Decision
2016, Vol. 3, No. 1, 40–53

2325-9965/16/$12.00 http://dx.doi.org/10.1037/dec0000042

Agency and Rationality: Adopting the Intentional Stance Toward Evolved Virtual Agents

Peter C. Pantelis
Rutgers University – New Brunswick and Indiana University – Bloomington

Timothy Gerstner, Kevin Sanik, Ari Weinstein, Steven A. Cholewiak, Gaurav Kharkwal, Chia-Chien Wu, and Jacob Feldman
Rutgers University – New Brunswick

The interpretation of other agents as intentional actors equipped with mental states has been connected to the attribution of *rationality* to their behavior. But a workable definition of "rationality" is difficult to formulate in complex situations, where standard normative definitions are difficult to apply. In this study, we explore a notion of rationality based on the idea of evolutionary fitness. We ask whether agents that are more adapted to their environment are, consequently, perceived as more rational and intentional. We created a 2-D virtual environment populated with autonomous virtual agents, each of which behaves according to a built-in program equipped with simulated perception, memory, and decision making. We then introduced a process of simulated evolution that pressured the agents' programs toward behavior more adapted to the simulated environment. We showed these agents to human subjects in 2 experiments, in which we respectively asked them to judge their intelligence and to dynamically estimate their "mental states." The results confirm that subjects construed evolved agents as more intelligent, and judged evolved agents' mental states more accurately, relative to nonevolved agents. These results corroborate a view that the interpretation of agent behavior is connected to a concept of rationality based on the apparent fit between an agent's actions and its environment.

Keywords: agency, simulated evolution, social cognition, theory of mind

Supplemental materials: http://dx.doi.org/10.1037/dec0000042.supp

When we interpret an object in the world as being alive and having a mind—what Dennett (1987) called "taking the intentional stance"—we will interpret its behavior in a qualitatively different manner from the behavior of inanimate objects without intentions (Leslie, 1987, 1994). But what makes a particular object seem animate? What does an object's

This article was published Online First October 5, 2015.

Peter C. Pantelis, Department of Psychology, Rutgers University – New Brunswick, and Department of Psychological and Brain Sciences, Indiana University – Bloomington; Timothy Gerstner, Kevin Sanik, and Ari Weinstein, Department of Computer Science, Rutgers University; Steven A. Cholewiak, Gaurav Kharkwal, Chia-Chien Wu, and Jacob Feldman, Department of Psychology, Rutgers University.

Timothy Gerstner is now at ALK Technologies, Princeton, NJ. Kevin Sanik is now at Amazon, Seattle, WA. Ari Weinstein is now at the Princeton Neuroscience Institute, Princeton University. Steven A. Cholewiak is now at the Vision Science Program, University of California, Berkeley. Gaurav Kharkwal is now at Bloomberg LP, New York City, NY. Chia-Chien Wu is now at the Department of Computer Science, University of Massachusetts at Boston Center for Computational Neuroscience and Neural Technology, Boston University.

This research was supported by the Rutgers IGERT program in Perceptual Science, NSF DGE 0549115 (http://perceptualscience.rutgers.edu/). We also owe thanks to the Robert J. Glushko and Pamela Samuelson Foundation and the Hellenic University Club of New York, who provided additional funding for this research and its presentation at scientific conferences. Thank you to Drs. Daniel Kennedy, Randy Gallistel, Rochel Gelman, and Brian Scholl for helpful comments.

Correspondence concerning this article should be addressed to Peter C. Pantelis, Department of Psychological and Brain Sciences, Indiana University – Bloomington, 1101 East 10th Street, Bloomington, IN 47405. E-mail: pcpantel@indiana.edu

behavior reveal about its mental state? These questions have historically received a great deal of attention in developmental psychology (e.g., Gelman, Durgin, & Kaufman, 1995; Gergely, Nádasdy, Csibra, & Bíró, 1995; Johnson, 2000; Kuhlmeier, Wynn, & Bloom, 2003; Onishi & Baillargeon, 2005; Williams, 2000) and philosophy (e.g., Goldman, 2006; Heal, 1996; Nichols & Stich, 1998; Stich & Nichols, 2003), and are receiving increasing attention in psychophysics with adult subjects (Barrett, Todd, Miller, & Blythe, 2005; Blythe, Todd, & Miller, 1999; Gao, McCarthy, & Scholl, 2010; McAleer & Pollick, 2008; Pratt, Radulescu, Guo, & Abrams, 2010; Pantelis & Feldman, 2012; Tremoulet & Feldman, 2000, 2006; Zacks, Kumar, Abrams, & Mehta, 2009) and in computational modeling (Baker, Saxe, & Tenenbaum, 2009; Burgos-Artizzu, Dollár, Lin, Anderson, & Perona, 2012; Crick & Scassellati, 2010; Feldman & Tremoulet, 2008; Kerr & Cohen, 2010; Pantelis et al., 2014; Pautler, Koenig, Quek, & Ortony, 2011; Thibadeau, 1986). Many of these past studies have relied on the direct parametric manipulation of the physical qualities of stimulus objects (e.g., their velocity or acceleration), and measurement of the resulting subjective percepts (such as perceived animacy). Here we employ a potentially richer approach, based on a virtual world populated by autonomous cognitive agents, depicted as triangular shapes moving about a virtual environment. These virtual agents elicit a strong impression of animacy and intentionality, behaving "intelligently" in a way that subjects can readily interpret (Pantelis & Feldman, 2012; Pantelis et al., 2014). In past studies, we asked how human subjects viewing these simulations infer the "mental states" underlying the agents' behavior. In the studies reported below, we introduce evolution into the simulation, and ask how adaptation—the degree to which the agents' behavior is tuned to their environment—influences subjects' judgments of intentionality and rationality. The goal is to test the hypothesis that the attribution of mental properties is based, to some degree, on the impression that the agent is adapted to its environment. This issue gets to the heart of the central cognitive faculty know as "theory of mind."

Agency and Mental States

The internal mental states of the agent (e.g., goals, intentions, beliefs, preferences, and emotions) *conditionalize* its behavior. That is, although the mapping from internal states to actions is typically an ambiguous, many-to-many relation (Malle, Moses, & Baldwin, 2001; Searle, 1984), certain actions become more or less likely depending on the agent's state. For this reason, the consideration of mental states plays a critical role in satisfactorily modeling another agent (i.e., possessing a "theory of mind"), and this is why an observer can, in principle, invert such a model to infer the hidden mental state of an agent on the basis its behavior (Baker et al., 2009; Pantelis et al., 2014).

Inferring the mental state of an observed agent by this method assumes that the agent exhibits behavior that is connected causally with its goals, or intentions, or beliefs, and so forth. If this assumption is not at least partially true, the agent's behavior will be unpredictable on the basis of its mental state, and, inversely, one will be unable to infer the mental state of the agent on the basis of behavior. In other words, unless the agent is "minimally rational" (Cherniak, 1981), observable agent behavior is nondiagnostic of its hidden states.

More strongly, if one does not construe the agent as being in some way rational—doing things *for a reason*—then it is meaningless to consider this object to be an agent, and its behavior ceases to be different from that generated by any other object class, guided by rules constrained by alternative considerations like Newtonian mechanics (Stewart, 1982; Williams, 2000). Positing that an object has a mind—a qualitatively different model for the generative processes producing the object's behavior—does not add explanatory power if this mind (and its beliefs, goals, intentions that lead to decisions) does not have any observable consequence vis-à-vis connecting means to ends.

Perspectives on Rationality

But beyond this minimum standard, what does it mean to be rational? The conventional normative answer is that a rational agent acts so as to maximize expected utility across possible outcomes (Wald, 1949; see Baron, 2004). Un-

der decision theory (which generalizes to game theory if there are multiple "players"), the agent takes into account all of the possible factors which are relevant to the problem in order to arrive at the optimal decision, which implies a body of knowledge and reasoning framework that "if not absolutely complete, is at least especially clear and voluminous" (Simon, 1955, p. 99). This burden can be difficult to meet in practice, especially when the set of relevant environmental variables is poorly defined, continuous, dynamic, or not fully observable to the agent—which of course it is in most naturalistic situations. The problem is further compounded when other agents are added to the environment, because the rational strategy will then depend on the possible strategies that can be adopted by these other agents (i.e., in a game theoretic setting).

An alternative perspective on agent choice relaxes the demands on the agent in terms of the amount of information, time, and computational power it is presumed to have at its disposal when making a decision (Kahneman, 2003; Simon, 1955). This notion of "bounded" rationality sacrifices the analytic precision and optimality of stricter mathematical models of rational choice, in favor of more approximate—but generally effective, and computationally tractable—methods of reasoning under uncertainty. Dennett (1987) simplifies the definition of agent rationality even further, and firmly embeds the concept of rationality into the environmental niche of the agent: What is (approximately) rational for the agent is what is adaptive with respect to its (empirical) evolutionary success, even if "the demands of nature and the demands of a logic course are not the same," and our cognitive and perceptual faculties may be "nothing more than a bag of tricks" (p. 51). This is a perspective which also flows through the work of Gigerenzer and Todd (1999), who coin the term "ecological rationality."

In this article we adopt this perspective, and use adaptive fitness in an evolutionary setting as a proxy for rationality. We argue that if the human capacity to interpret intentional actions relies on an assumption of rationality, then (a) subjects ought to be sensitive to the level of perceived rationality in observed agents, and (b) the more rational the behavior of the agent, the more effective subjects ought to be at inferring their mental states. Testing these predictions experimentally requires the following:

- a means for creating intentional agents to be used as stimuli;
- a definition of agent rationality that is well-defined and noncircular;
- a way to manipulate the level of rationality in stimulus agents; and
- a "ground truth" for assessing the accuracy of human subjects as they estimate the mental states of the agents.

The evolutionary paradigm we present below satisfies all of these criteria. The main element is a 2D simulation environment in which autonomous agents (nicknamed "IMPs" for "*I*ndependent *M*obile *P*ersonalities") compete in a simple game. These agents have modular perceptual and cognitive capabilities which determine their behavior and can be manipulated parametrically. Importantly, for the purposes of this study, the parameters governing the decision-making process of each IMP can also be *evolved*. We use an IMP's fitness within the environment as our operational definition of rationality, and let a simulated evolution select for this fitness.

Artificial Life, Simulated Evolution, and Ecological Rationality

Artificial life can be loosely defined as the modeling and simulation of biological processes or behaviors, with the goal of imitating life with increasing fidelity. The study of artificial life cuts across many disciplines and perspectives (Braitenberg, 1984; Carnahan, Li, Costantini, Tourè, & Taylor, 1997; Chaitin, 1970; Shao & Terzopoulos, 2007; Yaeger, 1994), and need not involve any evolutionary or genetic algorithms. But a particularly interesting set of questions arises when simulated Darwinian evolution is introduced into the artificial life simulacrum. The inclusion of evolution allows us to study how adaptive pressures can modify organisms' patterns of behavior, typically over a sequence of generations. Evolutionary approaches differ in many ways, but all involve the same basic collection of elements: (a) a representation of the agent that can be altered in some systematic way, (b) a method for generating a population of agents, (c) a means of assessing the "fitness" of each agent, and (d) a strategy for creating subsequent generations of agents based

on this assessment (i.e., "selection and genetic operators," Mitchell & Forrest, 1993).

The relative adaptiveness of various strategies can be assessed as being proportional to their respective frequencies in the evolved population (Bicchieri, 2009). Depending on the particular evolutionary algorithm and domain, this iterative process may ultimately converge to one optimal or approximately optimal strategy. In many other interesting cases, the resulting population of agents may have multiple modes representing coevolved stable equilibria, for example in the simulated coevolution of pursuit and evasion, or predators and prey (Cliff & Miller, 1996; Nolfi & Floreano, 1998; Reynolds, 1994). These treatments successfully model the concept of evolutionarily stable strategies first put forth by Smith and Price (1973), which merged the concept of the Nash equilibrium from game theory with the study of evolutionary dynamics. However, evolution—even the simulated variety—is a complex, noise-sensitive process. The outcomes are not always clean nor stable, and the dynamics can be chaotic (Nowak & Sigmund, 2004).

Our aim is not to demonstrate the particular effectiveness of our evolutionary algorithm versus any other, but instead to meet a minimal threshold of effectiveness for the purposes of the psychophysical experiments presented in this study: IMPs that have been subject to simulated evolution should, on average, be better adapted to the constraints of this domain than those which have not been subject to any such selective pressure. Given this, we can consider evolved IMPs to be operationally "more rational" than nonevolved IMPs (which have had the parameters of their internal programs set randomly). Our goal is then to compare human subjects' interpretations of evolved versus nonevolved agents in order to investigate the effect of the experimental manipulation of rationality on perceived mental properties.

To summarize the goals of the study, we aim to (a) present a virtual environment of agents, rich enough that rational behavior within it is difficult to define analytically; (b) introduce adaptive pressure into this environment in order to produce "evolved" agents; and then (c) investigate whether subjects are better able to interpret the intentional behavior of evolved agents, compared with nonevolved agents.

The IMPs and Their Environment

IMPs are virtual robots depicted on the computer screen as moving, isosceles triangles (see Pantelis & Feldman, 2012 and Pantelis et al., 2014 for previous experiments using autonomous virtual agents, but without an evolutionary component). In the tradition of Heider and Simmel (1944), the agents populating the virtual environment are rendered minimally as basic geometric shapes (see Figure 1), an approach that is designed to isolate the motion behavior of these stimuli as the critical aspect of the scene to be connected to how subjects perceive them (see also Gao & Scholl, 2011; Kerr & Cohen, 2010; Stewart, 1982; Thibadeau, 1986; Tremoulet & Feldman, 2000, 2006).

The ability of the IMPs to survive and thrive during each simulated generation of evolution involves successfully finding and collecting "food" in the environment while avoiding the harassment of other IMPs. An IMP's behavior is determined at all times by three factors: (a) which behavioral state it is in, (b) its knowledge of whether food and other IMPs are nearby (modulated by both perception and memory), and (c) its method for navigating this environment. An IMP can be in one of four states: "attack," "search," "flee," or "gather," each of which corresponds to a different subroutine

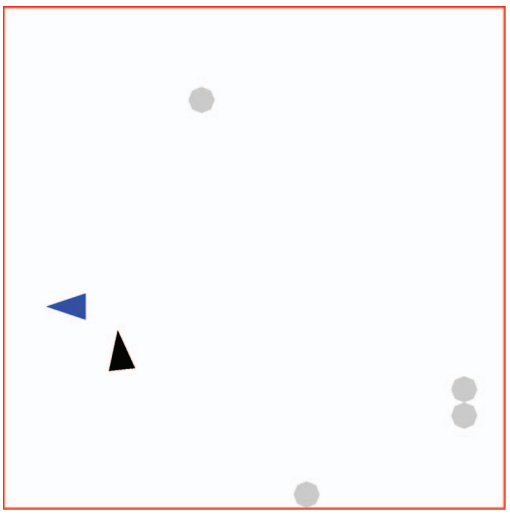

Figure 1. The virtual environment, in which the IMPs evolve. See the online article for the color version of this figure.

modulating behavior. What distinguishes one IMP from another, and allows these agents to be evolved, is a set of parameters that collectively determines the probability of transitioning to each of the four states at any given decision point. Each IMP's behavioral strategy is completely defined by this set of parameter settings. Some IMP strategies will be better than others, given the constraints on the IMP architecture, the structure of the virtual environment, and the rules of the "game." Through simulated evolution, we generate IMPs which, on average, behave more rationally than other IMPs which have not been selected by this process.

Modular Programming of the IMPs

Every IMP was programmed with the same modules governing perception, memory, and path planning:

Perception. Each IMP has a field of view modeled as a triangle (a 2D "cone") emanating forward from the center of its body. Unlike in (Pantelis et al., 2014), the IMPs employed in this experiment are allowed simulated vision such that the size, shape, and location of any object within its angular field of view (in this case, 60°) and within a distance of 3 1/3 IMP lengths is immediately perceived by the IMP and recorded in its mental map. In addition, if the perceived object is another IMP, the IMP becomes aware of its counterpart's internal state (attack, search, flee, or gather).

Memory. Using input returned from its vision module, the IMP continually updates a mental map of its environment. The environment is subdivided into a 150 × 150 grid, and the IMP believes each of the square subdivisions in this grid either contains another IMP, food, an obstacle (walls are considered to be obstacles), or is unknown.

Path planning. The IMP builds a 25 × 25 path planning grid, which subdivides the environment into square sections. The IMP traverses the environment by finding the shortest unimpeded path to the target location via this grid.

Decision making. Each IMP's decision making is governed by 45 parameters. One parameter determines how frequently the IMP makes a new decision, and was allowed to take on an integer value from 90 (i.e., the IMP updates its state/target every 90 frames, or 1.5 s) to 180 (i.e., the IMP updates its state/target every 180 frames, or 3 s); thus, all IMPs always stay in a given goal state for at least 90 frames before deciding to either remain in that state or transition to one of the other states, and readjusting its target location.

The other 44 decision making parameters increase or decrease the probability of the IMP transitioning to one of the four states (*attack*, *search*, *flee*, or *gather*) at the time of decision. These parameters collectively determine how an IMP weighs its own previous state, the state of the nearest other IMP, the distance to the nearest other IMP, the distance to the nearest food, and the IMP's current "health" (for a further explanation of "health," see next section) when making a decision. A detailed description of this decision-making module and its evolvable parameters can be found in Appendix A (in the online Supplemental Materials).

IMP Evolution

The simulated evolution began with a population of 100 randomly parameterized IMPs, and represented our pool of "nonevolved" IMPs.

Each iteration randomly selected two IMPs from the population and pitted them against each other. The IMPs could gain health by successfully collecting food, or lose health when struck by the counterpart IMP. How much health an IMP lost when struck depended on the respective states of the IMP and its counterpart. If an IMP's health ever reached zero, the IMP "died" and became a piece of food.

The two IMPs competed with each other for 3600 frames (the offline equivalent of a visually rendered 60 s scene presented at 60 frames per second). Afterward, the algorithm assessed the health of the two IMPs, and two new IMPs were created, each spawned from a previously surviving IMP with probability in proportion to the health of that parent IMP (if one of the IMPs had died in the competition, then the other IMP would necessarily be replicated twice). Then, the decision making parameters of each new agent were randomly tweaked, or "mutated," such that each new IMP bore some close "genetic" resemblance to its parent IMP, with slight alterations. The two new IMPs were reinserted into the population, reunited with the other 98

IMPs to create the next "generation" of evolution.

This procedure was repeated 3000 times, yielding a pool of 100 evolved IMPs. Further details about the evolutionary algorithm are included in Appendix A (in the online Supplemental Materials).

Evolution Results

Over the course of 3,000 generations of simulated evolution, many of the decision making parameters converged toward particular settings across the continually evolving population of 100 IMPs. It is beyond the scope of this article to discuss the evolutionary dynamics of each of the 45 decision-making parameters in detail, but we will mention one illustrative example.

Four of the parameters determined how relatively likely an IMP would be to transition into the four respective states (*attack*, *search*, *flee*, or *gather*), depending on how far away the IMP believed the other IMP was in the environment. A low setting on one of these parameters corresponded to a tendency to transition into the respective state when another IMP was nearby, and a high setting on one of these parameters corresponded to a tendency to transition into the respective state when another IMP was far away. As shown in Figure 2, the prevailing strategy toward which the IMPs evolved was to *attack* or *flee* when the other IMP was nearby, and to avoid *searching* or *gathering* in these situations, saving these behaviors for when the other IMP was at a safer distance.

Experiment 1: Discrimination of Evolved Versus Nonevolved IMPs

Method

Subjects. Fourteen undergraduates at Indiana University participated in the experiment and received course credit for their participation.

Stimuli. Each subject viewed the same set of 25 scenes, each 60 s in duration and generated in advance. The first scene was the same for each subject, whereas the other 24 scenes were presented in random order for each subject. Subjects sat approximately 2 feet from the display. Each prerecorded scene was presented within a 1440 × 1440 pixel window.

Each scene was populated with a black IMP, a blue IMP, and four pieces of food (depicted as gray octagons). A screenshot of this 2D environment is shown in Figure 1. This environmental configuration (60 seconds, two IMPs, four pieces of food) was identical to the evolutionary environment used to generate the evolved IMPs. All of the IMPs and food objects were initial-

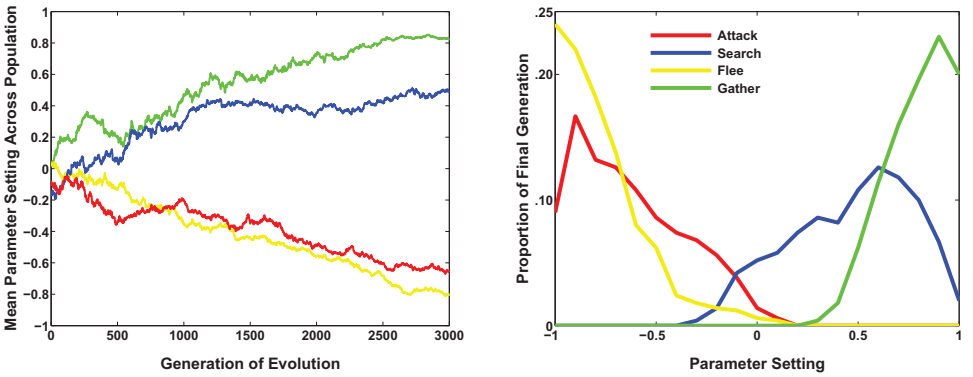

Figure 2. Evolutionary dynamics. Left: An illustration of how four of the decision-making parameters evolved over 3,000 generations of simulated evolution, with respect to the mean setting across the population of 100 IMPs. Right: Smoothed histograms of the frequency of the settings of these four decision-making parameters across the entire population of 100 IMPs after 3,000 generations of simulated evolution. See the online article for the color version of this figure.

ized at random locations in each scene. The "walls" of the environment were colored red.

When these scenes were generated and inspected, scenes were thrown out if one of the two target IMPs "died" over the course of the scene. A new scene was then generated to replace the discarded scene.

In half of the 24 randomly shuffled scenes, the black IMP was evolved (i.e., sampled from the 3000th generation of simulated evolution) and the blue IMP was nonevolved (i.e., randomly parameterized, or sampled from the 0th generation of simulated evolution). In the other half of trials, the black IMP was nonevolved and the blue IMP was evolved. Whether the IMP was evolved or nonevolved served as the primary independent variable, and we measured the effect of this categorical variable on the perceived intelligence of the agents.

Procedure. The experimenter first read aloud a plain English description of the IMPs and their environment (see Appendix B in the online Supplemental Materials), which described what was good or bad for the IMPs as they competed in the scenes the subject was about to observe, and briefly described the various possible IMP behaviors. The subject then observed 25 scenes (the first of which was treated as practice and not analyzed), in which one target IMP was colored black and the other was colored blue. The display froze at the last frame of each scene, cuing the subject to evaluate which of the 2 agents had behaved more intelligently over the course of the scene, on a scale from 1 (*definitely the blue agent*) to 6 (*definitely the black agent*). After making the judgment, the subject pressed the spacebar to move on to the next trial.

Results

We analyzed how subjects' rated the intelligence of the black IMP (compared with the blue IMP) on the 1–6 scale. Subjects' ratings of the black IMPs were, on average, 1.3 points higher on trials when the black IMP was evolved compared to trials when the black IMP was nonevolved, $t(13) = 13.1$, $p < .001$. Every subject's intelligence ratings discriminated between evolved and nonevolved IMPs better than chance (mean subject AUC = 0.73, $t(13) = 15.0$, $p < .001$; see Figure 3).

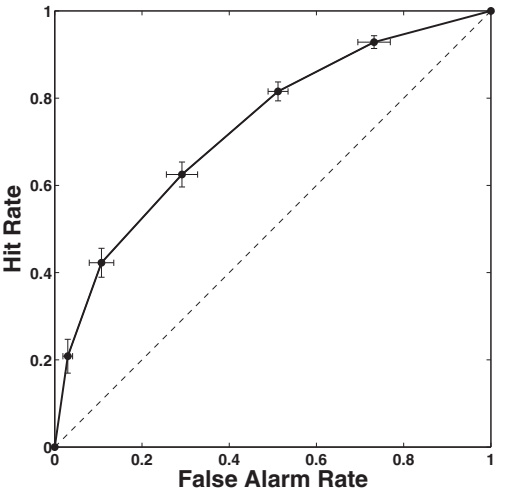

Figure 3. Experiment 1. ROC curve representing subjects' discrimination of evolved agents versus nonevolved agents, on the basis of perceived intelligence. Error bars represent standard error of the mean.

The evolved IMPs had been created to behave more adaptively in these scenes, and therefore should have performed better. And indeed, across the 24 scenes shown to subjects, the evolved IMP typically finished with a better "health" score than its nonevolved counterpart, $t(23) = 2.60$, $p = .02$. We attempted to analyze to what extent subjects may have used the observed performance of the IMPs, as they competed in the scenes, as a heuristic for assessing their intelligence.

We calculated the performance of a target IMP as the natural logarithm of the ratio of its health to its counterpart IMP's health at the end of the scene. We fit a linear regression model to each subject's data, predicting his or her ratings of the black IMPs' intelligence over the 24 trials from IMPs' performance. For the average subject, an IMP's performance was indeed a strong predictor of intelligence rating (mean $R^2 = .46$, $\beta = .82$, $t(13) = 9.15$, $p < .001$) and better as a single predictor of intelligence rating than the categorical variable of whether the black IMP was evolved or not (mean $R^2 = .19$). However, for the average subject, whether the IMP was evolved or not was a small but significant predictor of intelligence ratings, even when controlling statistically for IMP performance (mean R^2 of full model = .49, $\beta = .40$, $t(13) = 2.77$, $p = .02$). It appears likely that most subjects

assessed IMPs' intelligence heuristically by keeping a mental "score" of their relative performance, though a few probably employed some deeper thinking on the matter.

A comprehensive statistical analysis of how the 45 evolved IMP parameters (and their interactions) related to subjects' impressions of the IMPs' intelligence lies outside the focus of this article, and would also be statistically underpowered because of the high number of variables and the extensive multicollinearity among them. Much of this multicollinearity was the result of the manner with which the stimuli IMPs were generated, with many of the evolved IMPs converging toward a strong familial resemblance across many parameters. To partially overcome this limitation, we first turned our attention specifically toward the ratings subjects gave with respect to the 24 *nonevolved* IMPs featured in the scenes, for whom multicollinearity was less of an issue.

We constructed a multiple regression model predicting the average rating assigned by our subjects to the 24 nonevolved IMPs from a weighted linear combination of a subset of the 45 programmed parameters—in addition to IMP performance, which was already established as a dominant driver of subjects' ratings. We started with a single-variable model that predicted average intelligence rating from IMP performance alone. Then, we created 45 new candidate models by adding each of the 45 decision making parameters to the model. We tested each of these 45 candidate models with a split-half cross-validation (performed by splitting the subjects into training and test sets in every possible combination). We kept the most predictive variable (in terms of mean R^2 across all possible training/test splits) and added it to the model. We continued adding parameters to the model in this stepwise fashion until adding new parameters no longer appreciably improved the performance of the full model.

This statistical procedure resulting in a linear model that predicted subjects' ratings from 5 predictor variables at a cross-validated R^2 of .80. In addition to IMP performance (which predicted subjects' ratings by itself at $R^2 = .63$), nonevolved IMPs tended to be rated as more intelligent when programmed with (a) a tendency avoid transitioning from *search* to *flee*, (b and c) a tendency to transition to *attack* or *search* when the other IMP was *attacking*, and (d) a tendency to *flee* when the other IMP was *fleeing*.

We then considered the average ratings assigned to the *evolved* IMPs featured in the scenes, analyzed in the same fashion. In addition to IMP performance ($R^2 = .63$), evolved IMPs tended to be rated as more intelligent when programmed with (a) a tendency to switch states less frequently, (b) a tendency to avoid *gathering* when the other IMP was *searching*, and (c) a tendency to avoid *gathering* when it already had a high "health" score. This model (employing four predictive variables) predicted subjects' ratings at $R^2 = .70$.

Though these regression models "predicted" subjects' ratings well, we note that the sets of programmed parameters included in the two models did not overlap. It is difficult to say with confidence whether these particular parameters (selected from the much larger set) were actually robust predictors of subjective intelligence ratings, and whether it was appropriate in the first place to consider the sum perceptual consequence of the various programmed parameters as a linear combination. Thus, we consider these regression analyses to be exploratory, and instead highlight the two far stronger effects observed in Experiment 1: (a) Evolved IMPs were judged as more intelligent than nonevolved IMPs, and (b) the observed performance of the IMP (relative to its counterpart in the scene) was also a strong driver of subjects' intelligence ratings.

Experiment 2: Goal State Discrimination

In Experiment 2, the subject's task was to continually infer the underlying goal state—*attack*, *search*, *flee*, or *gather*—of the target (black) IMP. The task was identical to that used in Pantelis et al. (2014), which established that subjects are able to systematically estimate the goal state of a target IMP. Here we ask how this ability differs between evolved and nonevolved IMPs.

Method

Subjects. Eighteen undergraduates at Indiana University participated in the experiment and received course credit for their participation. One subject was excluded because of a failure to follow experimental instructions. Another subject

was excluded because his overall accuracy on the experimental task was much lower than any other subject, and not better than chance.

Stimuli. Each subject viewed the same set of 34 scenes, each 60 s in duration and generated in advance. The first four scenes were the same for each subject, whereas the other 30 scenes were presented in random order for each subject. Subjects sat approximately 2 feet from the display. Each prerecorded scene was presented within a 1440 × 1440 pixel window.

As in Experiment 1, each scene was populated with a black IMP, a blue IMP, and 4 pieces of food (depicted as gray octagons). In half of the 30 randomly shuffled scenes, both IMPs were evolved, and in the other half of trials, both IMPs were nonevolved. Because evolved and nonevolved IMPs behave according to systematically different programs, in the creation of the pool of stimuli we took care to create balance across the two conditions (evolved vs. nonevolved) with respect to a number of potentially confounding features.

To create this balance, we first generated an initial pool of 100 scenes for each condition (with every IMP in either population assigned as the target IMP in one scene and the counterpart IMP in another scene). We randomly selected 15 scenes from each pool and compared them along two dimensions: (a) The number of total times target IMPs in either condition transitioned among goal states, and (b) the amount of total time target IMPs in either condition spent in each of the four states. We resampled new candidates sets of stimuli from the two stimuli pools, and recompared them along these dimensions, until we found two sets of scenes that were well-balanced.

This rejection sampling procedure ultimately yielded 15 scenes featuring evolved IMPs and 15 scenes featuring nonevolved IMPs, where the nonevolved target IMPs (14.3 switches per scene) only switched states a total of 6 more times than the evolved target IMPs (13.9 switches per scene; $t(28) = .17$, $p = .87$), and the target IMPs spent approximately the same amount of time in each state (no sig. difference [$p > .60$] for each of the four states, see Tables 1 and 2). Thus, any observed difference in performance on this task would not be a result of systematic differences in how frequently IMPs switched states in either condition, or how much total time IMPs spent overall in one condition or another in either condition.

Procedure. In the first 2 scenes displayed to subjects, each IMP's true goal state was reflected in its color, changing several times over the course of each scene. Subjects were invited to speculate about what the changing colors of the IMPs meant. After these two scenes were complete, the IMPs and their states were explained to the subject. The experimenter read aloud a plain English description of the IMPs and their environment (the same as in Experiment 1; see Appendix B in the online Supplemental Materials), which described what was good or bad for the IMPs as they competed in the scenes they were about to observe, and briefly described the various possible IMP behaviors. After this explanation, subjects watched one more scene in which the color of the IMPs changed with their underlying goal states. These training scenes provided the opportunity to get a concrete sense of what the behaviors corresponding to these four goal states looked like, and what was meant by the "state" of the IMP.

Table 1
Average Confusion Matrix for Nonevolved IMPs

	Subject response						
Actual state	None	Attack	Search	Flee	Gather	FAR	*d′*
Attack (.19)	.019	**.142**	.674	.080	.086	.052	**.57**
Search (.44)	.084	.045	**.650**	.068	.154	.578	**.21**
Flee (.23)	.008	.083	.591	**.176**	.143	.071	**.60**
Gather (.15)	.052	.029	.446	.070	**.404**	.136	**.91**
	.050	.069	.610	.095	.176		

Note. Correct categorizations (hit rates) are shown in bold on the diagonal; also shown are false alarm rates (FAR) and *d′* calculated for each respective goal state. Mean proportion of IMP time spent in each state (i.e., base rate) is in parentheses. Overall response rates are included in the bottom row.

Table 2
Average Confusion Matrix for IMPs Sampled From the 3,000th Generation of Simulated Evolution

	Subject response						
Actual state	None	Attack	Search	Flee	Gather	FAR	*d′*
Attack (.19)	.008*	**.201****	.594**	.100	.097	.044	**.92****
Search (.45)	.072*	.042	**.721****	.078	.087**	.572	**.46****
Flee (.20)	.000*	.071	.657*	**.221***	.051**	.076	**.71**
Gather (.15)	.100*	.013*	.431	.040*	**.416**	.081**	**1.30****
	.049	.074	.640*	.106	.131*		

Note. Correct categorizations (hit rates) are shown in bold on the diagonal; also shown are false alarm rates (FAR) and *d′* calculated for each respective goal state. Mean proportion of IMP time spent in each state (i.e., base rate) is in parentheses. Overall response rates are included in the bottom row.
* Significant difference compared with nonevolved IMPs at $p < .05$. ** Significant difference at $p < .001$.

After these three training scenes, the subject saw 31 scenes (the first of which was treated as practice and not analyzed), in which one target IMP was colored black, and the other was colored blue. Subjects were instructed that they would be responding with respect to the black agent in each scene, and to indicate on the keyboard which state they thought this target agent was in at any given time. Four keys ('A,' 'S,' 'F,' and 'G') represented the four respective possible states; subjects were instructed to press a key as soon after a scene began as possible, and thereafter to press a key only when they thought the target IMP had transitioned into a new state.

The display froze at the last frame of each scene and the subject pressed the spacebar to move on to the next trial.

Results

Overall, subjects found the *gather* state to be the easiest to discriminate from the others (see Tables 1 and 2). The *search* state was the most difficult to discriminate, as subjects had a tendency to overestimate the proportion of time the IMPs were in this state (i.e., they had a high false alarm rate for this category).

When viewing scenes featuring nonevolved IMPs, subjects' responses matched the actual goal state of the target (black) IMP 40.2% of the time. But when viewing evolved IMPs, subjects made correct inferences 46.3% of the time. Every subject's performance was significantly facilitated when the IMPs in the scenes were evolved, $t(15) = 8.77$, $p < .001$; see Figure 4). This large and consistent effect (Cohen's $d = 2.19$) supported our hypothesis that the underlying intentions of an agent are easier to infer when this agent's behavior better conforms to subjects' expectations of rationality.

The enhanced performance in the evolved condition was not attributable to the better discrimination of any one particular goal state. With respect to the evolved IMPs, subjects demonstrated a significantly better hit rate for *attack*, *search*, and *flee*, and a significantly better false alarm rate for *gather*. Using *d′* as a metric for discrimination, subjects discriminated all four states better in the evolved condition, and discriminated *attack*, *search*, and *gather* significantly better.

Discussion

Evolved IMPs behave more effectively, on average, than their nonevolved counterparts, as would be expected as a result of the adaptive pressure provided by simulated evolution. Moreover, as established by the results of Experiment 1, they *appeared* to be more intelligent to subjects, demonstrating our main claim that the influence of evolution is cognitively available to observers. That is, evolved IMPs not only behave more rationally, they also convey the subjective impression of behaving more rationally. In Experiment 2, we asked how their enhanced rationality would influence subjects' more specific inferences about their intentions and beliefs. As the results showed, subjects are

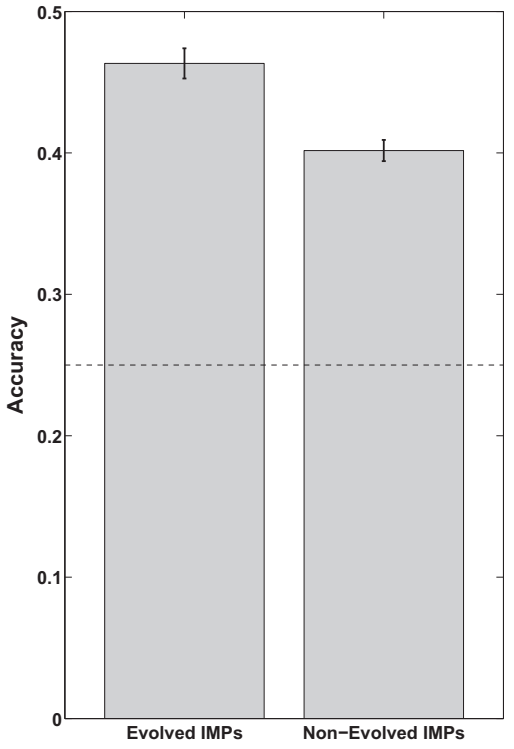

Figure 4. Experiment 2. Subjects' accuracy for estimating the underlying goal states of evolved and nonevolved IMPs in Experiment 2. Chance performance (25%) is shown with a dotted line, and error bars represent standard error of the mean.

indeed more effective at estimating mental states in evolved agents than in nonevolved ones, demonstrating that the improved rationality associated with adaptation conveys real benefits to observers attempting to understand their behavior.

More broadly, these experiments demonstrate that human intuitions about the mental processes generating an agent's behavior can be rigorously studied, and suggest that these intuitions feature a notion of rationality that goes beyond conventional normative models. As discussed above, some notion of rationality has already been extensively implicated in theory of mind (Dennett, 1987; Gergely et al., 1995; Baker et al., 2009). But exactly what rationality means in this context, and exactly what classes of behavior the system will interpret as rational, requires some clarification.

So how should one define the notion of rationality presumed by the inferential system? For all of the creatures found in nature—and also some found in artificial systems, such as our IMPs—it is quite difficult, if not impossible, to prescribe a normatively rational strategy on the basis of game theory or decision theory. The strategy space is too high-dimensional, and the set of relevant environmental variables too unpredictable and dynamic, to model tractably in normative terms. Moreover, as the enormous heuristics and biases literature can testify, a normative model of agent decision making can also be inaccurate as a descriptive theory (Gigerenzer & Goldstein, 1996; Johnson-Laird, 1983; Kahneman, Slovic, & Tversky, 1982; Krueger, 2014). Furthermore, when subjects fail to conform to normative models of rationality, some authors have argued that it is the normative model that has betrayed its own shortcomings, either by lacking robustness or efficiency, or making unsound assumptions about the nature of the problem (Cosmides & Tooby, 1994; Gigerenzer, 2008).

For these reasons, in this study we operationally defined agent rationality in a manner consistent with Dennett (1987): behaviors that have been selected by evolutionary pressure can be considered approximately rational, and more adaptive strategies are more rational—regardless of whether they are rational in a normative sense.

Perhaps, for the IMPs, there does exist one setting in their 45-parameter decision-making module which is provably optimal. We could not analytically derive this optimal IMP, but what we could do was simulate evolution—the "master of high-dimensional trial and error" (Taleb, 2012, p. 349)—to generate a pool of IMPs that was more rational than their nonevolved counterparts. Experiment 1 demonstrated that subjects *agree* that these IMPs are more intelligent, thereby demonstrating some level of consistency between our model of evolved rationality and subjects' expectations of rational agency. Experiment 2 then demonstrated that, as predicted, subjects are also better able to make sense of the behavior of more rational (evolved) IMPs. Agent rationality—as we have defined it—results in better discrimination of intentions.

In this study, we employed a definition of rationality based on evolutionary fitness, but the experimental approach we have used (the IMPs' environment) opens the door for testing a wide

range of hypotheses related to competing conceptions of rationality. For example, imagine if two sets of IMPs were created, one of which exhibited behavior prescribed by a normative theory, and another whose strategies were determined by a more naturalistic evolutionary process. Which set of IMPs would exhibit behavior that better matches human intuitions and expectations about agency? The modular nature of IMPs, and the ability to directly manipulate their cognitive and perceptual capabilities, invites novel experimentation of this kind.

Past experiments in the domain of theory of mind, dating to Heider and Simmel (1944), have frequently relied on stimuli generated according to the intuitions of either the experimenters or their subjects. This may result in too close a correlation between the nature of the independent and dependent variables—stimulus generation and subject response are both direct reflections of human intuitions about the subjective content of scenes, and therefore what is manipulated and what is being measured may be, at least to some extent, the same thing. The programming underlying the IMPs, admittedly, involves various design choices that were in part based on our judgments about the basic elements of cognitive architecture. But once the play button was hit in the simulation environment, what was displayed to subjects was entirely a result of their autonomous programs and the influence of adaptation on their parameters, and could not be known in advance. The apparently intelligent behavior of our IMPs was not created by intuition or by artistry in the animation, but rather was indirectly manipulated via evolution itself.

Our methodological approach also enables the creation of systematically varied agent stimuli, and allows for experimental paradigms in which the subject may interact with autonomous agents in real time. But most critically, we bring the psychophysics of theory of mind into closer analogy with the modeled process: in both cases, the goal is to estimate the qualities of a hidden process generating the observed world. Human beings rely critically on their ability to understand hidden mental processes, because the behavior of an entire class of objects in the world—agents—cannot be effectively explained or predicted according to alternate models like physics (whether naïve or rigorous). And, as we demonstrated in this study, when attempting to comprehend the behavior of agents, it is easier to make sense of the rational ones.

References

Baker, C. L., Saxe, R., & Tenenbaum, J. B. (2009). Action understanding as inverse planning. *Cognition, 113,* 329–349.

Baron, J. (2004). Normative models of judgment and decision making. In D. J. Koehler & N. Harvey (Eds.), *Blackwell handbook of judgment and decision making* (pp. 19–36). Malden, MA: Blackwell.

Barrett, H. C., Todd, P. M., Miller, G. F., & Blythe, P. W. (2005). Accurate judgments of intention from motion cues alone: A cross-cultural study. *Evolution and Human Behavior, 26,* 313–331.

Bicchieri, C. (2009). The handbook of philosophy of economics. In D. Ross & H. Kinkaid (Eds.), *The Oxford references library of philosophy,* (pp. 159–188). New York, NY: Oxford University Press.

Blythe, P., Todd, P. M., & Miller, G. F. (1999). How motion reveals intention: Categorizing social interactions. In G. Gigerenzer, P. M. Todd, G. F. Miller, & the ABC Research Group (Eds.), *Simple heuristics that make us smart* (pp. 257–286). New York, NY: Oxford University Press.

Braitenberg, V. (1984). *Vehicles*. Cambridge, MA: MIT Press.

Burgos-Artizzu, X. P., Dollár, P., Lin, D., Anderson, D. J., & Perona, P. (2012). Social behavior recognition in continuous video. *2012 IEEE Conference on Computer Vision and Pattern Recognition,* 1322–1329.

Carnahan, J., Li, S., Costantini, C., Tourè, Y. T., & Taylor, C. E. (1997). Computer simulation of dispersal by Anopheles gambiae s.l. in West Africa. *Artificial Life V, 5,* 387–394.

Chaitin, G. J. (1970). To a mathematical definition of "life." *ACM SICACT News, 4,* 12–18.

Cherniak, C. (1981). Minimal rationality. *Mind, 90,* 161–183.

Cliff, D., & Miller, G. F. (1996). Co-evolution of pursuit and evasion II: Simulation methods and results. In P. Maes, M. Mataric, J. Meyer, J. Pollack, & S. Wilson (Eds.), *From animals to animats 4: Proceedings of the fourth international conference on simulation of adaptive behavior*. Cambridge, MA: MIT Press Bradford Books.

Cosmides, L., & Tooby, J. (1994, May). Better than rational: Evolutionary psychology and the invisible hand. *The American Economic Review, 84,* 327–332.

Crick, C., & Scassellati, B. (2010). Controlling a robot with intention derived from motion. *Topics in Cognitive Science, 2,* 114–126.

Dennett, D. C. (1987). *The intentional stance*. Cambridge, MA: MIT Press.

Feldman, J., & Tremoulet, P. D. (2008). *The attribution of mental architecture from motion: Towards a computational theory* (Tech. Rep. No. TR-87). Piscataway, NJ: Rutgers University Center for Cognitive Science.

Gao, T., McCarthy, G., & Scholl, B. J. (2010). The wolfpack effect: Perception of animacy irresistably influences interactive behavior. *Psychological Science, 21,* 1845–1853.

Gao, T., & Scholl, B. J. (2011). Chasing vs. stalking: Interrupting the perception of animacy. *Journal of Experimental Psychology: Human Perception and Performance, 37,* 669–684.

Gelman, R., Durgin, F., & Kaufman, L. (1995). Distinguishing between animates and inanimates: Not by motion alone. In D. Sperber, D. Premack, & A. J. Premack (Eds.), *Causal cognition: A multidisciplinary debate* (pp. 151–184). New York, NY: Oxford University Press.

Gergely, G., Nádasdy, Z., Csibra, G., & Bíró, S. (1995). Taking the intentional stance at 12 months of age. *Cognition, 56,* 165–193.

Gigerenzer, G. (2008). Why heuristics work. *Perspectives on Psychological Science, 3,* 20–29.

Gigerenzer, G., & Goldstein, D. G. (1996). Reasoning the fast and frugal way: Models of bounded rationality. *Psychological Review, 103,* 650–669.

Gigerenzer, G., & Todd, P. M. (1999). Ecological rationality: The normative study of heuristics. In G. Gigerenzer, P. M. Todd, & the ABC Research Group (Eds.), *Ecological rationality: Intelligence in the world* (pp. 487–497). New York, NY: Oxford University Press.

Goldman, A. (2006). *Simulating minds: The philosophy, psychology and neuroscience of mindreading*. New York, NY: Oxford University Press.

Heal, J. (1996). Simulation, theory, and content. In P. Carruthers & P. K. Smith (Eds.), *Theories of theories of mind*. New York, NY: Cambridge University Press.

Heider, F., & Simmel, M. (1944). An experimental study of apparent behavior. *American Journal of Psychology, 57,* 242–259.

Johnson, S. C. (2000). The recognition of mentalistic agents in infancy. *Trends in Cognitive Development, 16,* 637–656.

Johnson-Laird, P. N. (1983). *Mental models: Towards a cognitive science of language, inference, and consciousness*. Cambridge, MA: Harvard University Press.

Kahneman, D. (2003). A perspective on judgment and choice: Mapping bounded rationality. *American Psychologist, 58,* 697–720.

Kahneman, D., Slovic, P., & Tversky, A. (Eds.). (1982). *Judgment under uncertainty: Heuristics and biases*. Cambridge, UK: Cambridge University Press.

Kerr, W., & Cohen, P. (2010). Recognizing behaviors and the internal state of the participants. *IEEE 9th International Conference on Development and Learning,* 33–38.

Krueger, J. I. (2014). Heuristic game theory. *Decision, 1,* 59–61.

Kuhlmeier, V., Wynn, K., & Bloom, P. (2003). Attribution of dispositional states by 12-month-olds. *Psychological Science, 14,* 402–408.

Leslie, A. M. (1987). Pretense and representation: The origins of "theory of mind." *Psychological Review, 94,* 412–426.

Leslie, A. M. (1994). ToMM, ToBy, and agency: Core architecture and domain specificity. In L. A. Hirschfeld & S. A. Gelman (Eds.), *Mapping the mind: Domain specificity in cognition and culture*. New York, NY: Cambridge University Press.

Malle, B. F., Moses, L. J., & Baldwin, D. A. (2001). Introduction: The significance of intentionality. In B. F. Malle, L. J. Moses, & D. A. Baldwin (Eds.), *Intentions and intentionality: Foundations of social cognition* (pp. 1–24). Cambridge, MA: MIT Press.

McAleer, P., & Pollick, F. E. (2008). Understanding intention from minimal displays of human activity. *Behavior Research Methods, 40,* 830–839.

Mitchell, M., & Forrest, S. (1993). Genetic algorithms and artificial life. *Artificial Life, 1,* 267–389.

Nichols, S., & Stich, S. (1998). Rethinking cocognition: A reply to Heal. *Mind & Language, 13,* 499–512.

Nolfi, S., & Floreano, D. (1998). Coevolving predator and prey robots: Do "arms races" arise in artificial evolution? *Artificial Life 4, 4* 311–335.

Nowak, M. A., & Sigmund, K. (2004). Evolutionary dynamics of biological games. *Science, 303,* 793–799.

Onishi, K. H., & Baillargeon, R. (2005). Do 15-month-old infants understand false beliefs? *Science, 308,* 255–258.

Pantelis, P. C., Baker, C. L., Cholewiak, S. A., Weinstein, A., Wu, C., Tenenbaum, J. B., & Feldman, J. (2014). Inferring the intentional states of autonomous virtual agents. *Cognition, 104,* 360–379.

Pantelis, P. C., & Feldman, J. (2012). Exploring the mental space of autonomous intentional agents. *Attention, Perception, & Psychophysics, 74,* 239–249.

Pautler, D., Koenig, B., Quek, B., & Ortony, A. (2011). Using modified incremental chart parsing to ascribe intentions to animated geometric figures. *Behavior Research Methods, 43,* 643–665.

Pratt, J., Radulescu, P., Guo, R., & Abrams, R. A. (2010). It's alive! Animate motion captures visual attention. *Psychological Science, 21,* 1724–1730.

Reynolds, C. W. (1994). Competition, coevolution and the game of tag. In *Artificial life IV: Proceedings of the fourth international workshop on the synthesis and simulation of living systems*. Cambridge, MA: MIT Press.

Searle, J. R. (1984). *Minds, brains and science*. Cambridge, MA: Harvard University Press.

Shao, W., & Terzopoulos, D. (2007). Autonomous pedestrians. *Graphical Models, 69,* 246–274.

Simon, H. A. (1955). A behavioral model of rational choice. *The Quarterly Journal of Economics, 69,* 99–118.

Smith, J., & Price, G. (1973). The logic of animal conflicts. *Nature, 246,* 15–18.

Stewart, J. A. (1982). *Perception of animacy* (Unpublished doctoral dissertation). University of Pennsylvania.

Stich, S., & Nichols, S. (2003). Folk psychology. In *The Blackwell guide to philosophy of mind* (pp. 235–255). Oxford, UK: Basil Blackwell.

Taleb, N. N. (2012). *Antifragile: Things that gain from disorder*. New York, NY: Random House.

Thibadeau, R. (1986) Artificial perception of actions. *Cognitive Science, 10,* 117–149.

Tremoulet, P. D., & Feldman, J. (2000). Perception of animacy from the motion of a single object. *Perception, 29,* 943–951.

Tremoulet, P. D., & Feldman, J. (2006). The influence of spatial context and the role of intentionality in the interpretation of animacy from motion. *Perception & Psychophysics, 68,* 1047–1058.

Wald, A. (1949). Statistical decision functions. *The Annals of Mathematical Statistics, 20,* 165–205.

Williams, E. W. (2000). *Causal reasoning by children and adults about the trajectory, context, and animacy of a moving object* (Unpublished doctoral dissertation). University of California, Los Angeles.

Yaeger, L. S. (1994). Computational genetics, physiology, metabolism, neural systems, learning, vision, and behavior or PolyWorld: Life in a new context. In C. Langton (Ed.), *Proceedings of the Artificial Life III Conference* (pp. 263–298). Reading, MA: Addison Wesley.

Zacks, J. M., Kumar, S., Abrams, R. A., & Mehta, R. (2009). Using movement and intentions to understand human activity. *Cognition, 112,* 201–216.

Received May 19, 2014
Revision received June 9, 2015
Accepted August 10, 2015 ■