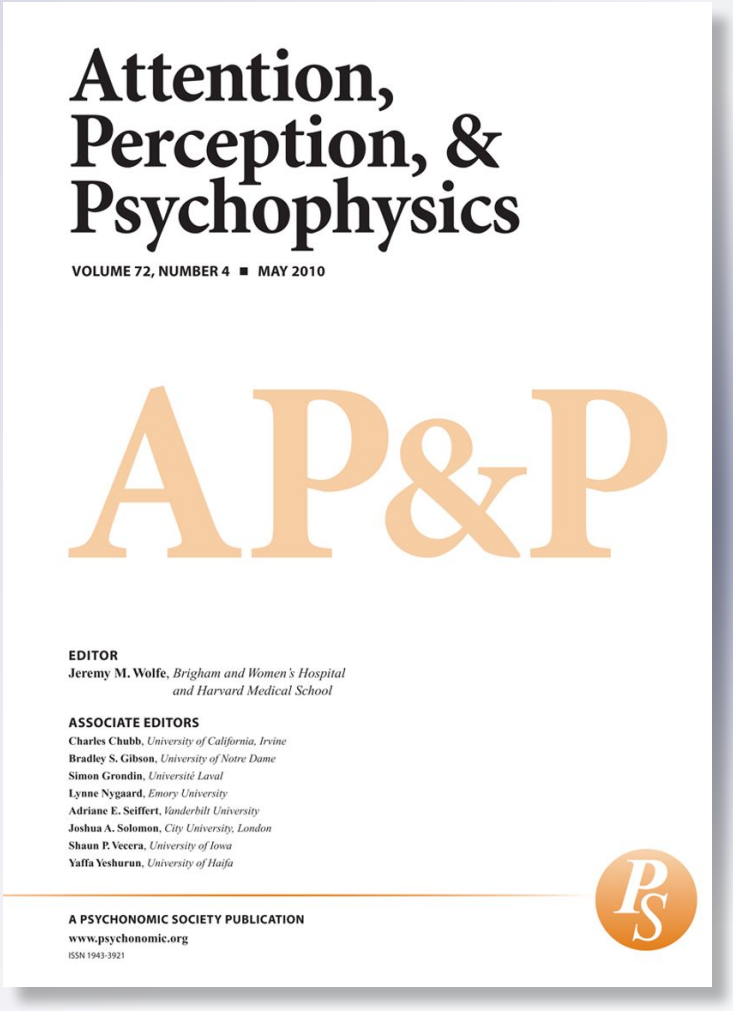
# *Exploring the mental space of autonomous intentional agents*

**Peter C. Pantelis & Jacob Feldman**

**Attention, Perception, & Psychophysics**

ISSN 1943-3921
Volume 74
Number 1

Atten Percept Psychophys (2012)
74:239-249
DOI 10.3758/s13414-011-0215-6




Springer

Atten Percept Psychophys (2012) 74:239–249
DOI 10.3758/s13414-011-0215-6

# Exploring the mental space of autonomous intentional agents

**Peter C. Pantelis · Jacob Feldman**

Published online: 18 October 2011


**Abstract** The ability to infer the intentions of other agents on the basis of their motion is a critical psychological faculty. In the present study, we examine a key question underlying this process, namely: What are the psychologically natural categories of intentional agents and actions? To investigate this question empirically, we use displays containing a number of autonomous, independently programmed agents moving about a two-dimensional environment and interacting with one another. Each agent behaves according to its own simple program, controlled by a small number of parameters that define its "personality." We probe participants' impressions of the similarities among the behaviors of the various agents, and then use multidimensional scaling in an attempt to recover the subjective mental space of agent types. An important variable underlying this space turns out to be a parameter that determines how the agent reacts to a nearby agent at one critical distance. A follow-up experiment suggests that variation along this parameter ultimately contributes to modulating a more fundamental perceptual dimension that reflects how "hostile" or "friendly" the agents appear to be.

**Keywords** Motion: Biological · Visual perception · Motion: *Other

Intelligent agents can and must distinguish between animate and inanimate objects they encounter in the world. Even infants make this distinction, apparently possessing a naive theory of other beings' mental states and intentions (Gergely, Nádasdy, Csibra, & Bíró, 1995; Johnson, 2000; Keil, 1994; Leslie, 1984). Socially intelligent agents naturally conceive of other humans as animate, mentalistic agents with independent perceptions and motivations. We further benefit from being able to infer the intentions of other agents in the environment. This is essential for understanding and predicting others' behavior, a prime skill whether for chess players contemplating their moves or for gazelles and lions engaging in mutual scrutiny on the African plain. The understanding of intentions is also critical for parsing of the perceptual flow of observed activities into meaningful, segmented events (Zacks, 2004; Zacks, Kumar, Abrams, & Mehta, 2009).

One particularly salient cue for intention is motion, although it is admittedly only one cue among many (Gelman, Durgin, & Kaufman, 1995). Ever since the famous study of Heider and Simmel (1944), it has been well known that participants will readily ascribe intentionality even to simple moving geometric figures moving about sparse environments. A handful of studies have shown that varying the motion of simple geometric figures along certain parameters (e.g., speed, trajectory) can influence the perception of animacy and intentions (Dittrich & Lea, 1994; Gao, Newman, & Scholl, 2009; Gao, McCarthy, & Scholl, 2011; Tremoulet & Feldman, 2000, 2006). Animate motion captures visual attention (Pratt, Radulescu, Guo, & Abrams, 2010). But the factors creating the impression of animacy, and the computational mechanisms of intentional interpretation, are still poorly understood.

In the present study, we investigate how adult human observers use visual motion to infer an agent's goals, behavioral dispositions, or other temperamental or psychological characteristics, given that the observer attributes animacy and intentionality to this agent. In any computational treatment of intention estimation (e.g., Baker et al., 2006, 2009; Feldman & Tremoulet, 2008), assumptions must always be made about the class of intentional agents to be considered as candidate models. For example, in a Bayesian formulation of the problem, priors must be defined over some hypothesis space

P. C. Pantelis (✉) · J. Feldman
Rutgers University,
New Brunswick, NJ, USA
e-mail: petercp@eden.rutgers.edu

of intentional agent types. But the empirical bases for such assumptions are still very much open questions. What are the natural psychological classes of agent types? What are the natural parameters underlying the space of intentional agents? In the present study, we aim to take a step forward in answering these questions, thus helping to provide an empirical ground for future computational models of the intention inference problem.

Indeed, the nature and structure of this intuitive mental space have been discussed only very speculatively in the literature. Barrett, Todd, Miller, and Blythe (2005) have argued that it probably includes such natural action classes as chasing, courting, following, guarding, fighting, and playing. Gao et al. (2009) and Gao et al. (2010) have investigated the perception of *chasing* in particular, which they argue is an especially simple and salient type of animate behavior. But, by design, their studies are limited to this one action class and do not attempt to survey the range of other types. Other studies have presented subjects with scenes constructed to resemble these different "natural categories" of dyadic interaction, and they demonstrate that participants are reliably able to categorize these scenes, even in degraded forms for which motion is the only salient cue (Barrett et al., 2005; McAleer & Pollick, 2008). But few empirical data address the question of how the mind naturally structures the space of possible actions.

In contrast with most previous experiments, the scenes we present to participants have *not* been preconstructed to convey particular categories of interaction. Our aim is to show participants a broad array of agent interactions—from a richer and more general collection of possibilities—in an attempt to allow participants' minds to impose *their own* structure on the agent space. The way we produce the desired scenes is also novel: We program the agents inhabiting these scenes to behave *autonomously*, which results in often chaotic multi-agent interactions that we cannot predict in advance.

In some ways, our experimental displays resemble the rich artificial life environments pioneered by Yaeger (1994) and Terzopoulos, Tu, and Grzeszczuk (1994). But the way such environments are understood by human participants—the psychology of the observation—has not been studied. For example, virtual agents in some of these environments have been endowed with hierarchies of goals (Shao & Terzopoulos, 2007), but the ability of human observers to perceive these goals or intentions, according to the motion of the agents, has never been investigated empirically.

In Experiments 1 and 2, we show participants scenes in which autonomous agents drawn from a larger set interact. We then ask participants to implicitly judge the similarities among these "agent stimuli." From these similarity judgments, we attempt to extract the natural clusters and cleavages present within the stimulus space of intentional behavior, with multidimensional scaling (MDS).

Attempting to infer the dimensionality and structure of a mental space from similarity judgments is a well-established application of MDS with a long history (Torgerson, 1958; Shepard, 1980). Of course, the dimensions of the inferred space need not correspond to spatial dimensions, and the dimensions uncovered by an MDS solution are indeed often difficult to interpret. Experiment 3 is explicitly designed to help clarify the results of Experiments 1 and 2 by unraveling the "semantics" of the features uncovered by the MDS.

Displays were programmed using the breve Simulation Environment (Klein, 2002), an open-source software package freely available at www.spiderland.org.

## Programming lifelike automata

In designing and coding the agent behaviors, we aimed to employ a simple programming scheme that would impose minimal structure on the agents' interactions but, nonetheless, would be capable of producing a rich variety of lifelike agent behaviors.[1] We programmed the triangular agents to behave autonomously, each running its own independent program. Inspired by the work of Braitenberg (1984), we aimed to create rule-governed agents that—notwithstanding the simplicity of their programs—yield vivid and lifelike behaviors that give participants a strong impression of intentions.

*Agent design* Rather than presenting participants with prefabricated animations, we populate simulations with autonomous agents and then allow these simulations to run for a predetermined length of time (15 s). Each agent starts off at a random location within the simulation environment. The agent always orients one vertex of its triangular body (that which lies on its axis of symmetry) in the direction of its movement, inducing the impression that this leading vertex is the agent's "head" (see Tremoulet & Feldman, 2000). When an agent collides with either another agent or the edge of the scene, it "bounces off" (i.e., it is assigned a random velocity vector) for one iteration of the simulation before reverting back to its normal program.[2]

[1] Note that the programming scheme we employ here is only one possible choice among many. The design of lifelike agents is a complex and multifaceted problem that extends far beyond the scope of our research. For us, these simple automata are merely tools for aiding an empirical study of the perception of intention.

[2] In Experiment 1, this sometimes resulted in jerky and unnatural-looking behaviors at agent collisions, so in Experiments 2 and 3, we changed collision behavior slightly: Agents in these experiments bounced off each other for a full .2 s at a random velocity vector. In no experiment was this "bouncing" designed to reflect or simulate actual physical laws.

This "normal program" is that of a simple reflex agent. At each iteration of a simulation, an agent finds the nearest *other* agent within the scene and then sets its own acceleration toward or away from that agent according to a set of six parameters contained in its program. These six parameters, which we will refer to as *Acceleration at Radius 1* through *Acceleration at Radius 6*, control the direction and magnitude of the agent's acceleration at six concentric distance ranges (0–0.6, 0.6–1.1, 1.1–2.3, 2.3–4.6, 4.6–8.0, or >8.0 cm [or, in degrees of visual angle (DVA), approximately 0–0.8°, 0.8–1.4°, 1.4–2.8°, 2.8–5.9°, 5.9–10.1°, or >10.1°]). Positive values indicate acceleration toward (approaching), whereas negative values indicate acceleration away (receding). For example, a positive value of *Acceleration at Radius 2* would mean that the agent in question will accelerate toward the nearest other agent when that agent falls within the second innermost ring. Figure 1 shows a schematic of the six concentric distance ranges in which the six parameters respectively operate, embedded around an agent within a snapshot from a trial of Experiment 1.

These six parameters, here denoted $\mathbf{P} = P_1 \ldots P_6$, collectively control the agent's behavior relative to other agents, and in this sense fully characterize its apparent "personality"—that is, its repertoire of behaviors relative to other agents. A pool of 12 random agents was created by randomly selecting six parameters, each from the continuous interval [−1, 1], for each of 12 agents.

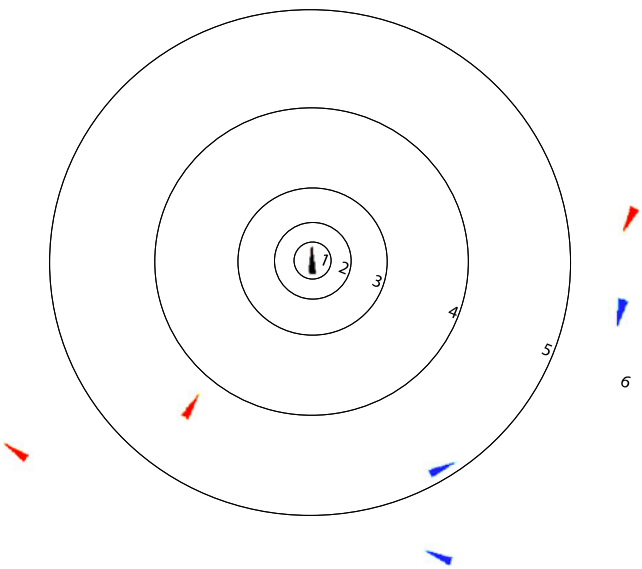



**Fig. 1** Screenshot from Experiment 1 (black background inverted to white), with black circles and numbers superimposed onto the scene to help illustrate the programming scheme for the automata. The center agent in this scene accelerates toward or away from the nearest other agent in the scene. The direction and magnitude of this acceleration depends on the distance to this nearest other agent, with possible distances divided into six zones illustrated above. Analysis of correlations between the six parameters controlling each agent's behavior at these respective other-agent distances and the empirically determined MDS dimensions revealed that Zone 5 was the most psychologically relevant

For example, Agent 1 was assigned parameter vector $\mathbf{P} = [-0.56, 0.77, 0.80, -0.30, -0.82, 0.12]$ At each iteration of the simulation, the respective $[x, y]$ locations of this agent and the nearest other agent in the environment, $L_{\text{self}}$ and $L_{\text{nearest}}$, and the Euclidean distance to this nearest agent, $D$, was computed. The agent's acceleration $A_{\text{self}}$, itself an $[x, y]$ vector, was then set by the following rule: if $D$ fell within distance interval $i$,

$$A_{\text{self}} = P_i \frac{C(L_{\text{nearest}} - L_{\text{self}})}{D}.$$

Here, $C$ is a constant adjusted for each computer on which the simulation was run so that the overall measured average speed of agents moving about the simulation environment would be about the desired rate of 3 cm/s (3.8 DVA/s). The manipulation of acceleration allows for changes in an agent's direction and speed to vary in abruptness.

Under this programming scheme, one example agent might consistently accelerate away from other agents; another might consistently approach. Another—for example, the agent with $P = [-0.82, -0.80, 0.20, 0.44, 0.85, 0.55]$, with negative values in inner radii but positive in outer ones—would approach from afar but then veer away as it got near. Depending on how any two nearby agents are programmed, their interaction might resemble chasing/fleeing, one pushing the other, or even one circling the other. Such rich interactions emerge from our simple scheme, without having to be explicitly programmed into the code.

## Experiment 1

### Method

*Participants* Eight students between the ages of 18 and 24 participated in an approximately 1-hr experimental session in exchange for course credit.

*Stimuli* Scenes were presented to participants on a 1,440 × 900 LED display, on a 15-in. MacBook Pro laptop with a 2.2-GHz dual-core processor. The simulation environment itself was a rectangular window measuring 33.0 × 16.5 cm (horizontally subtending approximately 40 DVA), and the viewing distance was approximately 45 cm. The triangular agents had bases of .23 cm and heights of .92 cm (subtending approximately .25 × 1 DVA).

*Procedure* In each 15-s scene, the participant observed seven agents interacting: three red, three blue, and one white. The reds behaved according to the same parameters as the other reds, the blues according to a different set of parameters, and the lone white according to a third set of parameters. The agents were drawn from a larger 12 agent pool; thus, there were 220 possible triads of these 12 agents.[3] For each scene, one of these 220 triads was selected at random, and each of the three programs in the selected triad was randomly assigned to red, blue, or white. Each participant saw 220 such scenes, exhausting the possible triads.

Participants were openly encouraged to construe the triangular agents as animate. At the end of each scene, they were asked, "Is the white agent behaving more like a red, or more like a blue?" They answered by clicking on a button in a dialog box.

We constructed a 12 × 12 symmetric distance matrix for each participant, to be fed into the individual differences multidimensional scaling (MDS) algorithm (INDSCAL/ALSCAL; Takane, Young, & de Leeuw, 1977). Within this matrix, an agent was assigned a distance of 0 from itself. Since two different agents appeared in the same trial of an experimental session 10 times, the distance in this matrix between any two agents was initially set at 11.

If the participant chose "red," then the agent whose programming was used for the red agents in this trial was made to be closer together (more similar) in this distance matrix with that of the white agent, and likewise if the participant chose "blue." That is, the distance between these two agents in the matrix was reduced by 1. Previous studies have used similar methodologies to gauge participant similarity ratings of visual stimuli (e.g., Kahana & Bennett, 1994; Pantelis, van Vugt, Sekuler, Wilson, & Kahana, 2008).

## Results and discussion

We derived a two-dimensional MDS solution in order to visualize the space of agents that participants (on average) perceived (see Fig. 2). For this amount of points in the space, the INDSCAL algorithm allows for fits of two to five dimensions. Deriving higher dimension solutions will always result in better fits to the experimental data. However, a higher number of dimensions would be even more difficult to interpret than the two condensed dimensions we present, and even a five-dimensional fit would probably be a condensed version of the true amount of psychologically relevant dimensions in this agent space (which could hypothetically be even higher than the total number of agents in our sample). Additionally, a scree plot of various MDS fits did not produce a clear "elbow" favoring one particular number of dimensions over another (Fig. 3).

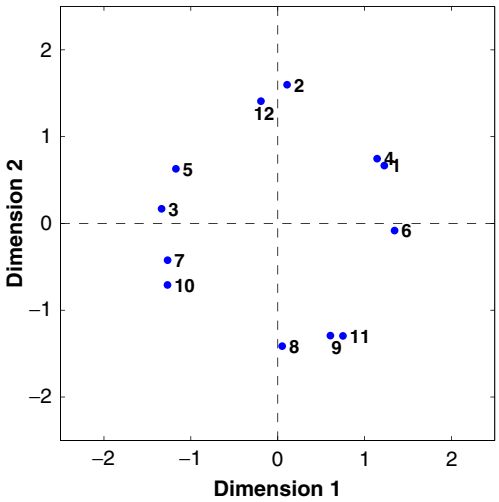



**Fig. 2** The two-dimensional MDS solution for the 12 agents, fitting data from Experiment 1

A 2-D solution allows for the easiest visualization of the interagent distances, an important motivation for using the MDS analysis in the first place. If interesting structure emerged only in higher dimensional fits for these data, this might have justified using these MDS solutions. However, we actually found the clearest and most interesting structure within a 2-D fit.

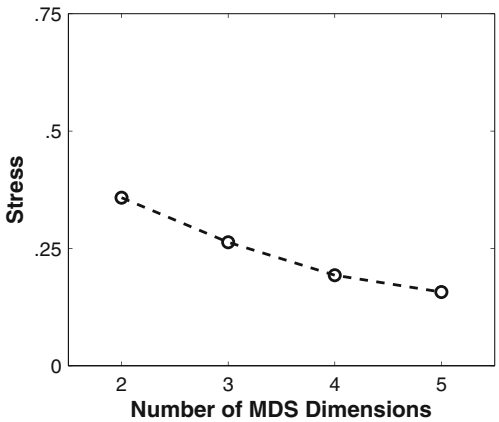



**Fig. 3** The scree plot for MDS fits of two to five dimensions, for pooled data from Experiments 1 and 2

[3] Strictly speaking, because the status of the white agent in each trial is special and, as a result, during a given trial the participant cannot respond that he or she actually believes the blue and red agents to be most alike, 660 possible arrangements actually exist. Rather than show all 660 possibilities, we randomized the procedure so that no agent type would be more or less likely to be "white" during a trial. Nevertheless, doing this presents a source of noise in the data, and we altered the procedure in Experiment 2 to address this issue.

The most striking aspect of the space is its ring-like structure, similar to what one would observe in a 2-D MDS plot of the color wheel (see Shepard, 1980). The significance of this ring structure was not immediately clear, in part because MDS dimensions are in general not self-explanatory but rather pull out subjectively primitive parameters. For clarity's sake, it should be noted that despite the superficial resemblance between the MDS ring structure and the circular radii governing agent behavior (as in Fig. 1), this is coincidental, since there is no relationship between them. Experiment 3, presented below, was designed to help clarify the nature of the parameter exhibited in this ring.

The goal of the present experiments was not, per se, to see how the somewhat arbitrary parameters with which we programmed the agents mapped to participants' percepts of the agents' behaviors. Rather, we had aimed to infer the structure of the perceptual space itself. Nonetheless, relating these parameters to the MDS dimensions was a useful step in understanding the 2-D MDS space.

Participants' perception of the agents' behaviors arises from some complex interaction of its underlying programming and the chaotic interaction with other agents that arises during each unique simulation. This contributed to there being many individual differences between participants' results; few participants' distance matrices showed obvious correlation. However, one of the six parameters with which we programmed each agent—*Acceleration at Radius 5*—was indeed strongly correlated with one of the MDS coordinates (see Table 1). This parameter controlled how an agent behaved when the nearest other agent was between 4.6 to 8.0 cm (5.9 to 10.1 DVA, or about five to nine times the length of an agent's body) away from it. This finding is addressed further in the Experiment 2 discussion.

## Experiment 2

In Experiment 2, we adjusted the basic methodology of Experiment 1 in hopes of reducing the amount of noise in the data and making the experiment more engaging for participants. The most significant change was to allow the participant to control one of the agents in each simulation via the mouse. The chance to interact with the simulated agents would, we expected, allow the participant to glean more information about the other agents' behaviors during the short 15-s display time and thus promote stronger impressions of the agents' "personalities" than was possible in Experiment 1.

**Table 1** Correlations ($r$[10]) between programmed parameters (rows) and MDS dimensions (columns), in Experiment 1. Bold font represents $p < .01$

| *P* | MDS Dimension 1 | MDS Dimension 2 |
|---|---|---|
| Acceleration at Radius 1 | −.070 | .384 |
| Acceleration at Radius 2 | −.275 | −.074 |
| Acceleration at Radius 3 | .527 | .199 |
| Acceleration at Radius 4 | .411 | −.375 |
| Acceleration at Radius 5 | **−.801** | .093 |
| Acceleration at Radius 6 | .459 | .197 |

### Method

*Participants* Seven students between the ages of 18 and 23 participated in an approximately 1-hr experimental session in exchange for course credit.

*Stimuli* We presented scenes to participants on an eMac with a 17-in. (16 in. viewable) monitor and an 1,152 × 864 display. The monitor refresh rate was 80 Hz, and the computer had a 1.25-GHz processor. The simulation environment itself was a rectangular window measuring 25.4 × 16.5 cm (horizontally subtending approximately 31.5 DVA), and the viewing distance was approximately 45 cm.

Experiment 2's scenes were populated with triangular agents of the same size and were programmed under the same scheme as in Experiment 1. We used the same pool of 12 agents from Experiment 1, each of which had been created with six randomized parameters within the programming scheme.

Additionally, the participant controlled one agent with the mouse—a white circular agent .46 cm (.6 DVA) in diameter. The automatic agents were programmed to react to the participant-controlled agent in the same manner as they would to any other triangular agent in the simulation.

*Procedure* In each 15-s scene, the participant observed six agents and controlled one agent. Two agents were red, two were green, two were blue, and the participant-controlled agent was white. The reds would behave according to the same parameters as the other reds, the greens according to a different set of parameters, and the blues according to a third set of parameters. The agents were drawn from a larger 12-agent pool; thus, there were 220 possible triads of these 12-agent programs. For each scene, one of these 220 triads was selected at random; then, each of the three programs in the selected triad was randomly assigned to red, green, or blue. Each participant saw 220 such scenes, exhausting the possible triads.

Participants were openly encouraged to construe the triangular agents as animate, and they were instructed that how agents of a certain color behaved during one trial would have nothing to do with how they behaved in subsequent trials. At the end of each scene, they were asked

to determine which color of agent behaved *least* like the other two—that is, which was most different: red, green, or blue? They responded by key press, at which point the next trial began.

As in Experiment 1, we constructed a 12 × 12 symmetric distance matrix for each participant, to be fed into the individual differences MDS algorithm. For each trial, the two unchosen agents in the odd-one-out procedure were made more similar within this distance matrix.

## Results and discussion

Once again, we derived a 2-D MDS solution in order to visualize the space of agents that participants (on average) perceived, and we once again observed a ring-like structure in the space (Fig. 4).

The MDS solutions for the two experiments—processed representations of participants' raw similarity matrices—were correlated with each other. Dimension 1 of Experiment 1's MDS was strongly correlated with Dimension 2 of Experiment 2's MDS [$r(10) = .713$, $p < .01$]. Dimension 2 of Experiment 1's MDS was weakly (and negatively) correlated with Dimension 1 of Experiment 2's MDS [$r(10) = -.551$, $p < .063$]. (The direction of these correlations is arbitrary and unimportant, but helpful in relating the 2D MDS spaces presented in Figures 2 and 4.) These correlations provide some assurance of the robustness and psychological reality of the subjective mental spaces that we have uncovered.

As is shown in Table 2, Dimension 2 in Experiment 2 correlated significantly with parameter 5 of the agents' programming—that is, *Acceleration at Radius 5*. This corroborates one of the results of Experiment 1, where Dimension 1 had been correlated with this same parameter.

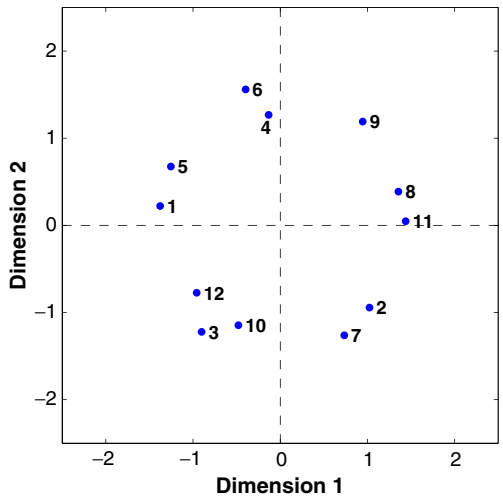



**Fig. 4** The two-dimensional MDS solution for the 12 agents, fitting data from Experiment 2

**Table 2** Correlations ($r$[10]) between programmed parameters (rows) and MDS dimensions (columns), in Experiment 2. Bold font represents $p < .05$

| *P* | MDS Dimension 1 | MDS Dimension 2 |
|---|---|---|
| Acceleration at Radius 1 | −.129 | −.147 |
| Acceleration at Radius 2 | −.529 | −.050 |
| Acceleration at Radius 3 | .096 | .195 |
| Acceleration at Radius 4 | .548 | .200 |
| Acceleration at Radius 5 | −.310 | **−.619** |
| Acceleration at Radius 6 | .249 | .233 |

Apparently, how an automaton reacts within a certain critical distance range—about five to nine times the length of an agent's body—played a psychologically important role in determining participants' judgments about its behavior. For an intriguing—though admittedly very rough—comparison, in Heider and Simmel's famous vignette, the two triangles faced off at a range of about five body lengths, while the circle watched anxiously from a distance of about nine body lengths.

We wondered whether the prominence of this parameter in participants' judgments was actually an artifact of the frequency with which interactions at this distance actually occurred in the displays. But the data do not bear this out. Because the entire displays were recorded (10 frames/s), we could assess the proportion of the time the interagent distance between any automaton and its nearest other agent was within each of the six intervals corresponding to the six underlying programmed parameters. The two most common distances between an automaton and its nearest other agent during a simulation were 0–.6 cm (0–.6 agent lengths) and 2.3–4.6 cm (2.5–5 agent lengths). The distance 4.6–8.0 cm (5–9 agent lengths) was only the fourth most common interagent distance. We conclude that the pivotal role of this interagent distance is not an artifact, but rather reflects a genuine cognitive focus on behavioral interactions at this distance.

In light of this finding, we examined reported data from the previous study of Barrett et al. (2005), in which participants acted out various intentional behaviors (playing, chasing, etc.) with "arrowhead"-shaped agents under their control. As in our study (but unlike in Heider & Simmel, 1944), the agents used in Barrett et al. (2005) were of uniform size and shape, allowing for an "apples-to-apples" comparison with our study by using "agent lengths" as our distance metric. The authors reported that, within the virtual scenes that participants dynamically created, interagent distance across the various action types was about 300 pixels, on average—approximately six times the length of an agent. Such an interagent distance falls comfortably within the range we found to be most psychologically salient in Experiments 1 and 2.

## Experiment 3

The results of the first two experiments were qualitatively similar, and we therefore choose to pool data from all 15 participants for the following analysis and discussion. The 2-D MDS solution for these pooled participants reveals an even cleaner ring structure (see Fig. 5). But what does it mean as we travel around this ring?

In the combined MDS, Dimension 1 is connected to how an agent behaves when the closest other agent is between five to nine agent lengths away (i.e., *Acceleration at Radius 5*). Agents that accelerate away from the nearest other agent at this distance tend to be low on MDS Dimension 1; agents that accelerate toward this nearest other agent at this distance tend to be high on MDS Dimension 2.

The meaning behind Dimension 2 is somewhat less straightforward. When analyzing the data from Experiments 1 and 2 separately, Dimension 2 was uncorrelated with any of the agents' programmed parameters. However, when pooling data from the experiments, we found that Dimension 2 was indeed correlated with *Acceleration at Radius 2*—the parameter that governs how an agent behaves at .6–1.1 agent lengths from its nearest other agent [$r(10) = -.614$, $p < .05$].

We could stop here, using these correlational results as interpretations for the two MDS dimensions. But in Experiment 3, we delve further in an attempt to make better sense of the MDS solution we have produced, in a manner that attaches intuitive semantics to our 2-D MDS space.

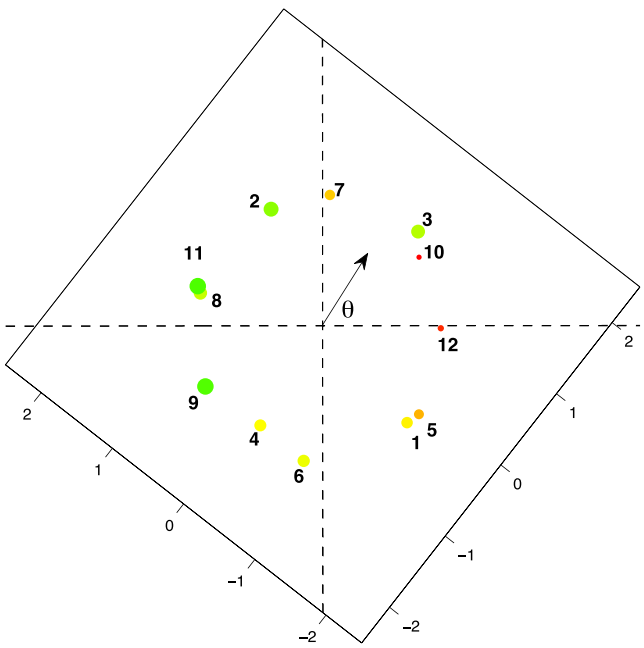



**Fig. 5** Pooled two-dimensional MDS from Experiments 1 and 2. The 2-D space is rotated about the origin so that the cosine of the agent angle (relative to the horizontal) best predicts how participants rated the agents along the "hostility" versus "friendliness" dimension. *Smaller* (redder) *circles* represent agents rated as more hostile, and *larger* (greener) *circles* represent agents rated as friendlier

The MDS procedure creates orthogonal dimensions, and as a result, there is no linear correlation between MDS Dimensions 1 and 2. However, a consequence of the 2-D ring structure means that knowledge of where an agent lies on one MDS dimension almost precisely predicts the *absolute value* along the other dimension. That is, the two MDS dimensions carry some mutual information.

Another consequence of the ring structure is that the *x*- and *y*-axes (representing MDS Dimensions 1 and 2) may indeed be arbitrary. If one preserves the orthogonality of these axes but rotates them about the origin, any such rotation will be a rigid transformation, preserving the ring structure and all of the distances among the points representing the 12 agents. The variance of the distribution of these points along the two respective axes will also be preserved. Thus, any pair of orthogonal axes placed in this 2-D space will work as well as any other pair. The question is: Which rotational orientation of these axes best enhances the interpretability of the MDS?

The methodology we employed, though innovative, resulted in participants making judgments about agent behavior from noisy data. Furthermore, the MDS space we produced is of very low resolution—only 12 data points—and is averaged across all participants. Thus, it is reasonable to expect that this MDS "ring" reflects a fairly coarse yet robust qualitative assessment made by participants about the agents—a simple perceptual dimension that cuts through this methodological noise. We propose that "hostility" versus "friendliness" might be this dimension, emerging from the interaction of the two MDS dimensions we inferred from Experiments 1 and 2. We turn to further psychophysics to provide evidence for this hypothesis.

### Method

*Participants* Seven students between the ages of 18 and 24 participated in a session that was approximately .5 hr, in exchange for course credit.

*Stimuli and procedure* We presented scenes to participants under the same viewing conditions as in Experiment 1. We again populated the simulations with the pool of 12 agents employed in Experiments 1 and 2. During each trial, the participant watched seven agents interacting for 15 s. Six of the agents were colored red and behaved under programs randomly selected from the pool of 12. The seventh, critical agent was colored blue, and the participant was instructed to attend to it. At the end of each trial, the participant was asked, "On a scale of 1–5, 1 being *most hostile*, and 5 being *most friendly*, how do you rate the blue agent?" The participant indicated his or her response on the keyboard.

Each of the 12 agents in the pool was assigned the blue color for eight of the session's trials, for a total of 96 trials presented in random order.

## Results and discussion

We first normalized each participant's responses, then calculated each participant's mean normalized response for each of the 12 agents observed over the experimental session. Then, averaging across participants, we were able to get a sense of how friendly versus hostile participants perceived each of the 12 autonomous agents. Figure 5 shows, on a gradient from small to large (red to green), what these perceptions were. The most hostile agents seem to be those that were high on MDS Dimension 1 and low on Dimension 2, whereas the friendlier agents tended to be low on Dimension 1 and high on Dimension 2. Agents low on both dimensions were quite neutral. Figure 5 shows the space in a rotated coordinate frame so that the horizontal dimension optimally reflects the friendliness versus hostility dimension. (All of the interagent distances and relationships have been preserved; only the "ring" has been rotated.) In the rotated space, the projection of each agent's position onto the horizontal (i.e., the cosine of its angle relative to the horizontal) reflects its position along the friendly/hostile dimension. We regressed the participants' mean friendliness rating against this variable and found a close fit ($r(10) = -.768$, $p < .01$, Fig. 6). These data corroborate our hypothesis that the ring variable essentially reflects the degree of perceived friendliness or hostility each agent exhibited.

The consideration of this friendliness dimension helps to make sense of the agent space in several additional ways. For one, perceived friendliness is linearly correlated with both MDS Dimension 2 [$r(10) = .598$, $p < .05$] and *Acceleration at Radius 2* [$r(10) = -.665$, $p < .05$] (see Fig. 7). There also appears to be a nonlinear, yet straightforward, relationship between perceived friendliness and *Acceleration at Radius 1*—the programmed parameter that governs the behavior of agents at the very closest distances (see Fig. 8).

Furthermore, as a byproduct of the rotation of the MDS space that we performed in the just-considered analysis, we produce two new MDS dimensions (the new *x*- and *y*-axes of Fig. 5). Whereas the old MDS dimensions could be connected to only one (or possibly two) of the programmed parameters, one of the newly adjusted MDS dimensions indeed correlates linearly with three of them (see Table 3). Though it may not be surprising that rotating the space allowed for the discovery of these additional statistical relationships, we note that this finding cannot be compared directly to the results of Experiments 1 and 2 because Experiment 3 employs a different dependent measure.

Scrutiny of this hypothesized "hostility" versus "friendliness" dimension unlocks many additional details about the relationship between the agents' underlying programming and the way participants perceived them, suggesting that this is indeed an important subjective dimension organizing the perceptual space. With only 12 data points, inferred from noisy underlying data, there were obvious limitations to the power of the correlational studies performed in Experiments 1 and 2. However, the perceptual effects of *Accelerations at Radii 1 and 4* were revealed not necessarily by the additional statistical power of Experiment 3, but by the fact that this experiment enabled us to rotate the perceptual "ring" (constructed from pooled data from Experiments 1 and 2) in a way that made the *x*- and *y*-axes suddenly more interpretable. That is, the *x*-axis in the original MDS orientation was not correlated with *Accelerations*

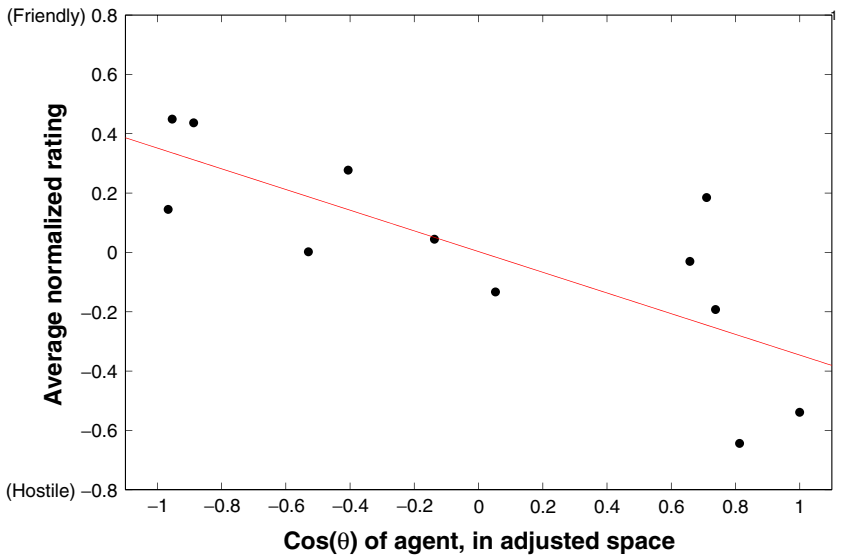



**Fig. 6** Cosine of the agent's angle in MDS space (see Fig. 5), plotted against how participants, on average, rated them (from *hostile* to *friendly*). The *line* indicates best linear fit

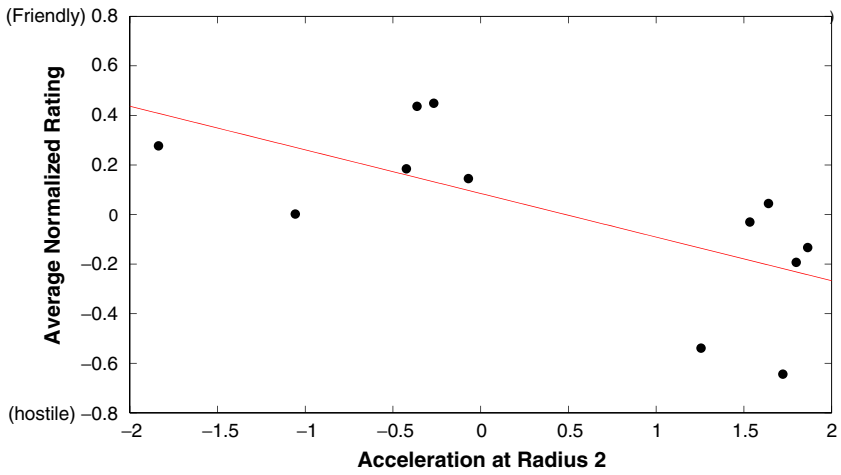



**Fig. 7** Agent values along programmed Parameter #2, plotted against how participants rated them (from *hostile* to *friendly*). The *line* indicates best linear fit

*at Radii 1 and 4*, but with this new rotation—done without a priori consideration as to how the rotation would affect the correlations between this axis and the programmed parameters, but rather to align it with the subjective dimension of "friendliness" versus "hostility"—these relationships suddenly emerge. This is perhaps the primary contribution of Experiment 3, which—more than helping to attach an intuitive semantic label to the MDS, which it also does—sets the ring in an immediately more interpretable orientation. We now observe that it is not merely *Acceleration at Radius 5* that contributes to participants' percept of agent personality, but additionally at *Radii 1, 2, and 4.*

## Further discussion of the "ring of agents"

Although the MDS solutions in this experiment are two-dimensional, the ring structure that emerges is not truly 2-D, but more accurately a 1-D manifold residing in a 2-D space. That is to say that locally, there is only 1 degree of freedom with which one can move along the ring: the angle, $\theta$, as one circles the origin.

The 1-D nature of this manifold reveals that the coarse "agent space" we have uncovered is probably best described by one complex dimension—that which changes as one travels along the 1-D manifold in either direction. Hostility/friendliness is a key component of this perceptual dimension, since "hostile" and "friendly" agents represent two poles on the manifold (i.e., the small and large [red and green] "poles" of Fig. 5). But there are segments along the ring (agents that are medium-sized [yellow] in Fig. 5, but nonetheless were judged to be maximally different by participants) for which we do not presently understand why participants judged them differently. They are neutral in terms of hostility/friendliness.

On our 1-D manifold, there are no "endpoints," as one would find on a 1-D number line—just as in the case of the surface of the globe (a 2-D manifold in 3-D space), where there is no "edge of the world." However, there may be modes or poles. On the globe, these poles lie on the two points where

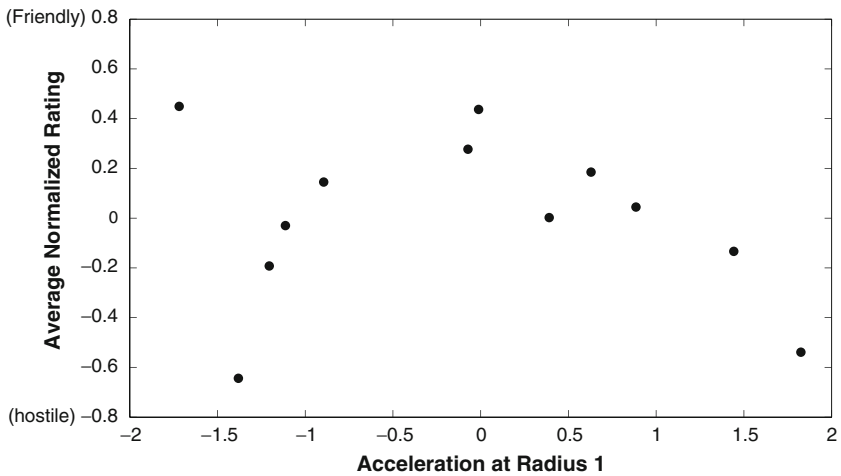



**Fig. 8** Agent values along programmed Parameter #1, plotted against how participants rated them (from *hostile* to *friendly*). The data trend in an inverted "U"

**Table 3** Correlations (*r*[10]) between programmed parameters (rows) and MDS dimensions in *rotated* space (columns). Bold font represents $p < .05$

| *P* | New MDS Dimension 1 | New MDS Dimension 2 |
|---|---|---|
| Acceleration at Radius 1 | .189 | .083 |
| Acceleration at Radius 2 | **.593** | −.232 |
| Acceleration at Radius 3 | −.215 | −.221 |
| Acceleration at Radius 4 | **−.617** | .048 |
| Acceleration at Radius 5 | **.597** | .442 |
| Acceleration at Radius 6 | −.350 | −.155 |

the rotational axis of the earth intersects with the manifold surface. In our agent space, two poles that we have discovered seem to represent hostile and friendly agents.

To summarize, we have identified two "poles" on our 1-D agent manifold—"hostile" and "friendly"—that can be considered the endpoints along a continuous dimension (hostility/friendliness). But it is unclear how many other salient poles the ring may have. Furthermore, these unknown poles need *not* reside at the endpoints of a continuous dimension.

## General discussion and conclusions

The present experiments were designed to probe the underlying structure of the agent space perceived by participants as they watched autonomously programmed agents interacting in dynamic scenes. In Experiments 1 and 2, the MDS approach succeeded in revealing certain aspects of this perceptual space: a ring-like structure, which—in Experiment 3—we attempted to connect to a dimension of perceived hostility versus friendliness in the agents.

The results of Experiments 1 and 2 revealed that one of the low-level parameters controlling the behaviors of the agents strongly contributed to this more abstract percept: that which controlled interagent reactive behavior at one critical distance. We concluded that this might have reflected one perceptually critical interagent zone upon which participants based their interpretations of the agents' intentional behavior. The results of Experiment 3, however, revealed that, indeed, several of the underlying agent parameters could be connected to the way participants organized the agent space.

If our triangular agents are indeed interpreted as "behaving"—a cognitive tendency that has been well established by a host of previous studies dating to Heider and Simmel (1944)—then the parameters with which they are programmed represent "behavioral dispositions" in a very literal sense. Our use of the term *personality* is obviously metaphorical, because the agents are not persons, and their behavior represents only a tiny microcosm of the behaviors that real persons manifest. However, within the admittedly limited range of actions of which these agents are capable, these parameters do literally modulate each agent's behavioral tendencies and in that sense determine its behavioral character.

We conclude that "hostility" versus "friendliness," or something akin to this dichotomy, produces an especially salient partition in participants' perceptual space of the agents' "personalities." In other words, after first surmising that an object in the world has intentions (i.e., is animate), a next step for the cognitive machinery might be an attempt to guess whether these intentions are good or bad, given the current perspective of the perceiver. That this would be an important perceptual distinction in this task agrees well with experimental findings from other domains. Todorov, Said, Engell, and Oosterhof (2008) found "valence" or "trustworthiness" to be a principal dimension along which participants evaluate faces.[4] In the domain of decision making, Burnham, McCabe, and Smith (2000) hypothesized that humans employ a "friend-or-foe" mental mechanism when engaging in games against other agents. Like Todorov et al., these authors relate such a dimension to the evaluation of another person's trustworthiness. Making such a classification allows for a player to tailor his strategy to the level of expected cooperation from an opponent (Oberholzer-Gee,Waldfogel & White, 2010).

"Friendliness," "valence," and "trustworthiness" are not, of course, equivalent concepts, as has been attested by an enormous amount of literature in social psychology. However, our concern is not with the full characterization of personality, but with the way behavior is reflexively classified on the basis of visual analysis of motion—and in this limited context, these complex attributes may well be conflated. Fiske, Cuddy, and Glick (2007) termed this a *friend-or-foe* judgment, or an evaluation of another agent along the broadly-defined dimension of "warmth." Such a judgment might well inform the simple but highly consequential binary social decision of whether to avoid another agent or approach it, although "snap" judgments such as this may not capture the full nuance of a more thorough and contextually enriched "mindreading" process. Whether or not the perceptual effects we present are encapsulated and cognitively impenetrable to higher level mindreading—in the manner articulated by Pylyshyn (1999) vis-à-vis the modularity of early vision—remains a provocative but open question, lying at the heart of

[4] The other principal dimension was slightly harder to define and was labeled "power/dominance" by the authors. Such a label does not appear to be relevant to our second MDS dimension.

fundamental debates regarding the boundaries between perception and cognition.

The present work represents one step in what we hope is a fruitful new direction. Programming agents autonomously, and asking how participants' interpretations of these agents' behavior relate to the actual programs they are carrying out, allows one to pursue a true "psychophysics of intention," in which we explore the relationship between the perceived intention and the "actual" intention present in the agent's autonomous program. In future experiments, by employing displays of potentially far more complex behavioral interactions, we hope to uncover correspondingly more complex structures in the intentionality percept.

**Author Note** Portions of this research were presented at the 2010 meeting of the Cognitive Science Society. This work was funded in part by the National Institutes of Health Grant NIH EY15888, the NSF IGERT program in Perceptual Science Grant NSF DGE 0549115, and a grant from the Hellenic University Club of New York. Special thanks to the Robert J. Glushko foundation.

## References

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

Baker, C. L., Tenenbaum, J. B., & Saxe, R. R. (2006). Bayesian models of human action understanding. In B. Scholkopf, J. Platt, & T. Hofmann, Advances in Neural Information Processing Systems 19 (pp. 99-106). Cambridge, MA: MIT Press.

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.

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

Burnham, T., McCabe, K., & Smith, V. L. (2000). Friend-or-foe intentionality priming in an extensive form trust game. *Journal of Economic Behavior & Organization, 43,* 57–73.

Dittrich, W. H., & Lea, S. E. G. (1994). Visual perception of intentional motion. *Perception, 23,* 253–268.

Feldman, J., & Tremoulet, P. D. (2008). *The attribution of mental architecture from motion: Towards a computational theory (TR-87)*. Piscataway: Rutgers University Center for Cognitive Science.

Fiske, S. T., Cuddy, A. J. C., & Glick, P. (2007). Universal dimensions of social cognition: Warmth and competence. *Trends in Cognitive Science, 11,* 77–83.

Gao, T., McCarthy, G., & Scholl, B. J. (2010). The wolfpack effect: Perception of animacy irresistably influences interactive behavior. *Psychological Science. 21(12)*, 1845–1853. doi:10.1177/0956797610388814.

Gao, T., Newman, G. E., & Scholl, B. J. (2009). The psychophysics of chasing: A case study in the perception of animacy. *Cognitive Psychology, 59,* 154–179.

Gao, T., & Scholl, B. J. (2011). Chasing vs. stalking: Interrupting the perception of animacy. *Journal of Experimental Psychology: Human Perception & Performance, 37(3)*, 669–684. doi:10.1037/a0020735.

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: 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.

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 mentalisic agents in infancy. *Trends in Cognitive Development, 16,* 637–656.

Kahana, M. J., & Bennett, P. J. (1994). Classification and perceived similarity of compound gratings that differ in relative spatial phase. *Perception & Psychophysics, 55,* 642–656.

Keil, F. C. (1994). The birth and nurturance of concepts by domains: The origins of concepts of living things. In L. Hirschfeld & S. A. Gelman (Eds.), *Mapping the mind: Domain specificity in cognition and culture*. New York: Cambridge University Press.

Klein, J. (2002). Breve: A 3 d simulation environment for the simulation of decentralized systems and artificial life. In R. K. Standish, M. A. Bedau, & H. A. Abbass (Eds.), *Proceedings of Artificial Life VIII, the 8th International Conference on the Simulation and Synthesis of Living Systems* (pp. 329–336). Cambridge: The MIT Press.

Leslie, A. M. (1984). Infant perception of a manual pick-up event. *British Journal of Developmental Psychology, 2,* 19–32.

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

Oberholzer-Gee, F., Waldfogel, J., & White, M. (2010). Friend or foe? Cooperation and learning in high-stakes games. *Review of Economics and Statistics, 92,* 179–187.

Pantelis, P. C., van Vugt, M. K., Sekuler, R., Wilson, H. R., & Kahana, M. J. (2008). Why are some people's names easier to learn than others? The effects of face similarity on memory for face–name associations. *Memory & Cognition, 36,* 1182–1195.

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

Pylyshyn, Z. (1999). Is vision continuous with cognition?: The case for cognitive impenetrability of visual perception. *Behavioral and Brain Sciences, 22,* 341–365.

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

Shepard, R. N. (1980). Multidimensional scaling, tree-fitting, and clustering. *Science, 210,* 390–398.

Takane, Y., Young, F. W., & de Leeuw, J. (1977). Nonmetric individual differences multidimensional scaling: An alternating least squares method with optimal scaling features. *Psychometrika, 42,* 7–67.

Terzopoulos, D., Tu, X., & Grzeszczuk, R. (1994). Artificial fishes: Autonomous locomotion, perception, behavior, and learning in a simulated physical world. *Artificial Life, 1,* 327–351.

Todorov, A., Said, C. P., Engell, A. D., & Oosterhof, N. N. (2008). Understanding evaluation of faces on social dimensions. *Trends in Cognitive Sciences, 12,* 455–460.

Torgerson, W. S. (1958). *Theory and methods of scaling*. New York: Wiley.

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.

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: Addison-Wesley.

Zacks, J. M. (2004). Using movement and intentions to understand simple events. *Cognitive Science, 28,* 979–1008.

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