Motivating the Imputation of Agent Cognition

Sven A. Brueckner, H. Van Dyke Parunak

Altarum
3520 Green Court, Suite 300
Ann Arbor, MI 48105 USA
{sven.brueckner, van.parunak}@altarum.org

Abstract

A cognitive model of an agent or a multi-agent system greatly enhances high-level human interaction with engineered systems. Modern work on "BDI agents" emphasizes explicit representation of cognitive attributes in the design and construction of an agent, but there are several circumstances in which this approach cannot yield the desired perspicuity, including fine-grained agents without explicit internal representation of cognitive attributes, agents whose inner structures are not accessible, and the emergent properties of swarms of interacting agents of any type. We propose that cognition can legitimately be imputed externally to a system irrespective of its internal structure, and we demonstrate this notion in examples, including swarm-based emergent path planning for unpiloted vehicles.

Introduction

Agents and multi-agent systems are engineered artifacts. They are constructed by people, in order to solve problems for people, and to be effective they must be understood and controllable by people. This simple observation accounts for the extensive popularity of cognitive metaphors in motivating and designing agents. People deal with one another in cognitive terms such as beliefs, desires, and intentions. Our intuitions and predictive instincts are tuned by generations of experience to manipulate epistemological, praxiological, and axiological categories, and we find it most natural to interact with computer systems in the same terms. This reality accounts for the growing popularity of so-called "BDI agents," those whose internal representations include structures that map explicitly onto cognitive characteristics (for examples, see (Haddadi and Sundermeyer 1996)).

In some cases, such a design approach to cognition is not feasible. The internal structure of some agents (e.g., neural networks or reactive architectures) may not lend themselves to interpretation as cognitive primitives, or may be unknown. Also, increasing use is being made of collections of agents with emergent properties, for which there is no central location where such a representation could live. Still, people must interact with such systems. How can the cognitive gap be bridged in these cases?

In this position paper we first motivate the need to impute cognition to individual agents and communities externally, based on their behavior, independent of their internal structure and then we illustrate the notion in a few examples.

Motivation

The growing popularity of BDI architectures attests to the intuitive need that people have to relate to their agents in cognitive terms. The motivation for this paper is the recognition that this need persists even when it is difficult or impossible to satisfy it through explicit representations of cognitive constructs in an agent or multi-agent system. We summarize three increasingly common circumstances that can frustrate the design approach to cognition, discuss the contribution that an analytic approach to cognition can make to research in multi-agent systems, and briefly summarize previous work on which our approach rests.

When does the design approach fail?

The design approach to cognition fails in at least three cases: non-symbolic agents, black-box agents, and decentralized agent communities with emergent properties.

Non-Symbolic Agents.—Explicit representation of cognitive constructs is usually grounded in a logical model of these constructs (a "BDI logic," e.g., (Rao and Georgeff 1998; M.Wooldridge 2000)), which in turn guides the construction and manipulation of symbolic structures in the agent's reasoning. There is increasing interest in models of agent reasoning that are numerical rather than symbolic (for example, in the swarm intelligence community (Parunak 1997; Bonabeau, Dorigo et al. 1999)). Reasoning in such agents may consist of neural networks, evaluation of polynomial or transcendental functions, or even weighted stochastic choice, with no clear mapping between the internal representation and useful cognitive concepts.

Black-Box Agents.—Agent internals may be invisible to the humans who need to interact with an agent or agent system, and thus not helpful in reasoning about the system's behavior. This "black box" status can occur in two ways. First, the population of an open system (Hewitt 1991) may

be constructed by many different people, using different models. As long as the sensors and effectors of individual agents are compatible with the environment shared by the agents, their internals can be inaccessible (and should be, to take full advantage of the modularity that agents can offer). Second, it is increasingly difficult to distinguish between synthetic and natural agents. A distributed military simulation, for example, may include both carbon-based and silicon-based agents, and a trainee interacting with the system through a computer terminal may not be able to tell the difference. The internal representations and processing of people (carbon-based agents) are in general not accessible to outside observers. Even introspection is notoriously unreliable.

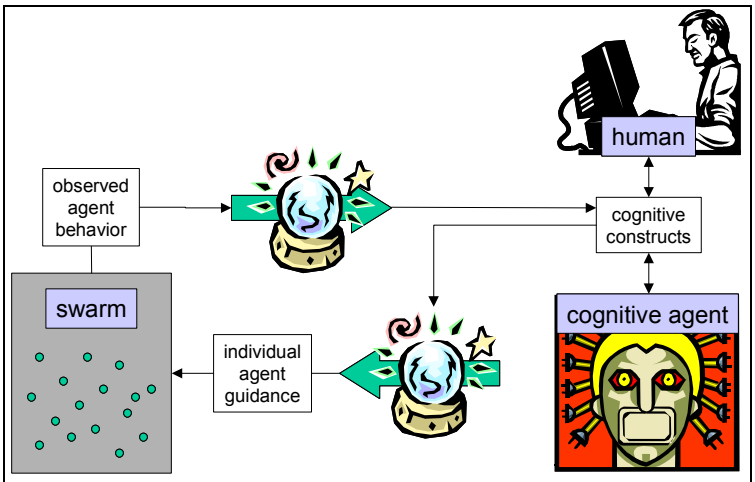

Figure 1: Human users and cognitive agents may interact with swarms of agents through cognitive constructs.

Emergent Behavior of Swarms.—A human stakeholder's interest in a multi-agent system is often at the level of the system as a whole rather than the individual agents, and cognition at this level can be difficult to represent. Consider, for example, resource allocation. A single agent seeks to maximize its own utility, but the purpose of the system as a whole is to maximize the net productivity of the entire set of agents, which usually requires that some individual agents receive less than they want. A human relying on a multi-agent resource allocation system will want to be assured that the system (qua system) wants to maximize overall productivity, and may inquire concerning actions it intends (again, as a system) to take. Team coordination is an active area of research. Some efforts take the design approach to cognition by maintaining explicit cognitive constructs such as team plans and intentions, for example in a team infrastructure to which all participants have access (Pynadath, Tambe et al. 1999). In other cases, coordination may take place emergently through the shared problem domain (Parunak, Brueckner et al. 2002) or through a non-symbolic infrastructure such as digital pheromones (Brueckner 2000), and in these cases another approach is needed to provide human stakeholders with the cognitive interface they require.

Why should agent researchers care?

The imputation of cognition to agents and agent systems will be useful to researchers and developers in MAS at several levels as illustrated in Figure 1.

Human interfaces to the three kinds of agents and agent systems outlined in the previous section depend on the ability to represent these systems cognitively. Our insights will enable the development of interfaces that make non-symbolic or black-box agents, or swarms of agents, more understandable and controllable to the human user. They will also reduce the bandwidth required to communicate between operators and systems, by permitting both status reports and commands to be expressed in cognitive terms rather than by transmitting large segments of the system state. This economy is particularly critical for swarming systems, in which the information needed to characterize the system's state is typically not localized spatially.

Hybrid agent systems combining conventional BDI agents with the kinds of agents and systems described in the previous section are becoming more common, and require interoperability between BDI agents and these other agents and systems. The imputation of cognition facilitate agent-to-agent interaction in two ways: as a tool used by a BDI agent to interpret the behavior of an agent or swarm without designed cognition, or as a "wrapper" around such agents or swarms to enable them to communicate with BDI agents on their own terms.

Introspection is the ability of an agent to reflect on its own behavior and that of groups in which it is involved, and is directly supported by an analytic approach to cognition. Even BDI agents may make use of such methods, for two reasons. First, full prediction of the dynamics of interaction is difficult and sometimes formally impossible, so that agents in a MAS must resort to interpretation of the group's actual dynamics in order to understand the full effect of their actions. Second, while a BDI agent may rationally plan its interactions with the real world, the world may constrain the effects of those interactions in unexpected ways (Ferber and Müller 1996), so that it may sometimes be helpful for such an agent to observe the effects of its own situated actions and interpret them cognitively.

In applications of real-world complexity it is very important that the effort directed at solving a problem is applied as effective as possible. For instance, commanding a battlefield requires to anticipate the next actions of the adversary and preparing the appropriate response. If we can infer the intention of the adversary from the previously observed operation, then we would be able to significantly reduce the number of possible next actions that we have to consider. Thus, imputed cognition may serve as a **search heuristic** and **predictor**. In the case of an adversary with limited rationality (a realistic assumption in real-world applications), predictions based on imputed cognition may be even out-

perform perfectly rational behavior, because we may be able to take advantage of our adversary's imperfections. In our battlefield example, the opposing forces may have chosen a strategy that is less optimal in the current scenario. So, if we were to assume a perfectly rational adversary, we would not be able to correctly anticipate the next moves.

What precedents are there for this approach?

We are not the first to suggest that cognitive characteristics can be applied usefully for the analysis of entities that were not designed from a cognitive perspective. Early work by Daniel Dennett (Dennett 1971; Dennett 1987) and John McCarthy (McCarthy 1979) emphasized the value of what Dennett called the "intentional stance," which was thoroughly analytical rather than design-oriented. In fact, Dennett contrasted it with the "design stance" that describes how a system is designed. This approach has recently been revised in the work of Jonker and colleagues (Jonker, Snoep et al. 2002; Jonker, Treur et al. 2002). Our approach is unique in using a non-symbolic dynamical systems formalism as the foundation for doing this imputation (Parunak and Brueckner 2003) (which readers should consult for technical details).

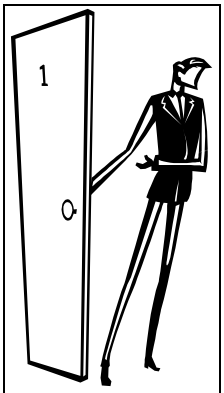

Figure 2: Intentions may be frustrated.

Examples of Imputed Cognition

In the following we present an example of cognition imputed to a single BDI agent whose intention is thwarted by its environment and we discuss the cognitive interpretation of a multi-agent system with emergent functionality.

A Frustrated BDI Agent

Consider the single-agent scenario illustrated in Figure 2. Internally, the agent has the explicitly represented intention to open the door and, because it is its belief that the door opens towards it, the agent tries to pull the handle.

Unfortunately, for the agent, the belief about the direction in which the door opens is wrong and consequentially the door does not open. Let's assume that the agent is not very smart and thus keeps pulling.

Now consider an external observer who finds the agent pulling the door handle. Without any access to the internal representation of the agent's beliefs, desires, and intentions, the observer should assume that the agent *wants to hold the door shut*!

This example demonstrates that imputation of cognition from observed behavior may provide the wrong answer, because what we observe is the result of the interaction of the agent's or agent system's explicit or implicit intentions with the constraints and dynamics of the environment.

Even though we may get the wrong answer, we need to consider the purpose of the external imputation of cognition based on observed behavior. For instance, we may want to use the derived intention to predict the future, in which case the wrong answer (agent wants to keep the door shut) is the right predictor (door stays shut).

While the chosen example may be overly simple, false interpretation of intentions based on observed behavior because of interaction effects with the environment is a common phenomenon. Simulation studies of racial segregation have shown for instance that even if the individual human have a very high tolerance threshold for racial diversity and thus the intention not to segregate, racially segregated neighborhoods emerge very robustly through the interactions in the society (Sander, Schreiber et al. 2000).

Emergent Path Planning

To illustrate the imputation of cognition to a community as opposed to an individual agent, we consider the ADAPTIV architecture for emergent path planning with digital pheromones (Parunak, Bruckner et al. 2002). The foundation of this system is a network of place agents, each responsible for a region of geographical space. The structure of the network is such that two place agents are connected just when the geographic regions they represent are adjacent. Each place agent hosts avatars, agents that represent targets, threats, and vehicles in its region. These avatars deposit primary pheromones indicating their presence and strength. The place agents maintain the levels of these pheromones using the pheromone dynamics of aggregation, evaporation, and propagation.

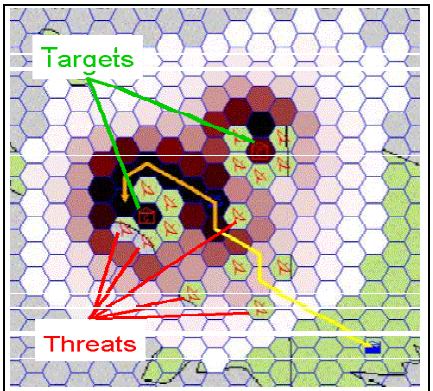

Figure 3: Alternate Decisions.—The system has currently decided on the western target, but a backup path has concurrently formed leading to the northern one.

As vehicles move through the space, their avatars continuously emit ghost agents that swarm over the network of place agents. Ghost agents execute an ant-inspired path planning algorithm (Parunak 1997), seeking target pheromones and avoiding threat pheromones. When a ghost finds a target, it returns to the issuing avatar, depositing a secondary pheromone. The aggregate strength of these ghost target pheromones emergently forms a path from the vehicle to the selected target.

Management of highly distributed systems of this sort is increasingly important for military and other applications, and other agents (both silicon and carbon) would benefit greatly by being able to relate to such a system at a cognitive level. What can be said about the

beliefs, desires, and intentions of such a system?

Beliefs.—As anticipated in our theoretical discussion, ADAPTIV's state can be correlated with the state of the world at several different levels, representing different degrees of knowledge on the part of different agents.

The presence of an avatar at a place constitutes belief by that place that there is an entity of the corresponding type in the region for which the place agent is responsible. This knowledge is local to an individual place agent. The system "knows" about these entities in the sense that it contains that knowledge, but not every member of the system has access to it.

Diffusion of primary pheromone from an avatar's place to neighboring places conveys knowledge to them of the presence of the entity nearby, and by sensing the gradient to neighboring place agents, they can estimate its direction. However, the pheromone field decays exponentially, and so extends only a limited distance.

Ghosts that encounter targets or threats embody that information in their movements and thus in the places where they deposit secondary pheromone. This pheromone condenses into a path that reaches from the avatar that sent out the ghosts to the target. At this point the issuing avatar has long-range knowledge (the best step to take toward one particular target, in light of any threats that may exist along the way).

Desires.—Concentrations of attractive and repulsive pheromones mark place agents as desirable or undesirable, and the attractors for the system consist of states in which avatars are located at place agents marked with locally maximal deposits of attractive pheromone and locally minimal deposits of repulsive pheromone. Since pheromone concentration is strongest at the place agent occupied by the depositing avatar, these maxima correspond to the locations of the targets and threats. The dynamic model warrants the semantics that the guided avatar wants to reach targets and avoid threats.

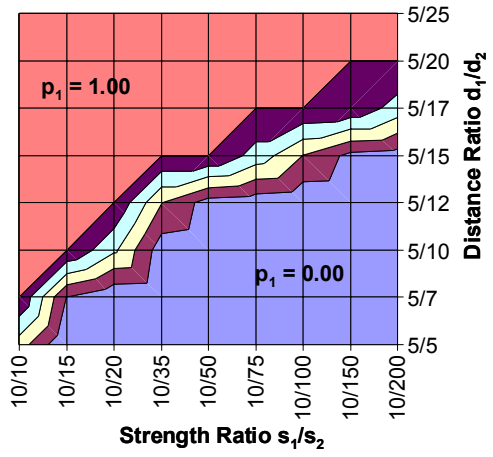

Figure 4: Attractors for alternate desires.—The vehicle is between two targets, and selects one or the other based on their relative distances and strengths.

Intentions.—The system's decision is manifested by the emergence of a dominant pheromone path leading around threats and to a particular target. Figure 3 shows the secondary pheromone field laid down by ghosts (shaded from dark to light), and the resulting path around the threats and to the higher priority target on the west. A secondary peak in the pheromone field leads to the northern target. If another threat were to be discovered blocking the primary path, the secondary peak would quickly capture the path, manifesting a decision at the system level. This decision cannot be attributed to any individual agent, but is an emergent property of the system. Nevertheless, using dynamical concepts, we can legitimately describe it as a decision, and other agents (synthetic and natural) can reason about it in cognitive terms.

Figure 4 shows a subset of the state space that illustrates basins of attraction for two decisions. The vehicle avatar is between two target avatars. The strength of target 1 is 10, and its distance to the vehicle is 5 units, and the strength and distance of the second target vary as indicated along the axes. On the upper left of the space, where the two targets are of comparable strength and target 1 is closer than target 2, the path forms to target 1 with probability 1. In the lower right, as target 2 grows stronger and closer compared with target 1, the path forms to target 2 with probability 1. The transition region between these two attractors is very narrow. Over most of the parameter space, the decision can be predicted with great certainty based on knowledge of the relative strength and distance of the different targets.

The emergent dynamics of pheromone-based path planning can seem surprisingly sophisticated from a cognitive point of view. Figure 5 shows two paths generated in approaching a target that is surrounded by air-defense installations. In the upper image, the defensive cordon has an opening at the far end, which the path finding mechanism readily locates and through which it guides the vehicle. The lower image shows the vehicle's path when the cordon

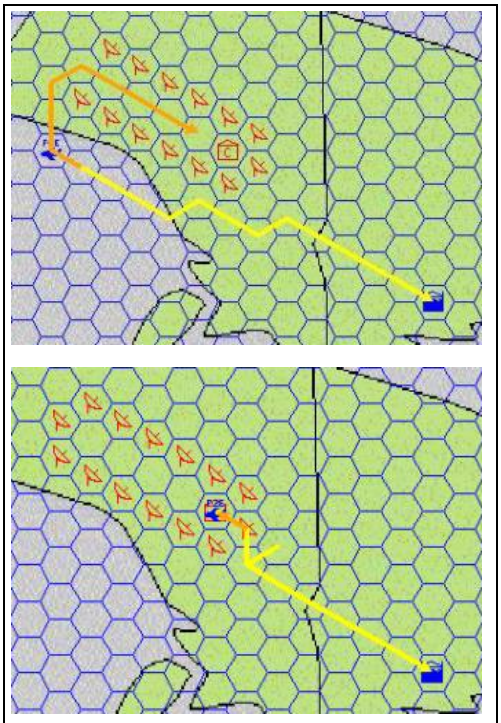

Figure 5: Imputed Indecision.—When the path is forced to penetrate a repulsive barrier, it takes longer to form, leading the vehicle to vacillate in a manner reminiscent of a person faced with a difficult choice.

completely blocks the way to the target. The path now forms through the nearest point of approach to the target. However, because of the repulsion generated by the air defense units, the path takes longer to form. In the meanwhile, the vehicle pauses in front of the barrier, and even makes a false move toward the northeast, until the path is strong enough to lead it in. The behavior is reminiscent of a person faced with a difficult but necessary choice, who hesitates through indecision before taking the plunge. This system has no explicit representation of the cognitive notions implicit in the idea of "a difficult but necessary choice," but its behavior is completely consistent with such cognitive notions, and it is appropriate to draw on such notions in describing it.

Conclusions

The dominant approach to cognition in agents is based on design, and most often on explicit representation of cognitive structures within individual agents. It is often desirable for other agents (both synthetic and natural) to be able to interact cognitively with agents and agent systems whose design is either inaccessible or else does not explicitly support cognition. When dealing with nonsymbolic agents or swarms with emergent properties, it is more natural to base this imputation on the observed behavior of the system, using concepts from dynamical systems theory.

Acknowledgments

This work was supported in part by DARPA (JFACC under con-tract F30602-99-C-0202, NA3TIVE under contract N00014-02-C-0458). The views and conclusions in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the Defense Advanced Research Projects Agency or the US Government.

References

Bonabeau, E., M. Dorigo, et al. (1999). Swarm Intelligence: From Natural to Artificial Systems. New York, Oxford University Press.

Brueckner, S. (2000). Return from the Ant: Synthetic Ecosystems for Manufacturing Control. Computer Science. Berlin, Germany, Humboldt University Berlin.

Dennett, D. C. (1971). "Intentional Systems." Journal of Philosophy **8**: 87-106.

Dennett, D. C. (1987). The Intentional Stance. Cambridge, MA, MIT Press.

Ferber, J. and J.-P. Müller (1996). Influences and Reactions: a Model of Situated Multiagent Systems. Second International Conference on Multi-Agent Systems (ICMAS-96).

Haddadi, A. and K. Sundermeyer (1996). Belief-Desire-Intention Agent Architectures. Foundations of Distributed Artificial Intelligence. G. M. P. O'Hare and N. R. Jennings. New York, NY, John Wiley**:** 169-185.

Hewitt, C. (1991). "Open Information Systems Semantics for Distributed Artificial Intelligence." Artificial Intelligence **47**: 79-106.

Jonker, C. M., J. L. Snoep, et al. (2002). "Putting Intentions into Cell Biochemistry: An Artificial Intelligence Perspective." Journal of Theoretical Biology **214**: 105-134.

Jonker, C. M., J. Treur, et al. (2002). "Temporal Analysis of the Dynamics of Beliefs, Desires, and Intentions." Cognitive Science Quarterly (forthcoming).

M.Wooldridge (2000). Reasoning about Rational Agents. Cambridge, MA, MIT Press.

McCarthy, J. (1979). Ascribing Mental Qualities to Machines. Philosophical Perspectives in Artificial Intelligence. M. Ringle, Harvester Press.

Parunak, H. V. D. (1997). "'Go to the Ant': Engineering Principles from Natural Agent Systems." Annals of Operations Research **75**: 69-101.

Parunak, H. V. D., S. A. Brueckner, et al. (2002). Synthetic Pheromone Mechanisms for Coordination of Unmanned Vehicles. First International Conference on Autonomous Agents and Multi-Agent Systems (AAMAS 2002), Bologna, Italy.

Parunak, H. V. D., S. Brueckner, et al. (2002). Co-X: Defining what Agents Do Together. Workshop on Teamwork and Coalition Formation, AAMAS 2002.

Parunak, H. V. D. and S. A. Brueckner (2003). Imputing Agent Cognition from Dynamics. Autonomous Agents and Multi-Agent Systems (AAMAS 2003), Melbourne, Australia.

Pynadath, D., M. Tambe, et al. (1999). Toward team-oriented programming. Workshop on Agents, theories, architectures and languages (ATAL'99), Springer.

Rao, A. S. and M. Georgeff (1998). "Decision procedures for BDI logics." Journal of Logic and Computation **8**(3): 293–344.

Sander, R. H., D. Schreiber, et al. (2000). Empirically Testing a Computational Model: The Example of Housing Segregation. Simulation of Social Agents: Architectures and Institutions, The University of Chicago.