterized by their tendency to appear purposive or otherwise goal directed, or rather their actions were. This creates an obvious complication, since the stability of the distinction between broad and narrow agency requires that we have some means for distinguishing between actions that are actually purposive and those that only seem to be purposive.

There seem to be two approaches available to us. First we might construct the distinction between narrow and merely broad agents upon some sort of principle that we build into the agent or agents in virtue of our attempts at modeling. This is a pragmatic or definitional approach, for it forces us to make a decision and live with it. But, there is no reason to think that this attribution of purpose by fiat will carve the class of agents into those which are actually goal directed and those which are not. In point of fact it will probably draw the class of narrow agents too narrowly. This might not strike all of us as significant, but if our immediate goal is to better understand the concept of agency, then this is a clear setback. The second possible approach is to be extremely liberal about what we take to be goal directed action. We might allow that any action which seems to be purposive be counted as purposive. This approach has the opposite difficulty from the pragmatic approach; it will probably draw the class of narrow agents too broadly. These considerations would seem to suggest that we might be unable to actually frame the distinction here conceived. Nevertheless one wants to maintain that glaciers are not agents in the way that football players are agents, despite the fact that both are capable of uprooting a young tree. So, while it is possible to frame the distinction between narrow and broad agency, and perhaps give examples of each, we cannot give any reliable extensional definition of narrow agency.

One further strategy might ignore the agent in favor of evaluating the action and its effects. This approach seems to be validated by reflecting upon the fact that for some types of objects we understand them in terms of what they do rather than what they are. Clearly a full understanding of a knife requires that we consider how it cuts and a full understanding of a flute will require that we hear it played. By analogy, perhaps a proper understanding of an agent requires that we look at what the agent does. For our purposes we have defined an action as some event which results in some change or evolution of a system of which the agent is a part. We must now try to characterize just how the agent participates in the action.

2. Direct and Indirect Agentive Relations

Not all agents are related to actions as causes are related to effects. Indeed the actions of an agent in either the broad or narrow sense may fall short of being sufficient to produce a specifiable effect. Instead we must recognize that sometimes the explanation of an event or an action requires that we allow for the combined action of several unrelated agents (e.g. producing a change in the GDP of Bolivia), or perhaps the concerted action of several individual agents, each singularly incapable of producing the effect in question (e.g. winning the World Series). Call the relationship between agent and the effect of an agent's action the *agentive relation.* We say of an agentive relation that it is *direct* when the agent is the immediate and proximate explanation for an evolution or change in the system, what philosophers call the efficient cause. An example of direct agency might be a lumberjack felling a tree. Call the relationship between the agent and the resulting state of the system *indirect* when the action of the agent plays a role in the explanation of the change in the system, but is not a sufficient cause.

The class of indirect agency relations can be further subdivided in those which *facilitate* and those which *initiate*. Facilitating agents are required in order to make some change in a system, usually more directly due to the actions of other agents, possible. One clear example is a chemical catalyst. Here the chemical agent participates in the reaction but not as a proximate cause of the change in the system. Facilitating agents play a necessary but not sufficient role in explaining the evolution of the system. Initiating agents play neither role. Rather they are a contingent first link in a chain of events. Their status is further complicated by the fact that they may perform some role for which they are the direct and proximate cause but play an initiating role within a larger pattern of change within the system. The initiating agent is related in a contingent way to the final outcome of this larger change. For example consider the assassination that set off the escalation of tensions in the Balkans which ultimately resulted in the outbreak of the First World War. The assassin was directly agentive in the killing of the Austrian Arch-duke. But was the assassin a direct agent relative to the war? No. There were points at which the chain of events could have been broken. And there were other possible events, any of which might have set the same chain of event in action.

A further sort of agentive relation remains to be discussed: *collective agency*. Consider the different sorts of activities that we attribute to individuals as opposed to collective. Individual ants might forage for food, care for their young, move individual bits of soil, etc. But, an ant colony is responsible for the creation of an anthill, a complex structure that requires complex patterns of division of labor to create and maintain. Likewise, we might speak of individual workers at a construction sight. No one worker builds the skyscraper, but the construction workers, or the construction company does. We speak of communities, societies, nations, organizations, and networks as both composed of agents and as agents themselves. However, our ability to fully describe collective agency will often depend upon how clear our picture of the composing agents can be made. Alternatively, certain circumstances might warrant our ignoring agency below the collective level. In either case the task of understanding agency itself remains central.

The recently discussed approach of trying to distinguish types of agents based upon the sort of agentive relation they participate in has the advantage of being external to the agents themselves. We do not face the epistemological problem of distinguishing genuinely purposive from seemingly purposive acts. However, it is unclear if this alternative typology allows us to make any progress either. In the face of this uncertainty, we might ask if there is a way to distinguish between direct and indirect agency? In order to make such a distinction and apply it to particular cases, we need to be able to answer the following question: what action context or effect context are we relating to the agent? In the case of the assassin we saw that instances of direct agency could be embedded within cases of indirect agency. So within the totality of the complex set of events all occurring when a system changes or evolves, the same agent performing the same individual actions can be plausibly seen as directly agentive relative to one description of a larger action context (the assassination) and indirectly agentive relative to another description of the action context (World War I). In the former instance however there was a clear sense in which the actions of the assassin were directed at a goal. In the latter instance, however, no one would suggest that the assassin's explicit aim was to set off precisely that complex set of events that followed the assassination. What this tells us is that our ability to ascribe direct and indirect agentive relationships is to some degree at the service of whether the behavior of an agent is or is not directed at a goal. This is because action contexts are sometimes distinguished on the basis of whether the agent or agents involved are pursuing a goal, and further by asking what goal they are pursuing. This means that while we can frame a distinction between direct and indirect agentive relations, and give examples of each, we remain unable to provide a clear principle upon which to isolate and distinguish the types. Again, the obscure notion of goal directedness is central.

3. Goals and Rationality

Having recognized the centrality of an agent's goal orientation to the distinctions between broad and narrow agents and direct and indirect agency, we can now make an in-principle distinction between agents that have the ability to pursue goals and those that do not. In order to genuinely pursue a goal an agent needs to have the capacity to represent the goal in question. This capacity or the lack of such a capacity gives us the ability to determine which agents are candidates for narrow agency. Likewise, only agents with the capacity for goal directed action can be related to their actions directly. The capacity for goal directed action in an agent must become our focus, not because we began by presuming that all agents are goal directed, but rather because the capability or failure of an agent to be goal directed is central to a proper understanding of the notion of agency itself. To this extent, our analysis provides an extra-theoretic confirmation and grounding to those theories which emphasize the significance of goals to agents of a less than mundane sort.

Before moving on, there is one sense of 'goal directed' in which we are not immediately interested. Specifically the sense in which biological organisms, understood teleologically, can be seen as pursuing certain goals which are central to the propagation of a species. This is not to say that a narrow agent will never adopt reproduction as a goal, rather teleological goals are only as-if goals. An organism acts for the sake of a biological goal in the same way that a river flows for the sake of reaching the sea. In short, there is a goal but not a purpose.

The capacity of an agent to select and pursue goals has traditionally been associated with other notions which play a significant role is discussions of agents. Autonomy, intelligence, and rationality come to mind. Of these, rationality has a certain conceptual pride of place, for it is rationality which allows an agent to connect actions with goals. This is because it is in virtue of the requirements of rationality that an agent is constrained to act only in those ways which are literally consistent with the pursuit of their goals. This is why there is such a strong connection between the notion of success and the notion of rationality. But conceptually, these notions are not always the same thing.

III. Conceptions of Rationality

We begin by noting a simple fact about rationality: it is an elusive concept--perhaps more elusive than the concept of agency. In part this stems from its complexity. We can sensibly attribute rationality to persons, agents, and the like, but also to thinks like systems or patterns of organization. For example we might say that a system for distributing goods and services within an economy is rational or irrational. Further, rationality exhibits a fundamental duality; sometimes it is employed descriptively and at other times is operates normatively. Rationality is both a term of approbation and a means of demarcation. To say of an agent that it is rational is in one sense to praise it, in another sense to attribute certain characteristic to the agent. It remains unclear, however, whether the rational agent deserves the praise in virtue of the characteristics so attributed, or for some other reason.

1. Rational Agents and Rational Acts

Having noted the significant role of the normative side of rationality, let us turn to the descriptive. Agents are actors, but we need to be able to distinguish agents from their actions. If we are considering the attributive sense of rationality, we should consider whether a rational agent is an agent which possesses a specific characteristic (rationality), or which performs actions which possess the feature of being rational actions. In short, we are asking if rational agency involves rational agents, or agents performing rational actions. The need for such a distinction can be seen in the definition of 'rational agent' proposed by Russell and Norvig: "A **rational agent** is one that does the right thing. … At first approximation we say that the right action is the one that will cause the agent to be the most successful"(31). Here there is a clear emphasis upon the action, while there is no discussion of what it is about the agent itself that explains its tendency to do what is right. By these lights, any agent that succeeds is rational. However, Russell and Norvig later define an 'ideal rational agent' which has the capacity for selecting and performing actions relative to its perceptions of the environment and some internal knowledge base (33). Here the weight of ideal rationality is placed upon the internal workings of the agent, its pseduo-psychological workings. While this is only one possible pair of definitions, we believe that they reveal a fundamental feature of our concept of rationality as related to agents*: there are at least two distinct kinds of rational agents.*

We have previously noted that the capacity of an agent to pursue a goal is a central feature of the interesting sorts of agents. The ability to distinguish which agents have this capacity is a necessary condition for making sense of our proposed distinctions between broad and narrow agents on the one hand and direct and indirect agentive relations on the other. If agents can be distinguished from their actions then we should also distinguish *goal-oriented agents* from *goal-oriented actions*. The former will be those agents which are capable of pursuing goals, where this will probably involve some sort of capacity to represent the goal or goals. Goal oriented actions can be actions which are produced by goal oriented agents, in virtue of their goal orientation, or possibly actions which facilitate the attainment of a goal –defined as such by something other than the agent. We may be able to judge whether an agent is genuinely capable of being goal oriented, perhaps in virtue of its relative complexity or simplicity, but we would be hard put to mark a firm distinction between the two sorts of goal directed action. Unless we have some antecedent access to the motivation behind the action they are effectively indistinguishable. But this has been our problem all along.

2. Internal Rational Agents and External Rational Agents

In the case of actions which attain a goal, which succeed, there is no way, absent insight into motivations, to distinguish between actions which were performed because the agent recognized that there was a means-to-end connection between the act and the goal, and actions which were performed because of some less complex set of dispositional characteristics attributable to the agent, where the limit case is random action. From an *external* point of view, agents which achieve goals are alike in all immediately relevant respects. These are agents that "do the right thing." It requires an *internal* point of view to evaluate whether an agent acted the way it acted because it was implementing a program or a procedure designed or developed for the purpose of attaining the goal.

What this reveals is that a successful agent need not be an agent with specific characteristics of any sort. Some successful agents are agents who by accident or design lack the capacity to antecedently represent means-end relationships but who though action alone achieve their goals. The other sort of successful agent will be an agent who by accident or design is capable of representing means-end relationships, selecting or simply pursuing determined ends, and in virtue of those ends is capable of selecting a means either from a pre-determined set of options, or dynamically developed, given an awareness of the agent's environment and situation. We would, without fear of being contradicted assert that the latter sort of agent is an instance of a rational agent. Indeed it counts as rational either from an internal or external point of view. But to distinguish it from the alternative let us call such an agent *internally rational*. And let us call a successful agent absent the proposed internal machinery *externally rational*.

Some people might wish to deny that externally rational agents should be counted as rational at all. But those objectors would need to provide a basis for such a claim that would ground their objection from an external point of view. We have already seen in some detail that it is precisely those features which distinguish internally rational agents from other agents which have been so difficult to pin down conceptually. For they are the same features which allow us to mark the narrow agent from the broad and the direct agentive relation from the indirect. Admittedly if one is designing an agent-based system or model, one can build in the difference—at some level. But if we wish our model to be a model of something actual, e.g. a community, an organization, a society, a distribution network, an economy, a political system, etc., as a means to better under stand the actual something, then it is an open question whether we should attribute characteristics to agents in the model which we cannot justifiably claim that citizens, persons, animals, consumers, or politicians have. We are to a certain extent restricted to the *external* point of view in such matters.

3. What we can discover from the internal–external distinction

We are not proposing that there is any sense in which the external should supplant the internal notion of rationality**.** Our claim in this regard is that the ability to distinguish agents from acts opens the door to the possibility of agents which we cannot fail to label as rational while at the same time recognizing that this does not require that the agent or agents possess any particular underlying traits or characteristics. Recognizing a distinction between internally and externally rational agents also reveal certain features about how they are related. First, for an agent to be externally rational does not require that it is also internally rational. Second, if an agent is internally rational, then *ceteris paribus*, it should also be externally rational. Third, if we admit that internally rational agents are necessarily more complex than externally rational agents, then we need to posit an explanation for this difference. One mark of an internally rational agent is that it is capable of performing a set of behaviors that can be thought of as implementations of rules relating means to ends. But where do these rules originate? Why should some rules be internally represented as rules for achieving a goal? The obvious answer seems to be either that there were agents which achieved the goal in the past, or that the rules are created anew by the agent upon the occasion of having a certain goal. In the first case, the previously successful agents could not have been internally rational agents. This suggests that internal rationality is functionally parasitic upon external rationality. In other words, rules for success are derived from the as-if goal seeking and achieving behaviors of externally rational agents. In the second case, the agent would need to be some sort of hybrid agent. An agent capable of generating alternative behaviors for itself would need to have the ability to filter those with a high likelihood of success from those with a lower probability of success. To do so would probably require that the agent be able to internally simulate what amounts to a trial and error procedure and test for the external rationality of the options. In either case, the internally rational agent is possible because there is a hypothetical external agent which either actually or hypothetically was able to do what the internally rational agent is attempting, but without possessing the internal attributes proper to an internal agent. Fourth, externally rational agents are indistinguishable from irrational agents except at the level of behavior. They may be constrained by the features of their environment, or by their specific capacities for action, but they need not have the capacity to represent a goal. And assertions that they have achieved a goal are always made from an external viewpoint which functions to interpret their actions. Fifth, in distinction then to the external agent whose only claim on rationality is a function of its success, the internally rational agent both represents the goal, and represents its behaviors as directed toward the goal in question. Whether the internally rational agent succeeds or not is immaterial to whether the agent counts as rational.

4. Autonomy reconsidered in light of the External-Internal distinction

Agents of the interesting sort are usually understood to be autonomous as well as rational. What autonomy amounts to is a matter of some disagreement, but its derivation is from the Greek terms for self and law or rule, thus to be autonomous is to give the law to oneself or to be able to determine one's goals and/or actions. If we accept that there is a relationship between being autonomous and being rational, should the introduction of a distinction between externally rational agents and internally rational require that we revise our understanding of autonomy?

For an internally rational agent there are many axes along which the agent might move from being determined to being autonomous. For instance an internally rational agent might have a pre-determined goal or it might be capable of selecting goals for itself. Likewise, once a goal is determined, the agent might be able to select a specific goal seeking behavior in response to the goal, or it might be permitted to select from a repertoire of possible behaviors. Or its choice of behavior might be determined. The degree to which an internally rational agent is autonomous is a function of its relative complexity. An agent capable of detecting changes in its environment or in its own states might be able to employ that information in the selection of its goals or its behaviors. Alternatively, any internally rational agent that lacks this capacity for perception will not be capable of such radical change. Nevertheless a non-reactive, non-perceptive, internal agent with a single goal and a single behavior could still plausibly be seen as autonomous provided that the rule and behavior were in some sense intrinsic to the agent.

In the case of an externally rational agent, the agent will behave in a manner consistent with its capacities and limited by the constraints of the environment. From the external point of view this might be interpreted as the agent having the capacity to select goals and behaviors, but *there need not be such deliberation actually taking place*. However, when the external agent is free of extrinsic active control by another agent, it seems fair to say that at a minimal level the external agent is autonomous. The curious thing is that the autonomy of the internally rational agent consists in its ability to set goals and act according to rules while the autonomy of the externally rational agent consists not in it actually giving itself rules, but in their being no other possible source for the determination of its behavior. This means that autonomy, too, is subject to the external-internal distinction.

IV. Three Paradigms of Rational Agency

In this final section our aim is to draw some morals from the preceding analyses as they apply to the wider general project of modeling agents. We have seen that the notions of agent and rationality are far less precise that we should require of other notions which we would attempt to introduce into a formal system such as a model. Our analyses to this point strongly suggest that the basis for this lack of precision resides in an epistemological difficulty. We have no epistemic access to the internal characteristics of agents except in cases where the agents are products of our own design. The problem with a design based approach is that we cannot objectively justify attributing specific characteristics to agents, when the agents we are modeling are actual actors, be they persons, organizations, nations, etc. Our ability to build capabilities into agents pushes us to take up what we have called an internal point of view, while our epistemic situation requires that we take up an external point of view.

In light of this problematic, we suggest that instead of talking univocally about agents, we should adopt a *looser* sort of discourse about agent types or agent *paradigms*, where we can then sort models into models operating relative to one paradigm or another. We therefore wish to propose (without claiming to exhaust the possibilities) three paradigms for rational agency: The human paradigm, the procedural paradigm, and the emergent paradigm.

Paradigm One: The Human Paradigm

At an intuitive level one paradigmatic instance of a rational agent is a human being. One plausible interpretation of the claim that an agent is rational is that the agent is either acting as a human would act or deliberating as a human would deliberate. Likewise we employ our native ability to represent the minds and intentions of other human beings when we attribute mental or quasi-mental states such as belief, desire, intention, etc. to an agent. We say, as it were, that the agent is acting as a human being would if the human being had such and such a goal or desire.

Since we are human and we attribute our own capacities unto agents that seem to act in “rational” ways, we can easily come to think that we do in fact have a sort of epistemic access to the internal procedures that govern the behavior of the agent. But this assumption of access is an artifact of the human paradigm. This is not to say that this de-legitimates the human paradigm, rather attending to its status as an artifact helps us make determinations about the reliability of models that assume this sort of access is possible and already in play. It is within the human paradigm that we are able to mark out the distinction between broad and narrow agents and direct and indirect agentive relations. It also allows us to distinguish between internally and externally rational agents, recognizing that some agents may be capable of both sorts of rationality. Human agents undoubtedly fall into this hybrid category considered either individually or collectively. In this sense the development of models relative to the human paradigm will always be subordinate to developments in psychology and the social science. At the same time we must note that advancement in these areas probably takes the form of a change in the dominant human paradigm model.

Paradigm Two: The Procedural Paradigm

As we have seen, there is a further sense in which to be rational is to conform to certain procedures or rules. We have already labeled this type of rationality, internal rationality. To be an internally rational agent is in some sense to be a rule following agent, where the notion of rule is understood in a general way. Models that are developed relative to the procedural paradigm will obtain the sort of access to the internal operations of an agent by designing those operations into the model.   If a model assumes that agents have a single overriding goal and a preset repertoire of possible behaviors by means of which to pursue the goal then the model is squarely procedural. Rationality for such an agent resides in their recognition of and obedience to certain predetermined rules relating means to ends, or to the determination of goals relative to pre-determined values or sets of values, and so on. The procedural agent is not intrinsically creative, but neither is it useless. Its worth lies in the ability of procedural agents to perform specifiable tasks according to specifiable behaviors or sets of behaviors.

One important limitation upon such an approach is that it is so focused upon individual agents. The procedural paradigm is therefore not an apt paradigm for representing collective agency except at those levels of abstraction where collectives are understood to act as a single agent. An example of a procedural paradigm agent might be something like an expert system for playing chess or some sort of symbolic rule governed artificial intelligence project. Since the individual agent is drawn in such fine detail the task of constructing a model of a larger population of agents by applying this paradigm seems daunting. Indeed the computational resources necessary to implement a model which contained several independent iterations of a procedural agent model would be quite extensive, and perhaps prohibitively so. The development of agent models relative to the procedural paradigm will depend in part upon our ability to come up with ways of capturing and representing the knowledge and capacities of the agent. This will depend in large part upon what sorts of knowledge about the world, about the agent itself, about how the agent is related to the world, etc. count as important relative to the task or tasks assigned to the agent. But these are the same minimal requirement for constructing internally rational agents: the capacity to represent goals and rules for behaving which promote goal attainment relative to the capacities and knowledge of the agent.

Paradigm Three: The Emergent Paradigm

If an externally rational agent is an agent for which it is impossible to determine if it acts for the sake of a goal, whether or not it makes any actual choices, and if it actually behaves in accord with any rules or guidelines, in what sense is it rational? Such an agent counts as rational just in case it is successful in attaining something that counts as a goal to an external observer. Externally rational agents are marked by our inability to know if they have any of the psychological or quasi-psychological attributes proper to agents native to either the human paradigm or the procedural paradigm. In fact externally rational agents are literally indistinguishable from agents that do, in fact, lack these attributes.

Rational agents modeled relative to the *emergent paradigm* are agents for which it is possible to be rational while it is assumed that they specifically lack those attributes traditionally seen as necessary for rationality.

One might wonder if this requires that we adopt a stance whereby these agents are merely called 'rational' as a courtesy, for in what sense could they *be* rational if the lack what amounts to the capacity for reason? The answer is that the rationality of an externally rational agent is emergent. The notion of an emergent property is an old one, dating back at least to the work of Samuel Alexander and the other so-called British Emergentists in the 1920s. Their view was that not all physical properties and causal powers were fundamental properties and powers, some properties and powers were held to only come into play at higher levels of complexity in the organization of the physical world. In more recent contexts the notion of emergence has played a role in discussions of how global patterns of organization can emerge from complex and chaotic systems. In the present situation we suggest that externally rational agents are only able to exhibit their rationality as a consequence of their presumably unconditioned actions. Their rationality is revealed in their acts not in their underlying nature. In this sense it *emerges from the activities of agency* rather than being dictated by the supposedly fundamental features of rational agents.

A paradigm that allows for rationality to be an emergent property of agents will tend to develop models in a *bottom-up* rather than a *top-down* manner. In a top-down structure the relevant concepts, relationships, and objects will be predefined. In a bottom up model, only minimal constraints are imposed and as the model develops over time patterns of order and organization will emerge. An example of such a bottom-up development can be seen in the construction of an anthill by a colony of ants. The individual ants engage in very simple behaviors none of which is directly oriented at constructing an anthill. Yet if the system is permitted to develop the result is a complex structure with an efficient assignment of different uses to distinct parts of the anthill and a division of labor among the various parts of its population. Where the cognitive science analogue for the procedural paradigm was symbolic rule governed artificial intelligence, the emergent paradigm has similar affinities with so-called *artificial life* projects.

What this means is that agents modeled relative to the emergent paradigm will be given a minimal characterization in terms of their capacities to act and interact with the environment and possibly with other similar agents. One virtue of such agents is that on the basis of a few simple dispositional rules a wide variety of complex behaviors can emerge. What these agents lack is anything which would approximate the capacity for complex representation of goals or the environment; they exhibit no sign of intelligence, or of being governed by principles which are recognizable as intrinsically rational. But what they demonstrate is the ability to behave in ways that seem goal oriented and perform actions that would count as achieving those goals to an observer who seeks to interpret their actions. A further virtue of such an approach to modeling agents is that at root these agents are quite simple, easy to represent within a formal system, capable of mass iteration, and not computationally intensive in the way that procedural agents are. This approach is well suited to the representation of collectives, populations, or organizations. No single agent bears the burden of needing to be rational or plan or coordinate the collective behavior of these agents, their collective behavior is emergently rational, emergently organized, in the way it would be had it been planned by rational agents of other sorts.

V. Modeling Strategies

No doubt the project of agent modeling has many motivations. We sometimes model to describe and better understand what actual agents do. At other times we might with to develop a model which allows us to make predictions so that we may anticipate and take actions as occasioned by this foresight. Alternatively, we might want to represent what is, allow it to evolve and refine this development in the service of certain goals or purposes.

If our aim is to model rational agents which are essentially human beings, and what we are especially interested in are those subjects which speak in large degree to the psychological characteristics of these agents: their desires, their fears, their motivations, then modeling these agents relative to the human paradigm seems to be called for. This can be accomplished quite simply by using human beings to model human being. We see such modeling in political and advertising focus groups, in role-playing exercises, in war games, and perhaps even in theatre or literature.

If our aim is to try to capture the characteristics of agents capable of what we would ordinarily call rational thought, taken individually, then agent modeling relative to the procedural paradigm may be called for. Examples of such approaches are easily found in the annals of AI research or in any home computer with a chess program. As we suggested before, this approach is fundamentally limited in two respects: first, the approach is difficult to iterate which precludes its applicability to large populations of agents. Second, since these agents are limited to a predetermined set of possible behaviors, they lack a certain capacity to creatively react to changes in their situation.

However, if our aim in modeling agents is to represent the features of agents without appealing explicitly to their psychology, especially if we want to model large populations then agent modeling relative to the emergent paradigm is called for. Two examples in the extant literature with affinities to this paradigm are so called "artificial societies" approach described by Epstein and Axtell in their *Growing Artificial* Societies (1995) and the ISAAC (Irreducible Semi-Autonomous Adaptive Combat) Simulation system developed by Andrew Ilachinski (e.g. see his 1997a and 1997b) of the Center for Naval analysis. The central feature of this approach is that it allows behavior patterns and the characteristics of agents, including their rational capacities to emerge from relatively simple representations. In this way models relative to the emergent paradigm are unique in their capacity to go beyond what is given in the formal model itself.

In conclusion, our proposal to reconsider the question of what it is to be an agent has led to the recognition of at least three distinct paradigms for modeling agents of a non-mundane sort. We have suggested that just as the project of modeling agents is always at the service of other goals and purposes, the models we thereby seek to create are best conceived relative to these three paradigms. Each paradigm takes a different feature or set of features of rational agents as essential and provides a framework for modeling the sorts of agents in question. While less than perfectly precise, these paradigms were distinguished in response to a more general epistemological problem: the near impossibility of independently determining the aptness of the agent model to the agents it is intended to represent.   Each paradigm can be seen as an attempt to resolve this difficulty.

Works Cited

Anscombe, G. E. M. (1963). *Intention*, 2nd Ed. Basil Blackwell.

Austin, J.L. (1979a). “A Plea for Excuses”. in *Philosophical Papers*, 3rd Ed., Oxford: Oxford University Press, pp. 175-204.

Austin, J. L. (1979b). “Three Ways of Spilling Ink,” in *Philosophical Papers*, 3rd Ed., Oxford: Oxford University Press, pp. 272-287.

Epstein, J and R. Axtell (1995) *Growing Artificial Societies: Social Science from the bottom up.* Brookings Institution Press/MIT Press.

Ilachinski, A (1997a) “Irreducible Semi-Autonomous Adaptive Combat (ISAAC): An Artificial-Life Approach to Land Warfare,” 399pp, CNA, Research Memorandum 97-61.10, First Revision, August 1997 (DTIC Report ADA362371)

Ilachinski, A (1997b). “A Concise User's Guide to ISAAC-FL: ISAAC's Mission-Fitness Landscape Mapper Program”, 31pp, CNA, Annotated Briefing 97-88, September 1997 (DTIC Report ADA362401)

Luck, M. and M. d'Inverno. (2001) “A conceptual framework for agent definition and development”. *The Computer Journal*, 44(1):1—20.

Norman, T. J. and D. Long (1995). “Goal Creation in Motivated Agents,” in *Intelligent Agents: Theories, Architectures, and Languages*. (LNAI Volume 890)

Russell, S. and P. Norvig (1995). *Artificial Intelligence: A Modern Approach*. Prentice Hall, Englewood Cliffs, NJ.

Sowa. J. F. (1998). *Knowledge Representation: Logical, Philosophical, and Computational Foundations.* PWS Publishing Company.

Tarski, A (1944). “The Semantic Conception of Truth and the Foundations of Semantics,” *Philosophy and Phenomenological Research* 4: 341-376.

Wooldridge, M and N. R. Jennings (1995) “Intelligent Agents: Theory and Practice,” *Knowledge Engineering Review*, 10(2): 115-152

Wooldridge, M and N. R. Jennings. (1998) "Pitfalls of agent-oriented development," in *Proceedings of the Second International Conference on Autonomous Agents (Agents 98)*, , Minneapolis/St Paul, MN, pp. 385--391