The Attribution of Mental Architecture from Motion: Towards a Computational Theory

Jacob Feldman and Patrice D. Tremoulet
Dept. of Psychology, Center for Cognitive Science
Rutgers University - New Brunswick, Piscataway, NJ 08854

Recently there has been great interest in how observers attribute mental properties—beliefs, intentions, goals, cognitive capacities, and so forth—to other agents in our environment. In many cases, such attributions are based solely on patterns of motion. For example human observers tend to interpret certain entities as *animate* (living), and others not, based solely on their motion trajectories, and whether they seem to suggest intentional or goal-driven behavior. In this paper we consider such attributions from a computational point of view, and we ask how information derived solely from observable behaviors might formally support the attribution of a particular *computational architecture* to a target agent. We develop a mathematical theory of the inference of such an architecture, which we call the *attributed mental architecture* (AMA). Within the theory, particular mental faculties, such as a perceptual capacity or the possession of a goal, can be characterized mathematically in terms of formal properties on the attributed mental architecture. We give theorems concerning minimal conditions for the inference of particular types of mental faculties, including *perceptual capacities, cognitive capacities*, and *intentionality*.

Keywords: animacy, computational architecture, goal attribution, intentionality, mental architecture, mindreading,

The meaning behind the motion

As any fan of Disney movies knows, all it takes is motion—at least, the right kind of motion—to bring things to life. When objects move in certain ways, they can seem to have goals, intentions, personalities: in short, to have *minds.* This point was brought home particularly vividly by the short film created in 1944 by the psychologists Heider and Simmel, which consists simply of three shapes (two triangles and a circle) moving about for a few minutes within a simple static environment. The three shapes perform an elaborate drama, whose plot—featuring acts of bullying, protecting, chasing, and escaping—is communicated entirely by the motion trajectories of the three rigid shapes. Heider and Simmel's subjects, asked to simply report what they saw, ascribed personalities, relationships, and emotions to the ensemble of shapes with remarkable unanimity. Needless to say, a

We are grateful to Rochel Gelman and Whitman Richards for many helpful discussions.

Please direct correspondence to Jacob Feldman, Department of Psychology (Busch Campus), Center for Cognitive Science, Rutgers University - New Brunswick, 152 Frelinghuysen Rd., Piscataway, NJ 08854, or by email at jacob@ruccs.rutgers.edu.

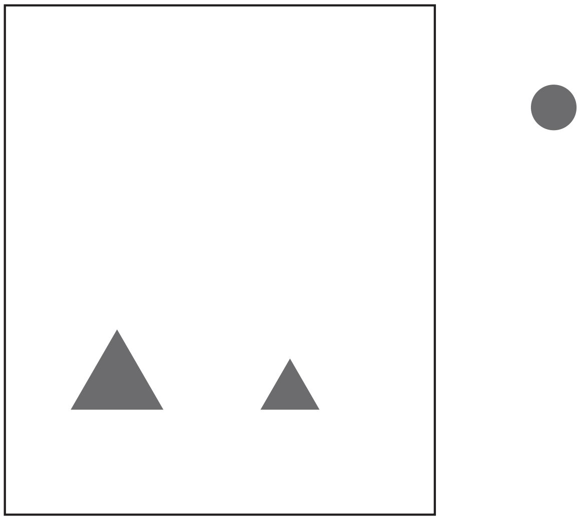

Figure 1. Illustration of a still frame from Heider and Simmel's (1944) film. When the shapes move, they seem to have goals, intentions, emotions, and even personalities. Without the motion, of course, the image conveys little.

static frame from the movie (Fig. 1) communicates none of these things.

In this paper we take up the issue of how observable patterns of behavior—for example, motion trajectories—can be interpreted as signaling particular mental capacities and faculties. Our own interest in this topic began with the perception of *animacy* from motion: how the visual system distinguishes living from non-living motion trajectories. As we summarize below, our studies of this topic led us to the conclusion that the essential quality that makes motion look animate is the impression of *goal-directedness* or *intentionality*. That is, motion looks alive when it appears to have been directed by some internal planning or motivation, rather than being the passive result of some exogenous force.

But terms like *goal-directed* and *intentional* are a bit vague. How do you know if an agent actually has a goal? After all, one cannot peer inside the brains of a viewed target in order to ascertain its mental state, or indeed whether it even *has* mental states. How then can you estimate, using only observable cues, the internal mental architecture of another agent? This is the question we seek to answer in this article.

Antecedents

Attributing mental properties to entities moving about the environment is a particular interest of human infants', who seem innately inclined to understand events around them as arising from mentalistic agents (Dasser, Ulbaek, & Premack, 1989; Gelman & Spelke, 1981; see Johnson, 2000 for a review). The sophistication with which infants seem to understand the underlying elements of mentality has led some authors to hypothesize an innate "naive psychology" (Carey, 1985) or "theory of mind" (Leslie, 1987) underlying their judgments—that is, a built-in sense of how beliefs, desires and goals work, and who is liable to have them (viz., animate agents). Infants' judgments of intentionality rely heavily on motion, and in particular, motion that is apparently contingent on the motions of other entities in the environment; for example infants apparently regard as animate an entity that moves in concert with the infant him- or herself, even when the object doesn't much resemble a human or an animals (Johnson, Booth, & O'Hearn, 2001). Gergely, Nádasdy, Csibra, and Bíró (1995) have suggested that children reserve intentional interpretations for those agents that seem to behave in an apparently *rational* manner—that is, in a way that seems to bring them systematically towards some perceivable goal (see also Csibra, Bíró, Koós, & Gergely, 2003). These two ideas—behavior contingent on the environment, and rational choice of action—are central to our formalization of intentionality attribution below.

The idea of understanding others' minds has also received quite a lot of attention from philosophers under the rubric "mindreading." Debate has centered on two classes of theories: simulation theories, in which we in effect run a simulation of another agent, using as a model resources drawn from our *own* minds (Goldman, 1992); and the "theory theory," in which we draw inferences about others' mental states with reference to a distinct theory of mental function, perhaps taking the form of a "folk psychology" (Stich & Nichols, 1992, 1995). This debate has even entered the realm of neuroscience, where the existence of neurons specially tuned to the recognition of organism's own actions carried out by conspecifics has been taken as evidence in favor of simulation theory (Gallese & Goldman, 1998).

Without entering too deeply into the nuances of this debate, we remark that our formalization given below has elements of both sides. We attribute models to agents, which could be thought of as simulations; but they are not conceived of as drawing from the observer's own planning resources, as they are in simulation theory, but rather from some more impoverished set of causal operations, as in some versions of the theory theory. We skirt over these issues here because (as will become obvious) our main goal lies elsewhere: in formally developing certain ideas that we see as critical to the problem. Exactly how our theory compares with others' accounts is not always clear to us, in part because of the inexplicitness inherent in non-formal theories. To us, the attribution of mental architecture based on observed actions is a problem subject to a mathematical account in the same manner as (and in fact running along similar lines as) the induction of language structure based on observed patterns of speech.

Our attempt to place the inference of mental states and intentionality on a computational footing has several important antecedents, including the classical literature on plan recognition in AI (Schmidt, 1976), which centered in the automatic interpretation of goals and plans from observed actions. Baker, Tenenbaum, and Saxe (2006) have provided one of the only computationally explicit accounts of how plans are estimated from observations, and shares many aspects of its motivation from the development below. Blythe, Todd, and Miller (1999) have proposed a system for classifying animate motions in terms of intent, proposing a heuristic combination of seven specific rules concerning the interaction of two moving targets (e.g., pursuing and evading tend to produce high relatively velocities, courtship relatively low relative velocities; etc.). While these rules are appealing in their concreteness, and probably represent realistic expectations about real terrestrial fauna, we intentionally pose the question at a more abstract level. We are interested in the essential structure of an intentionality attribution in a sense that is deliberately removed from any specific assumptions about the preferences characteristic of extant living species. Certainly there are meaningful regularities to how real animals behave under specific real circumstances, but more detailed consideration of these must necessarily follow, rather than precede, a careful statement of the problem itself. Similarly, we deliberately avoid delving into any algorithmic details, which would inevitably lend a certain ad hoc quality to any treatment; our discussion is strictly at the competence level, and at this stage we seek principled statements about the problem itself.

The interpretation of animacy from motion

In a famous 1959 paper, J. Lettvin and his colleagues (Lettvin, Maturana, McCulloch, & Pitts, 1959) described a neural circuit in the frog's visual system specialized for detecting dark spots moving over a light background—sensitive, that is, not to particular shapes, but rather to a particular kind of object motion. The implication was obvious: the frog's brain was trying to detect flies (i.e., prey). Indeed, it's hard to imagine a more important task for the frog's early visual system to be optimized for. Detecting fast-moving and skittish prey must be done rapidly. A quick-and-dirty classification into prey/non-prey categories, based on whatever cues are most informative, is essential. And motion is *very* informative—inanimate entities in the frog's natural environment rarely dart across the field of view. By the time more elaborate mechanisms for distinguishing food from non-food—such as by shape—could be carried out, the frog might have starved to death.

The bug-detector example is a hyper-simplified case of distinguishing animate from inanimate entities based on motion. In the more sophisticated classifications carried out by human brains, it is not just the presence of motion, but the *characteristics* of motion that signal animacy. We have explored these characteristics in a sequence of experiments involving moving dots in either empty or extremely spare static environments (Tremoulet & Feldman, 2000; Tremoulet, 2000; Tremoulet & Feldman, 2006; see Scholl & Tremoulet, 2000 for a summary of some of this work, and see Dittrich & Lea, 1994 for related work).[1]

[1] There is a large literature on what is referred to as *biological motion*, stemming from the pioneering work of Johansson (1973). In typical biological motion displays, observers see a number of moving points that can be integrated to form an impression of a walking person. Integration of such displays is known to be sensitive to constraints on joint mechanics (Chatterjee, Freyd, & Shiffrar, 1996) and is greater for biological than for mechanical motion (Shiffrar, Lichtey, & Chatterjee, 1997), so our ability to process these displays is probably related to the fact that they depict living things.

In these experiments, we began with displays containing a single rigid object, called the target, moving through an empty field (Fig. 2a). About halfway through the trial, the target would abruptly change speed and direction. We asked our subjects to rate the display for animacy (on a scale of 1–7, from "definitely not alive" to "definitely alive"). We found that the magnitude of the velocity and direction changes had systematic effects on animacy ratings. More acceleration, and larger direction changes, led to a greater sense of animacy. Of course, these are hardly Disney animations; the sense of aliveness in such simplified displays was comparatively subtle. But we were interested in reducing the motion required for "animation" (a word that means, literally, *endowment with life*) down to its bare minimum, so that we could investigate what qualities of motion were crucial in producing it.

Several other manipulations in this first experiment were revealing. First, we included trials where the moving target was a circle (Fig. 2a), and others where the target was an elongated rectangle, i.e. an object with a well-defined orientation. In the oriented case, the object sometimes maintained its alignment with its motion path after changing direction (Fig. 2b). This latter case produced the most animate ratings; it looks more "in control," as if the target *deliberately* altered its orientation in order to maintain path alignment. We also included a case where the rectangle was aligned with the initial motion path, but *didn't* turn after the path change, so that it was misaligned in the second half of the trial (Fig. 2c). These trials looked the least animate of all; even though the object changes velocity, its failure to align its orientation after the path change makes it look passive and non-intentional (some subjects said it looked as if it had been kicked) and thus inanimate.

This last manipulation is particularly important because the results contradicts one of the leading hypotheses about the source of animacy percepts: that it's all about energy sources. Stewart (1982) has proposed that motions are considered animate when they (as she put it) violate Newtonian laws, or (as we'd rather put it), when they fail to conserve visible energy (and see also Bingham, Schmidt, & Rosenblum, 1995 for related ideas). The idea is that animate entities—e.g. animals—have access to hidden energy sources, e.g. mechanical energy expendable by their muscles. Hence they can move on their own steam. Inanimate entities can only be pushed around by outside forces. So to determine animacy, all one would have to look at in a motion path is its energy profile. If energy is only dissipating, e.g. due to friction, then it's inanimate; if energy is increasing, and there is no visible external source of energy to explain the increase (e.g. a collision with another object), then the source must be *internal*, and the target must be animate.

It's a very interesting idea, and it has some additional support from several other lines of reasoning. First, human observers turn out to be very sensitive to whether collisions are Newtonian or not—that is, whether they preserve energy, like a passive collision among billiard balls (Kaiser & Proffitt, 1987). This determination involves a subtle comparison of the speeds and directions of the objects before and after the collision; exactly the sort of thing physics students are taught to do with an elaborate computation. But human observers can estimate by eye, apparently with great precision, whether the collision exactly obeyed the energy-conservation equations governing passive collisions. If it doesn't, the collision looks "wrong"—or, at least, wrong as a *passive* collision.

At the same time, the actual computational problem of determining the sources of motion energy when multiple objects move in concert is a very tricky one. When a hand and a bottle move together, which one is the move-er, and which one the mov-ee? This is surprisingly tricky—for example it's completely ambiguous if the hand and bottle stay together for the whole moving sequence. But Mann, Jepson, and Siskind (1997) have shown that you can solve this problem computationally by assigning hypothetical motion-energy-sources (which they call *body-motors*) to the entities in the scene as parsimoniously as possible. If Stewart's hypothesis is correct, then you can just declare animate all and only those entities that have body-motors, given the simplest possible assignment

Yet the emphasis in this literature in not on the *classification* of motions as animate or intentional, but on the integration of multiple motions into a coherent but non-rigid moving form—a process that is at least possible, though more difficult, when the form is an artifact and the motion mechanical. By contrast the displays in our experiment usually contain only a single moving element, so there is no question of integrating multiple motions. Rather our interest is on the qualities of the single unambiguous motion trajectory that influence whether it is perceived as a living thing.

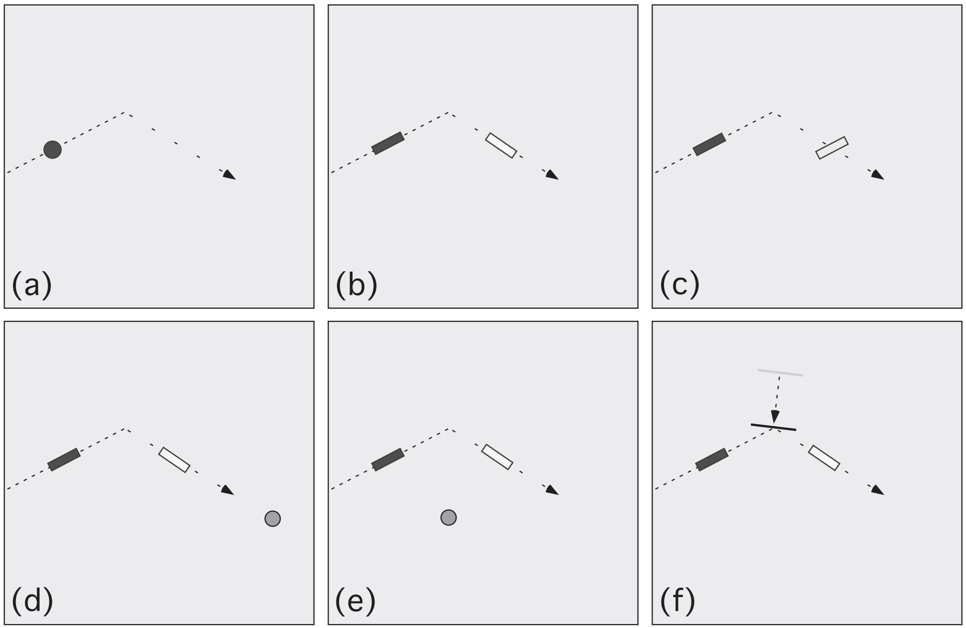

Figure 2. Examples of displays used in our animacy-rating experiments. In (a), a circular target moves at constant velocity, then abruptly turns and accelerates. This display is rated much more animate then a dot moving at constant velocity and never changing direction; generally animacy ratings increase with the magnitude of the speed and direction change. (b) The percept of animacy is enhanced if the target is an oriented rectangle, whose main axis always remains aligned with the motion path, but (c) animacy is reduced if the rectangle maintains its initial orientation, thus being *misaligned* with the motion path during the second half of the display. (d) The percept of animacy is even greater if there is a static dot (the *foil*) located along the target's path, as if the target altered its trajectory in order to avoid an obstacle. (e) This effect is diminished if the foil is located at an irrelevant location, which seems not to pertain to any goals of the target. (f) The percept of animacy is almost completely destroyed if a moving bar (the *paddle*) is introduced so that it seems to strike the target just at the point of the path change. In this case the target's change in speed and direction seems to have been caused exogenously by the paddle, rather than endogenously as an intentional act by the target.

consistent with the observed motions.

The problem with this as a complete account of animacy perception is that our experiments show that some factor *other* than energy profiles must be involved. The moving circle (Fig. 2a) and misaligned rectangle (Fig. 2c) displays have precisely the same energy profiles—because the motion is the same, with only the rigid shape differing (and the shape itself never turns, which would require energy). But the moving circle is rated more animate than the misaligned case. Why?

An answer was suggested by the next sequence of experiments. In this set, we added a static dot, called a foil, somewhere in the field. In some cases, the foil was somewhere along part of the moving target's path; Fig. 2d shows what we called the *Prey* condition, the term referring to the role apparently played by the foil. (We also included *Predator* and *Obstacle* conditions, where the foil was located in other locations also aligned with some part of the target's motion path.) In other cases, the foil could be in some other location not apparently related to the motion path (Fig. 2e, the *Irrelevant* condition). We found that the location of the foil made a big difference to the perceived animacy of the target. For example the Prey condition was rated much more animate than the Irrelevant condition. Again, this can't be explained by any account based solely on energy profiles, because the displays only differed by the location of a static element in the environment, which doesn't affect the energy analysis. Some other idea is required.

Goals

Our hypothesis was that the difference had to do with the *perceived goals* of the moving target. In the Prey condition, the fact that the new motion in the second half of the trial was directly towards the static foil created the impression in our subjects that the target *had something in its mind*—namely, an intention or desire to approach the foil. No such impression was created by the Irrelevant condition, because there is no spatial alignment between the foil's location and any part of the motion trajectory.

This hypothesis was strengthened by another series of experiment, where we added a different kind of foil, this time *reducing* the impression of goal-directedness in the target. The new foil, called a *paddle*, is a flat rectangular object that emerges from off-screen and "strikes" the target just at the point of the path change. The angle of the paddle is such that the target seems to deflect passively off of it after the collision, like a tennis ball being struck by a racket. The result is that the target no longer seems to be the causal source of the velocity discontinuity, and animacy ratings are greatly reduced.

So the critical variable influencing animacy judgments, we concluded, is whether the target appears to have a "goal." But what does this mean, exactly, and how can you tell? What does it mean in computational terms to attribute a goal to an agent? Indeed, what does it mean to attribute *any* particular mental state to an agent? These are the questions that led to the theory presented below.

Perception or cognition?

Before presenting the theory, a few other issues need to be addressed. One might well ask whether the determination of animacy by our subjects was an act more of *perception* or *cognition*. Did they as it were *see* the target as animate, or did they *decide* it was animate? Gelman, Durgin, and Kaufman (1995) have suggested that animacy judgments are a kind of "story telling;" we settle on the explanation or "story" that makes the most sense as an account of what we've seen. We agree with this idea—in fact it plays a major role in the theory below—but do not draw the conclusion that this means that it can't be perceptual. In fact many theories of perception also treat it as a kind of story-telling or explanation-construction (see Jepson & Feldman, 1996). Indeed the entire Gestalt tradition of *Prägnanz* or simplicity-based perception assumes that the perceptual interpretation drawn is in some sense the most reasonable or simplest account of the sense-data. Irvin Rock (1983) has demonstrated that many aspects of vision are akin to problem-solving, where the visual system seeks and often finds the best solution to the puzzle posed by the visual stimulus. Thus perception, too, might be a kind of "inference to the best explanation"—except that the explanation-construction process is automatic, unconscious, and involves a more limited and closed knowledge base (cf. Pylyshyn, 1999).

The question thus is whether this story-telling in the service of animacy classification is conscious or unconscious—-or, in more technical terms, whether it is a central process or an information-encapsulated input module. Rough classification of static visual images into animal/non-animal categories is known to be completed in as little as 150msec (Thorpe, Fize, & Marlot, 1996). There are several reasons to expect animacy interpretation to be a similarly early process, out of conscious control and legitimately perceptual in nature. First, as discussed above in the context of bug-detectors, it *ought* to be, in the sense that recognizing members of one's own species, detecting prey and predators, and coarsely classifying the intentions of animate targets (threatening, fleeing, etc.) are extremely urgent matters for any organism, and require the most rapid response possible. Hence one would expect the visual system to be optimized to extract whatever information about these things can be wrung out of the early motion signal. Second, all the essential elements that we've found influence animacy ratings have well-known neural hardware in visual cortex supporting them—the detection of motion direction, change in motion direction, and acceleration, and the integration of nearby context (such as our static foils) into motion computation.

Motion perception itself is not a simple matter, and involves far more than local signal-processing. Local estimates of motion have to be coordinated in a complex manner with each other in order to induce the most coherent and "reasonable" overall estimate. (For example nearby local motion estimates can be very divergent even for rigidly moving objects—the famous "aperture problem"—meaning that unless they were overridden in some intelligent way the world would look much more non-rigid than it actually does.) Hence intelligent (but unconscious and automatic) inference processes are certainly involved from the earliest stages of motion processing. Hence we feel that it is not much of a stretch to suppose that they may be influenced by more abstract—albeit unconscious—categories such as causality, animacy, and intentionality.

Of course, later conscious processes will have their say in the classification of what's been seen. But the question is: what's been seen? (Or as Lettvin et al. put it, "What does the frog's eye tell the frog's brain?") Our expectation is that input to the central system coming from the visual system has been optimized by evolution to support the most meaningful, high-level categories possible, including causality, intentionality, animacy, and all that. Understanding these aspects of the moving environment, and understanding them *quickly*, is critical for the organism's survival. Hence the question is not so much whether the system would be set up to extract them early, but rather how much such information *can* be extracted early? That is, what do a few simple motions by a target say about the target's mental architecture? The answer, certainly, is not *everything*: a moment's conscious thought will often overturn an immediate perceptual impression of animacy, as when observers of Heider and Simmel's film, perhaps after a pause, reject the strong impression that the shapes have emotions. After all, the observer knows they are artificial animations! But before this happens, we suspect, early processes have reflexively pegged the shapes as intentional. This very general impression can then be fleshed out, augmented, and completed— or perhaps overruled—by later conscious thought.

However, none of this really matters for the theory presented below. We are describing a formal process of classification of motion events (and more generally of observed behaviors). Whether this abstract process is instantiated in the brain as an unconscious reflexive system, or as a conscious process of reasoning, is essentially orthogonal to our statement of the theory. Our main goal is to understand the attribution of mental architectures as an abstract competence, and we defer discussion of how it might be implemented in the brain.

Attributed mental architectures

We will use the term *mental architecture* to denote the complex of cognitive and perceptual capacities, goals, intentions, decision rules, and behavioral tendencies with which an agent is endowed. Our aim, then, is to articulate a theory under which mental architectures may be *attributed* to agents based only on observable cues. When a mental architecture is attributed or inferred, we will refer to it as an *attributed mental architecture* or AMA. Thus an AMA is a "theory of the agent."

First principles

Our theory begins with two first principles, which we regard as axiomatic, and from which the more technical elements of the theory derive.

Principle 1. *Mental architecture is computational.*
By this we mean simply that the information-processing qualities of a mental system, from a functional point of view, are best and most completely described by computational models—that is, by computer programs or something isomorphic to them. This is in a sense the unifying credo of the the modern school of cognitive science, and will, we suspect, be considered obvious by most readers of this article. Cognitive scientists use this idea when they, for example, attempt to explain behavioral data via a computational model, linguistic judgments via a well-defined system of grammatical rules (another kind of computational system), motor programs via a particular kind of abstract neural network, and so forth. The common thread is that mental functions are explained as some sort of mechanistic process of information transformation. And all such processes are equivalent to some computer program (the famous Church-Turing thesis).

For our purposes, what this means is that mental faculties are really computational capacities, and that mental qualities (such as intentionality) ultimate have definitions that can be expressed in terms of formal properties of the underlying computer program. "Possessing mental quality X" means something like "possessing computational power Y," which in turn can be cashed out in terms of the internal structure of the machine. This is critical in allowing us to be explicit about whether a particular attributed mental architecture does or does not have a particular mental quality.

Most authors in the agency and intentionality literature are from the cognitivist school, and thus presumably roughly agree with our Principle 1. Yet most models of mental architecture in this literature are surprisingly inexplicit. Mental capacities attributed to agents are usually assumed to have some computational model underlying them, and yet these models are seldom spelled out in more than intuitive terms. We feel that this has impeded progress in explaining out exactly how various models of potential agentive capacities differ from each other, and more to the point, exactly what would be required to *attribute* them—that is, to estimate them from observations of the agent's observable behavior. The formal properties of what might be called the *intentionality induction problem* require a more concrete statement.

In our theory we have attempted to be as explicit as possible in our computational models of agents. This means, inevitably, that our agents will be very simple in form—in particular, finite automata—certainly *too* simple to be a complete model of any real animal, even an insect. However this explicitness has the benefit of allowing us to make definite statements about which computational models are actually consistent or inconsistent with particular behaviors, which models have particular mental faculties and which don't, and which computational models can be inferred from which types of behaviors and which cannot.

Principle 2. *The attribution of mental architecture is a kind of "inference to the best explanation."*

By this we mean that of all the mental architectures consistent with a particular set of observed behaviors, we will choose the one that is in some sense the most reasonable explanation of the observations, again echoing Gelman et al. (1995). Of course, *many* mental architectures will be consistent with any specific set of observations: this is the basic inductive ambiguity inherent in any interesting inference situation. We need a selection rule by which some one architecture will be inferred over all the other competing candidates. This Principle is intended to informally motivate the choice of a more formal selection rule.

More specifically, what we mean here is that a particular mental faculty will be attributed to a viewed agent when that mental faculty is a part of the *best* theory of that agent—and more emphatically, that a particular mental faculty will *not* be attributed unless it's a *necessary* component of understanding the agent. It isn't enough that a particular mental faculty is *consistent* with the observations; that's too weak. It has to be a quality without which one cannot explain what's been seen. Without some sort of principle like this, there is no reason not to attribute intentionality and animacy to every entity in the environment, even those that give no tangible evidence of controlling their own motions, or even those that don't move at all.[2]

Automata

We begin by assuming that, given a target object, observers attempt to explain its pattern of motion by attributing to it some simple computational architecture that governs its motion. In what follows, we assume that this architecture takes the form of a *deterministic finite automaton*[3] (see Lewis & Papadimitriou, 1981 for an introduction), an extremely simple type of computational system. Formally, a finite automaton consists of some set of states $\{S_0, S_1, \ldots\}$, including a designated *start state* S_0; and the capacity to move from state to state in predetermined ways depending on its input, chosen from some set of input symbols $\{a, b, \ldots\}$. Informally, an automaton is usually depicted by drawing its states as nodes, with labeled arrows connecting them to indicate possible state transitions. We further assume that in each state the automaton emits some fixed output, which we indicate by writing it underneath the corresponding state node. In what follows we will often assume that this output takes the form of a motion vector v representing the motion of the target (i.e., of the automaton itself). As a simple though artificial convention, we assume that an automaton makes actions only at discrete "ticks" of a clock, e.g. at successive video frames.

An extremely simple example is automaton A-1 (Fig. 3), which simply moves at a constant velocity, i.e., continuously emits the output v_0. (The unlabeled arrow indicates a state transition on no input, so in each frame A-1 continually returns to the starting state S_0, where it emits vector v_0; it is insensitive to any input.)

A slightly more interesting example is A-2 (Fig. 4). A-2 begins by moving at velocity v_0, but

[2] Of course, that's a perfectly coherent stance—called *animism* in theology—but it's not what we are trying to model.

[3] We note here an intriguing historical connection. Finite automata were first formalized by Kleene (1956), but the main ideas were derived from a 1943 article by W. McCulloch and W. Pitts. Six years later, these two were co-authors (with J. Lettvin and H. Maturana) of the famous 1959 article on classification of moving targets in the visual system of the frog—one of the main precursors to the research reported in the current article. Thus long before our speculations, the ideas of animacy detection and finite automata as computational architectures were already intertwined.

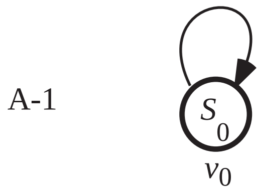

Figure 3. A simple automaton, which simply moves at a constant velocity v_0.

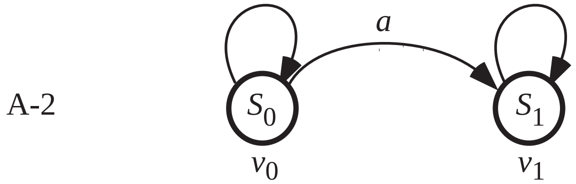

Figure 4. A slightly more complex automaton, which starts in state S_0 moving at velocity v_0, then when encountering input a moves to state S_1 at velocity v_1.

then changes to velocity v_1 when it detects the input a; hence, it can be said to have the capacity to detect input and to behave in a way that is functionally dependent on that input—that is, in effect, to *perceive*. While A-2 is still very restricted in its "cognitive" faculties, compared to A-1 it has taken one tiny step on the road towards the more elaborate mentality that characterizes animate agents.

Inference of an automaton

Our main hypothesis is that observers infer the mental capacities of a target via two steps:

1. Attribute to it a particular mental architecture, the AMA, based on its pattern of motion;
2. Classify the AMA with respect to the particular mental capacities it embodies.

The first step, inference of the appropriate automaton, is inductively ambiguous,[4] and a complete solution to it is outside the scope of this paper. However without completely solving it, we can still make certain assumptions about the form of its solution. In particular we assume that observers infer the *simplest* automaton consistent with some sequence of observed motion events, a notion we will refer to as *minimality.* The minimality assumption is in effect a technical realization of the notion of "inference to the best explanation" (Principle 2).

Formally, one automaton is strictly simpler than another if the former can be produced from the latter by deleting some of its constituent elements, such as its nodes or connections. More technically, a simpler automaton is one that has fewer nodes, but that preserves the transition relations in the original automaton—technically, a *homomorphic reduction* of the larger automaton. We will denote this relation by writing $A_1 \prec A_2$ to mean that automaton A_1 is strictly simpler than automaton A_2. An inferred automaton is *minimal* with respect to some set of target observations if no strictly simpler automaton accounts for the same observations. Hence our only assumption about automata induction is that no automaton will be inferred which contains any completely superfluous elements. Hence, for example, a target moving at constant velocity would be inferred to have architecture A-1—i.e., the AMA is A-1—because that simple architecture is sufficient to explain its motion, and no simpler interpretation is possible. Automaton A-2 is also consistent with a target moving at constant velocity—say, because input a was never encountered, or because $v_0 = v_1$—but A-2 is strictly more complex than A-1, and therefore non-minimal. For a constant-velocity target, A-1, the minimal choice, is the winning AMA. Subsequent inferences about the mental faculties of the target (in this case, presumably, that it doesn't have any) would refer only to this inferred automaton.

We next focus on the second step: classification of automaton architectures with respect to their "mental" faculties. Before considering more complex cases, and in particular before we can consider how an automaton interacts with its environment, we need to clarify just what constitutes the *input* to an automaton. We assume that the inputs to an automaton are elements of its environment that it detects: obstacles, other agents, and so forth. However, because we are interested in automata as *attributed* architectures—that is, as models of an observed target—the possibility exists that the target agent may perceive in its environment different inputs than we, the attributing observer, perceive in its environment. It may perceive finer distinctions than we do[5]; conversely, it may be blind to

[4] In fact, this problem is formally identical to the problem of inducing a regular language from finite examples, and hence is a (simple) example of grammar induction, with all the resulting well-known inductive difficulties.

[5] *This is the story of the bee/Whose sex is very hard to see/You cannot tell the he from the she/But she can tell, and*

distinctions that we perceive, thus conflating multiple distinct types of environmental inputs into one formal input symbol. Or, the automaton may perceive the same environmental elements that we do, but not "notice" them at the same time we do (such as the targets in the foil conditions in our experiments (Fig. 2d,e), where the target seems to "suddenly notice" a foil that has actually been present since the beginning of the trial). These possibilities complicate and ambiguate the attribution process enormously. Hence as a gross simplifying assumption, in what follows we will assume that the inputs to the automaton are precisely those elements of its environment that are visible to us, the attributing observer. Specifically, we will usually assume that the target automaton has access to the positions and velocities (both represented as vectors) of elements in its environment.

Three examples of automata embedded in simple environments are shown in Fig. 5. A-3 changes from initial state S_0 with velocity v_0 to state S_1 with velocity v_1 when it detects foil a; the change in state and motion is triggered by the foil. A-4 also changes to a new state S_1 upon detecting the foil a, but unlike in A-3 the new state entails the same output as the old state. Hence A-4's detection of a does not actually change its behavior, illustrating that internal mental states do not necessarily have counterparts in observable behavior. A-5 is a more complicated example in which the target moves at a different velocity in each state. It begins in state S_0 at velocity v_0, but then switches to state S_1 and velocity v_1 whenever it detects foil a; and to state S_2 and velocity v_2 whenever it detects foil b. Notice that velocities v_1 and v_2 are all "arbitrary," and don't relate in any way to the respective foils that trigger them. A more meaningful response to its environment requires a more intelligent type of agent, which we will introduced below.

It might be objected that *any* complex motion path will give rise to at least some attribution of mental capacity—e.g., some capacity to perceive or categorize inputs—simply because complex motion paths cannot be explained without complex multi-state automata with at least those properties. A-6 (Fig. 6) provides a counterexample. A-6 can "detect"[6] only one entity—a force field whose local magnitude is some force vector $\vec{F}(t)$. Its velocity output in response is simply that which obeys Newton's law $\vec{F}(t) = m\vec{a}(t)$. A simple example of a target with a complex motion trajectory that can be accounted for entirely by A-6 is a ping-pong ball, accelerating, decelerating, and abruptly changing direction, always subject entirely to the time-varying force field applied by gravity and a sequence of force-transmitting collisions with paddles, table, etc.

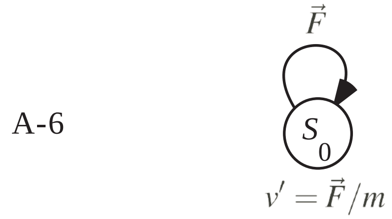

Figure 6. An automaton that responds passively to a Newtonian force field, emitting motion $\vec{F}/m$ for every input $\vec{F}$ (i.e., obeying $\vec{F} = m\vec{a}$.)

Typology of mental architectures

We now attempt to identify and formalize various mental capacities exhibited by automata. This list is not by any means intended to be exhaustive, but rather to illustrate how concrete mathematical properties of automata can serve as models of attributed cognitive competences.

Perception

A very basic mental function, exhibited by all the automata above except A-1 and A-6, is to be able to detect elements of the environment, a property we call *perceptuality*. Formally, an automaton is perceptual if it contains two distinct states S_i and S_j and some input a such that the automaton moves from state S_i to S_j on input a but not otherwise. Conceptually, an automaton is perceptual when its state is a function of its input. More loosely, we might say that a perceptual automaton's beliefs (i.e., its state) are affected by what is sees.[7]

Note that in this simple definition we don't mention exactly how the agent actually accomplishes

so can he (attr. Ogden Nash).

[6] Note that A-6 does *not* satisfy the definition of perception given below because it contains only one state.

[7] Admittedly, under these impoverished definitions it is not at all clear that an automaton has "beliefs" in anything resembling the usual sense; we mean the term suggestively.

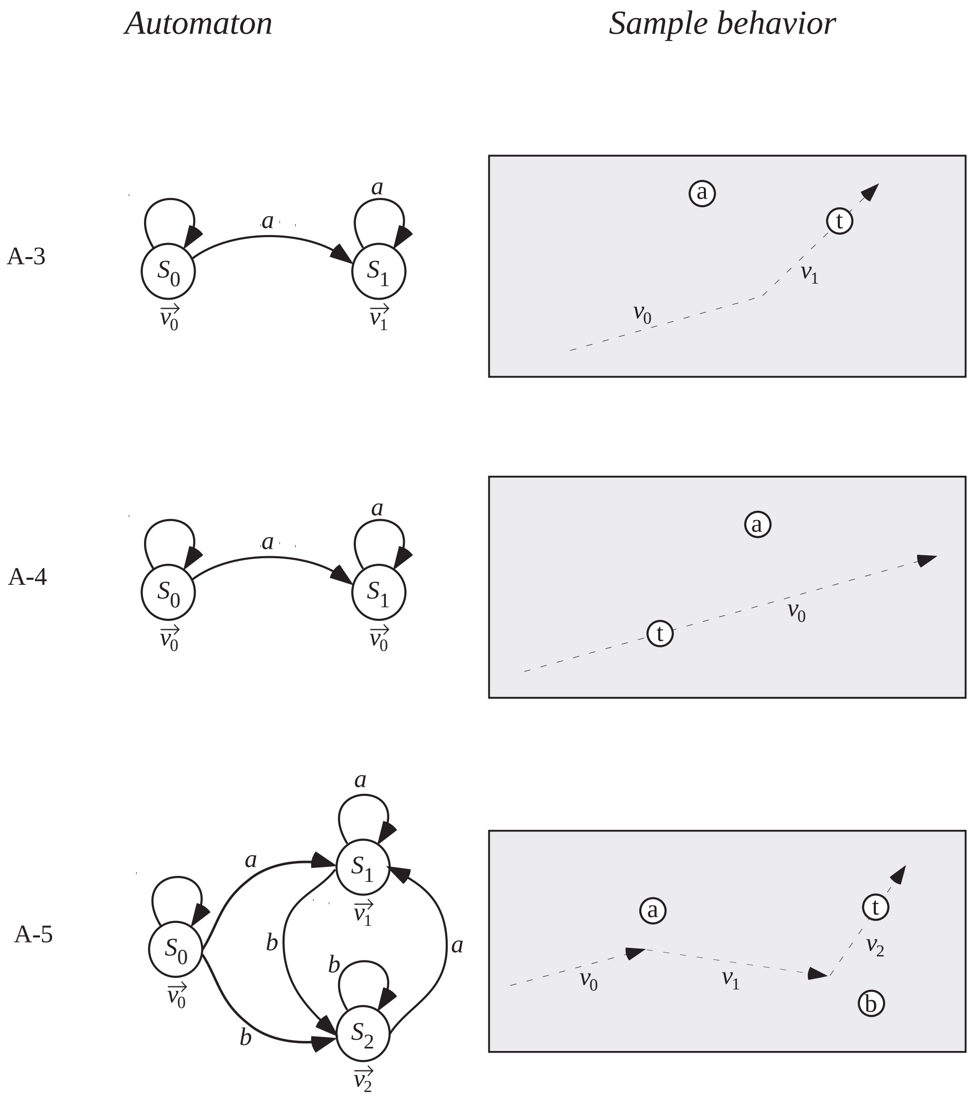

Figure 5. Three more automata and their typical motion trajectories.

the act of perception, which is of course a rich and complex question. Inside the agent, complex inference processes may be required to give rise to the simple state change reflecting recognition of input *a*. But these are invisible to the attributing observer. With respect to observable behavior only, the agent's detection of input *a* is reflected only in some outward actions necessitating the hypothesis of an internal state change, and hence this state change is the minimal necessary architecture that must be attributed to account for the capacity to perceive.

Categorization

A slightly more sophisticated mental capacity is the ability to categorize, which we call (somewhat awkwardly) *categoriality*. Formally, an automaton is categorial if it contains distinct states S_i, S_j, and S_k, and some input sequences *a* and *b*, such that starting from S_i the automaton arrives at state S_j on input *a* but S_k on input *b*. Conceptually, an automaton is categorial when its state is a *differential* function of its input; it "believes" one thing (is in one mental state) when it sees *a* and a different thing (a different state) when it sees *b*. Note that perceptuality in our definition is entailed by categoriality, though not vice-versa, because states S_i and S_j and input *a* in the latter definition fulfill the former definition.

We now give a simple theorem about categorial and perceptual AMAs, to illustrate how our formalism illuminates the conditions under which mental attributes can be inferred.

Theorem 1 *Every perceptual or categorial AMA can change its output.*

Proof sketch: This is equivalent to the claim that every perceptual or categorial AMA contains a structure isomorphic to A-2 (Fig. 4) with $v_1 \neq v_2$. We prove this by contradiction, using the assumption of minimality of AMAs, as described above. Assume conversely that every pair of connected states is as in A-4 (Fig. 5), i.e. with output velocities equal. But this is not minimal, because it is observationally equivalent to the strictly simpler structure A-1. But A-1 is not perceptual or categorial, because it has only one state. Hence every minimal perceptual or categorial automaton must contain A-2. (This completes the proof.)

The point here is that, because of the minimality assumption, it never makes sense to infer a mental capacity for which there is no direct observational evidence—and without some change in output there is in principle no evidence for any mental structure more complex than a single state. Theorem 1 shows, in effect, that the ability to change some observable aspect of behavior is absolute, mathematical prerequisite to appearing to have any mental capacities whatever.

Note that not every categorial *automaton* can change its output; above we mentioned one that cannot. Rather, the theorem states that no target without the ability to change its motion will ever be *interpreted by an observer* as having the capacity to categorize—i.e. will have a categorial AMA. Hence every attributed mental architecture that with this ability must be able to categorize. The proof hinges on the notion of minimality: though some non-motion-changing targets might have an interpretation that is categorial, this interpretation will not be minimal and hence will not be drawn.

In the particular case of moving agents—whose outputs consist of motion vectors—the theorem says that every perceptual or categorial moving agent is capable of changing its speed or direction. That is, it must have two distinct outputs that differ in at least some way. Hence if one is inferring mental states solely on the basis of motion trajectories, some change in speed *or* direction is a prerequisite to attributing complex mental capacities.

Intentionality

The term *intentionality* is used in many distinct ways in the literature, not all of which can conceivably be captured by a single formalism. We assume that the essence of intentionality is the possession of "intention:" that is, *goals* or *preferences* of some kind. How can this be captured formally?

One simple way to capture the way preferences guide actions, which will pursue, is a *utility function*. A utility function is a function that measures the degree of "happiness," satisfaction, or desirability that an agent attaches to a particular state of affairs. Commonly, this is operationalized as a function that maps events to real numbers, called utilities, expressing the agent's subjective preference for each event; higher numbers indicate more desired outcomes, lower numbers less desired. The principle guiding action is then that the agent seeks to maximize utility. More specifically, in our connection, we will suppose that an intentional agent chooses its actions in such a way as to maximize its utility. Of course, the observer has no way of knowing exactly what precisely an agent's utility function might be. Hence the central inferential problem is to determine whether an agent's actions are consistent with the maximization of *some* utility function. Hence our analysis is necessarily couched in abstract terms relating to general properties of utility functions, rather than to any specific expectation about what agents might tend to prefer. We can't assume we know that much about any agent. Instead, we attribute intentionality when the agent's actions satisfy more general properties of utility maximization.

In what follows, we develop a simple formalization of agents and their actions and utility functions, sufficient to capture some general properties of utility functions that are useful for inference. To help make the formal machinery as clear as possible, we will use as a running example a simple visual target whose inventory of possible actions is the set of simple motions in the plane—exactly like the moving targets in our animacy experiments. However we will attempt to keep the actual formalism as general as possible, so that the resulting theorems will be applicable to arbitrary agents with arbitrary domains of action. The theory applies to people buying groceries or playing the stock market, as much as it does to moving dots avoiding static dots on computer screens. But to make the meaning of the formalism as concrete as possible, we will consistently use the running example of a dot whose state is expressed by a position vector and whose actions correspond to velocity vectors. In the end we will prove a theorem concerning the attribution of intentionality to an agent, using the moving dot as the concrete example, but leaving it clear that this is simply a specific illustration, rather than a completely general example, of the domain to which the theory applies.

Inferring intention

Notation. In order to formalize an agent with sufficient detail to be able to capture intentionality, we need to express the *configuration* of the agent (e.g., the position of our moving target in the plane), and the *state of the world* (i.e. the target's environment, e.g., the configuration of foils, etc); and the possible actions that the agent may take (e.g. the target's motions in the plane). For these we adopt the following notation:

- X, the set of possible configurations of the agent, denoting a particular state by $x \in X$;

- E, the set of environments, denoting a particular environment by $e \in E$;

- V, the set of actions, denoting a particular action by $v \in V$.

An action v taken by an agent is really a change in its state x, meaning that V is a set of transformations mapping X to X. Specifically, after an agent at x takes an action v, its new state is

$$x' = x + v. \tag{1}$$

In the moving dot world, think of x denoting the position of the target, e the state of the environment, and v the velocity (speed and direction) the target adopts at a particular point in time.[8] We are assuming action at discrete time clicks, so we can think v as denoting a discrete position change Δx that maps x to $x' = x + \Delta x$. The crux of our argument below concerns how actions v $(= \Delta x)$ are selected by the agent. The interpretation of the agent as intentional or not will hinge on whether this selection function seems "intentional."

Utility

Our notion of intentionality is that agents act so as to maximize their "happiness." We capture the notion of happiness formally by means of a *utility function*, which is simply a function that indicates the agent's preference or satisfaction with a particular state of affairs. In our notation a "state of affairs" refers to the state of the world (e) and the state of the agent itself (x) in the world. Thus we formalize the utility function as a function mapping $X \times E$ to the real numbers R, i.e.

$$u(x, e) = r \tag{2}$$

gives the agent's subjective utility r concerning the state (x, e) of the agent and world. My subjective happiness at any time is a function of where I am and what the world is like. If I am in Cleveland and I don't like Cleveland, say, than my utility is low. If I leave Cleveland (x changes), or Cleveland improves (e changes), my utility will go up. Crucial in the following discussion is that the agent has no power to change the world e, but it *does* have the power to change its own state x, by executing an action v. (I can't change Cleveland, but I can leave.) An agent that consistently chooses v so as to increase its utility is acting intentionally.

Definition 1 (Intentionality) *Assume an agent A having configuration space X, environment space E, and action space V as defined above, and assume that its output $v \in V$ is always given by some selection function*

$$F_A(x, e) = v, \tag{3}$$

meaning that in configuration x in environment e, A always outputs $v = F_A(x, e)$. Then the agent is called intentional *if there exists some utility function*

$$u = u(x, e) \tag{4}$$

such that $F_A(x, e)$ always maximizes $u(x + v, e)$.

[8] By adopting the notation $x + v$ for the result of applying action v to state x, we are in effect assuming that V is an additive group, such as the space of vector motions. This is actually a more restrictive assumption than we really require, but adopting some more general notation (e.g. $x' = T_v(x)$, suggesting that x' is the result of applying transformation T_v to x) greatly complicates the notation without adding much in the way of substance.

A simple example is the *Prey* condition from our experiments, where the moving target apparently acted like a predator. (Again, the name of the condition refers to the role played by the foil.) Here, in the judgment of our subjects, the target acted as if it had a utility function that depended entirely on the distance d between itself and the foil, and specifically was a monotonically decreasing function of d. Thus among all its possible actions (motion vectors) it chose the one that minimized d on the next time step, which means taking the largest possible step towards the foil. Hence the larger this step (the larger its acceleration towards the foil), the more it appeared to be obeying such a utility function, the more it appeared to be intentional, and the higher the animacy rating.

The variety of intention.

The key idea in what follows is that the observer is trying to guess, based on observing particular actions by the agent, whether the agent is consistently maximizing some utility function. Ideally, what we would like to do is to articulate conditions under which a utility-preserving automaton is the best or simplest attributed mental architecture—much as we did above with perceptual automata and categorial automata—and then declare that intentionality will be attributed just under those conditions.

The problem is that utility functions can take a nearly unlimited variety of forms, simply because different types of agents can have radically different preferences—“different strokes for different folks.” We certainly don’t want to solve this problem by assuming that utility functions will closely resemble our own, or adhere any narrow preconceptions about what counts as a reasonable desire. We would like to be somewhat more inclusive. How then can we spot intentionality?

Our strategy instead is to begin with what seem like the most basic and universal attributes of utility functions—conditions that are required in order for them to coherently satisfy the role they play in inference—keeping the assumptions broad enough to encompass a vast range of specific preferences. Thus we can still have people who like broccoli and people that like sky-diving, animals that run from lions and animals that chase lions, etc., all within the very loose confines of our assumptions.

Utility-preserving transformations.

As discussed, utility functions can vary very freely from agent to agent, but even given this freedom one would not expect a utility function to be completely arbitrary. For example, we expect utility functions to exhibit certain kinds of consistency. Obviously, the same agent, under the same circumstances, with the same utility function, ought to make the same choice. More subtly, even when the circumstances are not *precisely* the same, we expect certain kinds of consistency in relation to other choices. For example, most preferences don’t really relate to the absolute situation at all, but rather to how the agent is situated *relative* to the environment. For example, a basketball player’s choices (i.e., his or her choice of motor movement at each point in time) don’t depend on the *absolute* position of the basket (which after all never changes), but rather by its position *relative* to him or herself, the positions of the other players relative to himself, etc. This imposes a certain type of consistency on the utility function: it means that choices will be the same whenever relative conditions are the same, regardless of the absolute state of affairs.

A clear example of this comes from our moving dot displays, where in many cases the targets’ utility appears to be determined relative to the location of the foil. The “prey” target darts in many different directions, but all of them happen to point away from the foil; this means that its motion direction relative to the foil is actually constant. And this constancy of (relativized) behavior is part of what gives the impression of goal-directedness.

In fact, this kind of relativization of behavior may be inevitable if the agent actually can’t *perceive* the absolute locations of entities in its environment, but only their locations relative to itself—i.e. in an egocentric coordinate system. Its utilities of course can only be computed based on its perceptions, so this leads directly to the utility function being similarly relativized. But more generally, even if the target knows its own position and positions in the environment in some more absolute sense, it may only *care* about its location relative to particularly important environmental elements.

Formally, this relativization is captured very generally by the notion of a *utility-preserving transformation*. A utility-preserving transformation t is a mapping that converts a particular configuration and environment (x, e) into *different* configuration and environment (x', e'),

$$(x', e') = t[(x, e)] \quad (5)$$

such that utility is invariant, i.e.,

$$u(x, e) = u(x', e'). \quad (6)$$

We define a *class* of utility preserving transformations T as a set of transformations in which each member $t \in T$ is utility-preserving for the agent in question. We will generally assume that each utility-preserving transformation t has an inverse t^{-1} that is also utility-preserving.

Again the moving-dot world provides a convenient and easily-visualized example. Take T to be the set of translations in the plane. A particular $t \in T$ is a particular translation, such as movement to the right by 5 meters. In our moving dot world, the single-dot foil itself it the only detectable "environment." So it is easy to imagine that this transformation preserves utility: all this means is that if you move the target and the foil both by 5 meters to the right, preserving their relative spatial positions, then the agent's state of happiness would be unchanged. Similarly for any other translation, such as movement left by 6 meters, or up by 3.14 microns. If the target only cares where the foil is relative to itself, then none of these changes make any difference. Hence not only is each particular t utility-preserving, but so in fact is the entire class T of planar translations.

Similarly, if relative position of the target and foil are all that matter, then the class of rigid rotations would also be utility-preserving, and more generally so would the class of transformations involving both translation and rotation (technically called isometries). Moving everything by 5 meters and rotating it by 37° doesn't change relative positions, and obviously the same can be said for any distance and angle.

Fig. 7 shows some simple examples. The configurations in Fig. 7(a) and (b) differ in both the position of the agent x, the configuration of the environment e (here represented by the position of the foil), and the action taken by the agent (v_1 vs. v_2). But the entire ensemble of position, foil, and action in the two cases differ only by a rigid rotation. Hence (again assuming that such transformations don't affect utility), the agent's action in the two cases is actually the same *relative* to the environment (the foil). In each case it is trying to get to a particular spot—defined, as it were, in foil-centered coordinates.

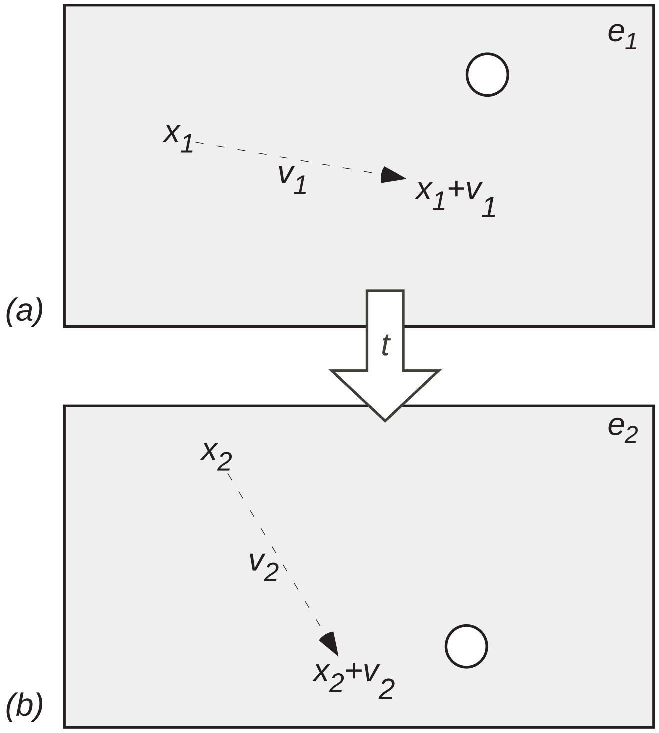

Figure 7. Illustration of the Intentionality Conditions (special case of identical actions). The figure shows two configurations of an agent (x) and environment (e), and the resulting actions v. The two situations differ by a rigid motion, which by assumption preserves utility. Hence the agent's action in (b) (v_2) is actually the same as in (a) (v_1), modulo such transformations. Hence in (b) the agent is executing the same action modulo the environment as in (a), and for the same reasons.

It might be objected that in many circumstances translation of the target and foil would not be utility-preserving. Perhaps moving 5 meters to the right meters to the right brings both target and foil closer to a tree, where the target might be able to hide, which would influence its escape plan and thus its motion choices. However in this case we have not really rigidly translated the entire environment—because we didn't move the tree. If, as assumed, the *entire* environment is moved—foil, tree, and all—then we would again expect utility to be preserved. Again this notion is simply a way of articulating that it is the target's configuration relative to that of the environment that matters to utility. Any transformation that by definition does not affect this relative configuration will not, ipso facto, affect utility.

There is nothing in general to guarantee that any particular transformation will be utility-preserving—we have been using rigid motions only as a suggestive example. More subtly, there is no guarantee that any particular utility function will submit to any utility-preserving transformations. However the example suggests that this is a very basic attribute of "reasonable" utility functions: that they are only affected by certain factors, and that transformations that don't change those factors won't affect utility. Hence in what follows we will consider what can be said about utility functions that *do* submit to some utility-preserving transformation class T. However we will not assume any *particular* transformation. This allows us to keep the assumptions abstract and non-specific, so that we can see what rules apply generally—although for concrete visualization it is still useful to keep the moving dots-world, unaffected by translation and rotation, in one's mind.

The mathematics of transformations that leave something invariant, called group theory provides terminology for expressing properties that are defined relative to some transformation, called *modulo*. A familiar example is division modulo some integer. Two numbers x and y are the same mod 10 if they leave the same remainder when divided by 10, e.g. 13 and 43; in this case we say that

$$13 = 43 \text{ mod} 10, \tag{7}$$

This is just another way of saying that 13 and 43 are the same if you ignore addition of multiples of 10—the "transformation" relevant here. In place of the somewhat awkward "mod" notation, we will use the symbol $\overset{T}{=}$ to denote "equality mod T", that is, we write

$$(x, e) \overset{T}{=} (x', e') \tag{8}$$

whenever configurations (x, e) and (x', e') differ by a utility-preserving transformation $t \in T$. By definition, $(x, e) \overset{T}{=} (x', e')$ implies $u(x, e) = u(x', e')$. The notation $(x, e)/T$ refers to the abstract point "(x,e) mod T," that is, to the position of point (x, e) in X, E-space after transformations in T are disregarded. (This is just like remainder 3 in the above example; division modulo 3 creates a space of remainders—0, 1, and 2—which you can think of as all that's left of the integers when you disregard the part of each integer that is a multiple of 3.) This notation is very useful for expressing the structure of the situation (agent and environment) disregarding changes that don't affect utility.

Preservation of preferences.

When an agent and environment change by a utility-preserving transformation, not only are utilities preserved—that's true by definition—but so are the agent's *choices of action*. This critical fact is captured by the following lemma.

Lemma 1 (Preservation of choices under utility-invariance) *Assume an agent A such that $F_A(x, e) = v$, and assume that t is a utility-preserving transformation with $t(x, e) = (x', e')$. Then*

$$F_A(x', e') = v. \tag{9}$$

In words, if A takes action v in configuration (x, e), then it will also take action v in the transformed configuration (x', e').

Proof sketch. By contradiction. Imagine, contrary to the theorem, that there exists a $\tilde{v} \neq v$ such that $u(x' + \tilde{v}, e') > u(x' + v, e')$. Note that t^{-1} (the inverse of t) exists and is also utility-preserving. Hence $u(t^{-1}(x' + \tilde{v}, e')) > u(t^{-1}(x' + v, e'))$. But because the latter is just $u(x + v, e)$, this violates the assumption that $u(x + v, e)$ is maximal. Hence $\tilde{v}$ as described cannot exist, completing the proof.

This theorem establishes that utility-preserving transformations preserve not just utilities but the entire preference rank-ordering of an agent's actions. Hence below this will allow us show that an agent who makes choices that obey some kind of utility-invariance can be described more simply—be given a simpler attributed mental architecture—than one that doesn't. This in turn gives a basis for inferring intention. If an agent obeys (some kind of) utility-preservation, then it must *have* a utility function. And if an agent has a utility function, then (by definition) it's intentional. If it's intentional, then it's probably animate.

Attributing utility functions.

When is it reasonable to infer that an observed agent's behaviors are the product of systematic utility maximization? As discussed above, we have to solve this problem without assuming that we know exactly what it likes and dislikes. The key instead, more modestly, is to look for consistency in its actions, implying consistency in its preferences.

Specifically, we look for a collection of actions all of which can be seen as manifestations of the same preference, transformed in various ways that affect the details but don't affect what the agent cares about—i.e., don't affect utility.

The simplest situation is an agent with two distinct behaviors observed at different times. Assume the agent emits action v_1 in situation (x_1, e_1), and v_2 in situation (x_2, e_2) (again see Fig. 7). What should the attributed automaton look like? The answer depends on how we regard the transformation t that maps (x_1, e_1) to (x_2, e_2).

Imagine first that we know *nothing* about t, and thus don't connect (x_1, e_1) and (x_2, e_2) in any particular way. In this case, all we can do is construct an automaton that simply parrots the observations, in effect saying "in (x_1, e_1) it does v_1 and in (x_2, e_2) it does v_2" (see the uppermost automaton in Fig. 8). This automaton has two states per behavior (plus a single extra start state), one of which reflects the agent's state before it perceived that it was in (x_i, e_i), and the other one after, when it is emitting v_i. Note that we can *always* use this procedure to create an automaton that fully explains the observed actions in a completely dumb way, spending exactly $2n+1$ states to explain n observed actions (one start state, one state for each pre-action configuration, and one state for each post-action configuration). This is the "raw" or uncompressed interpretation of the agent's actions, which we will denote by A_{raw}. This AMA attributes no real mental structure, but simply a matrix of rote behaviors.

Conversely, now imagine that we notice that the agent's two actions obey a very special relationship: utility equivalence (or, more correctly, imagine that we *guess* that the transformation mapping the first to the second is utility invariant). Formally, the relationship is defined by the following two-part condition:

Intentionality condition (special case of identical actions)

(a) $(x_1, e_1) \overset{T}{=} (x_2, e_2)$;

(b) $(x_1 + v_1, e_1) \overset{T}{=} (x_2 + v_2, e_2)$

Part (a) says that the agent's pre-action states were utility-equivalent, and part (b) says that its post-action states were utility-equivalent. If an agent obeys this condition, what automaton can we attribute to it? To see the answer, notice that if the first action, v_1, maximizes utility for the agent, then, by Lemma 1, so does the second action. The second action is consistent with the same utility rank-ordering as the first one. Formally, this means that we can redraw the attributed automaton in a simpler way. Specifically, we can relabel the agent's two actions in abstract "mod-T" space, where the two actions are in fact the *same* action, because, by hypothesis they differ only by utility-invariant transformation. In this space the pre-action points (x_1, e_1) and (x_2, e_2) both correspond to the same point, which we can call $(x_1, e_1)/T$ (we could just as well have called it $(x_2, e_2)/T$; the choice is arbitrary). Similarly, the post-action states $(x_1 + v_1, e_1)$ and $(x_2 + v_2, e_2)$ both correspond to the same point, labeled $(x_1 + v_1, e_1)/T$. When viewed through the prism of utility-invariance, the two observed actions were really the *same* action.

This leads to a very concrete simplification of attributed automaton. Because what had appeared to be two actions can now be represented as a single action repeated twice, we can represent the agent's mental architecture by an automaton with fewer nodes (Fig. 8). By Principle 2 (minimality), we prefer the simpler interpretation, i.e. the smaller automaton. But this simpler interpretation *entails* an attribution of intentionality, because it hinges on regarding t as utility-preserving, and only intentional agents choose their actions by maximizing utility. Thus Condition 1 sanctions the attribution of intentionality.

The following simple theorem generalizes this situation.

Theorem 2 *Consider an agent that is observed in n situations $(x_1, e_1) \dots (x_n, e_n)$, in which it executes respectively actions $v_1 \dots v_n$. Denote by A_{raw} the $2n+1$-state automaton describing these actions "verbatim," as described in the text (e.g. see Fig. 8).*

Now assume that $t \in T$ is some utility-preserving transformation satisfying the Intentionality Conditions (identical-actions case) given above, i.e.

$$t(x_i, e_i) = (x_j, e_j) \qquad (10)$$

$$t(x_i + v_i, e_i) = (x_j + v_j, e_j) \qquad (11)$$

for $i \neq j$. Then it is possible to attribute a 3-state automaton $A_{intentional}$ (see Fig. 8), such that

$$A_{intentional} \prec A_{raw}. \qquad (12)$$

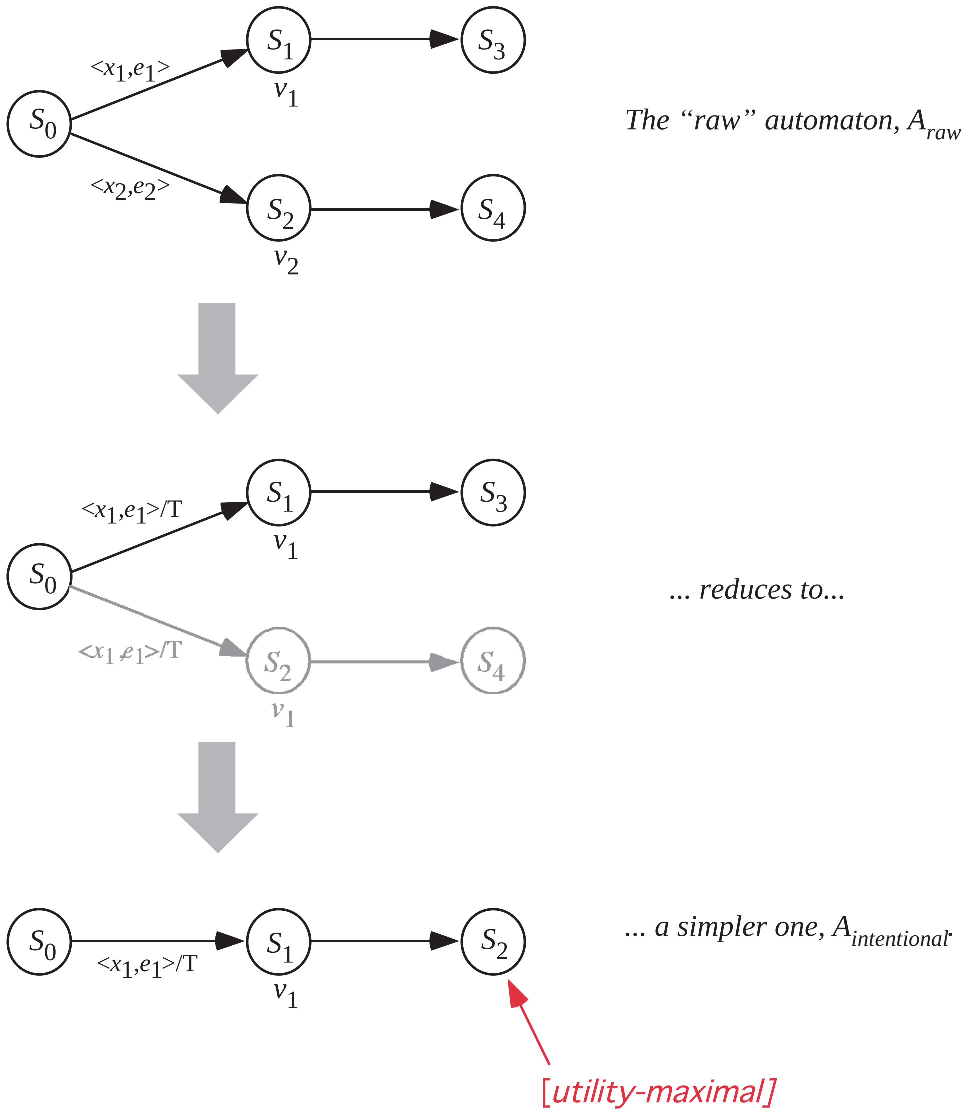

Figure 8. The collapse of A_{raw} to form $A_{\text{intentional}}$ under utility-preservation. Notice how the utility-equivalence enables a relabeling of A_{raw}'s states, making some of them equivalent to each other and thus redundant, which in turn leads to the re-expression as $A_{\text{intentional}}$.

Proof sketch: The initial state (S_0) is the same in A_{raw} and $A_{\text{intentional}}$. The next n states of A_{raw} map to one state (S_1) of $A_{\text{intentional}}$ because by assumption they are all equivalent under t, meaning that they all denote state $(x_1, e_1)/T$. Similarly the next n states of A_{raw} map to S_2 of $A_{\text{intentional}}$ because they are all equivalent to $(x_1 + v_1, e_1)/T$. This shows that A_{raw} maps homomorphically to $A_{\text{intentional}}$, demonstrating that $A_{\text{intentional}} \prec A_{\text{raw}}$ as claimed.

The Intentionality Conditions in this version led to an attribution of intentionality, but the requirement that we observe the same agent and environment multiple times is unnecessarily strong. We don't often observe an agent executing precisely the same action twice, let alone n times. However, if we weaken the requirement of identical initial states, a similar theorem applies. The more general case is this:

Intentionality condition (general case)

$$(x_1 + v_1, e_1) \overset{T}{=} (x_2 + v_2, e_2) \qquad (13)$$

As before, this condition leads directly to a simplification of the resulting automaton, as spelled in the more general case out by this theorem.

Theorem 3 *Consider an agent that is observed in n situations $(x_1, e_1) \ldots (x_n, e_n)$, in which it executes respectively actions $v_1 \ldots v_n$. Denote by A_{raw} the $2n+1$-state describing these actions "verbatim," as in Thm. 2 (see Fig. 10).*

Now assume that $t \in T$ is some utility-preserving transformation satisfying the Intentionality Condition given above,

$$t(x_i + v_i, e_i) = (x_j + v_j, e_j)$$

for $i \neq j$. Then it is possible to attribute an $n+2$-state automaton $A_{intentional}$ to the agent such that

$$A_{intentional} \prec A_{raw}, \tag{14}$$

(see Fig.10).

Proof sketch: The proof here follows the same outline as Thm. 2, except that here only the last n states of A_{raw} collapse to one state (S_{n+1}) of $A_{\text{intentional}}$ (because by assumption they are all equivalent to $(x_1 + v_1, e_1)/T$). This again suffices to prove that $A_{\text{intentional}} \prec A_{\text{raw}}$ as claimed. (This completes the proof.)

This theorem is very similar to Thm. 2; the only difference is that because the first configurations are not necessarily utility-equivalent, the corresponding states of the automaton do not collapse after rewriting as $A_{\text{intentional}}$. However, the *final* configurations are utility-equivalent, so the corresponding states do collapse, leading the required subset relation.

In words: if an agent obeys the Intentionality Condition, then it is subject to a more compact description than if it doesn't—specifically, the best description goes from $2n+1$ states to only $n+2$. When we attribute utility-invariance, we can rewrite the automaton in terms that are expressed only modulo T. Then all actions that are equivalent modulo T have the same names in mod-T space, and thus can be represented by a single node, replacing the larger set that was necessary in the non-intentional automaton A_{raw}. So even though here the two actions weren't themselves identical mod T, the states following the action *were* thus identical, and both were maxima of the same utility function. This equivalence of utility states leads directly to an equivalence of attributed mental states, and to the simplification of the attributed mental architecture, and thence to an inference of intention. The key is that a collection of actions on the part of the agent may appear disjoint and idiosyncratic if we don't understand its decision-making. But when we *do* understand it—or at least understand that some actions seem to be achieving the same net environmental conditions for the agent as other actions—then we can model the agent in a simpler way.

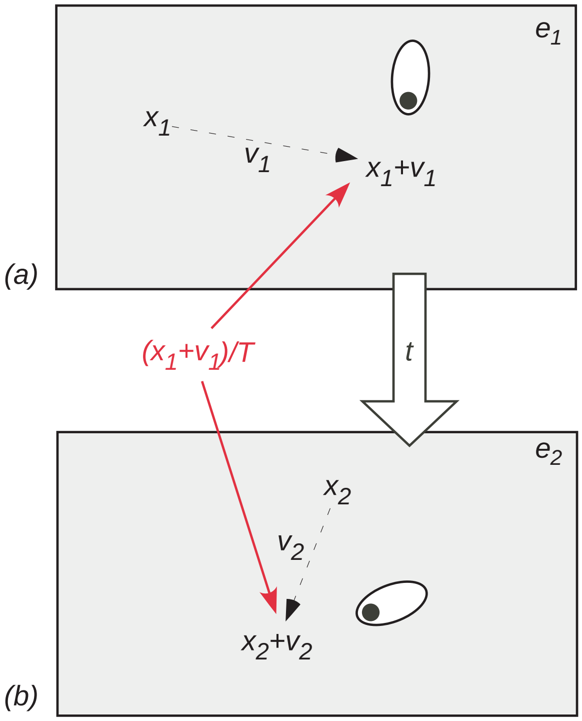

Figure 9. Illustration of Intentionality Conditions (general case). Here the two configurations (x_1, e_1) and (x_2, e_2) are *not* utility-equivalent, but the post-action configurations $(x_1 + v_1, e_1)$ and $(x_2 + v_2, e_)$ are utility-equivalent. (They are both $(x_1 + v_1, e_1)/T$, as indicated in the figure.) Hence these two agents are not taking the same action, but they have the same motivation—that is, they reflect the same utility function.

An agent that always flees at the sight of the foil is really only capable of one behavior—fleeing—regardless of how many superficially distinct actions this motion may appear to require if you don't understand the utility function. More generally, an agent that maximizes utility will have, as an automatic consequence, a wide range of behaviors that may seem unrelated to each other, but only if you don't see that they all achieve similar ends. When you don't understand its utility function at all, its diverse array of behaviors seems complex; when you do, you realize that some of them are systematically related to each other, and the attributed mental architecture required gets a notch

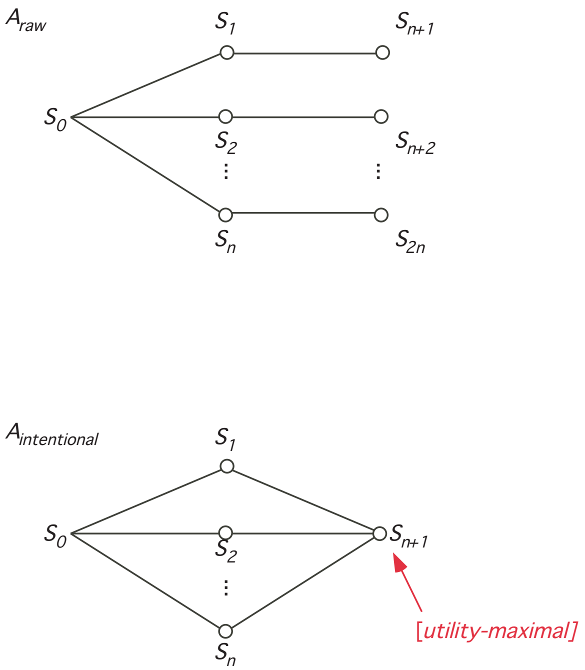

Figure 10. Schematic illustrations of A_{raw} and $A_{\text{intentional}}$ in the general case (see Thm. 3), illustrating the enumeration of states.

simpler. When an intentional interpretation is possible, minimization of the AMA points you in the direction of it. Aha, *that's* why it's doing what it's doing—*now* I understand.

Adequacy criteria for the attribution of intention. A useful analogy can be made between the attribution of intentionality, as formalized here, and the interpretation of structure from motion (SFM) as formalized by (Ullman, 1979). Ullman showed that it made sense to infer 3D structure from image motion if the image motion took a certain, very specific, form: namely, that it was consistent with a rigid rotation in depth. The key point is that not *all* image motion can be interpreted that way; if the motions were generated at random, it usually can't. Specifically, Ullman showed that three views of four points generally don't have any rigid interpretation at all. Hence if you see three views of four points that *do* have a rigid interpretation, you ought to draw that interpretation—you ought to see the points as rigid motion in depth. This sheds light on the underlying logic of 3D motion interpretation: three-dimensionality is seen when the dots move in a way that is consistent with a rigid model, precisely because most motions aren't consistent with any rigid model at all.

The analogy to intentionality interpretation is fairly direct, with utility-invariance playing the role of rigidity. Most sequences of actions aren't consistent with an intentional interpretation, so you can't draw one. And precisely for this reason, when a given sequence is consistent with intentionality, you should infer intentionality. The collapse to a simple automaton model in Thm. 3 after two actions by an intentional agent is exactly like the collapse of Ullman's three potentially unrelated sets of four points to a single rigid 3D model. In this light, intentionality can be seen as a kind of "non-accidental" property of action sequences; it's atypical of random actions, but it's expected of utility-maximizing ones.

The formal analogy can be clarified a bit further if we re-express the Intentionality Condition as a formal constraint on the second action, stating exactly what condition it must satisfy in order for the two to be consistent with a common intentional interpretation. To do this, we observe that transformations t act uniformly on both environments and the agents embedded within them. Formally, this means that the Intentionality Condition (Eq. 13) can be rewritten as a system of two simultaneous equations

$$t(e_1) = e_2, \tag{15}$$

$$t(x_1 + v_1) = x_2 + v_2, \tag{16}$$

in both of which the t represents the same common transformation to both environment and agent. We first use Eq. 15 to "solve for t." Denote by $t_{e_1 \rightarrow e_2}$ the transformation mapping e_1 to e_2, i.e. t satisfying

$$t_{e_1 \rightarrow e_2}(e_1) = e_2. \tag{17}$$

Then we plug this into Eq. 16 to obtain

$$t_{e_1 \rightarrow e_2}(x_1 + v_1) = x_2 + v_2. \tag{18}$$

Finally by subtracting x_2 from both sides we obtain

$$v_2 = t_{e_1 \rightarrow e_2}(x_1 + v_1) - x_2. \tag{19}$$

This statement takes the form of a condition on the agent's second action (v_2) that entails intentionality, in exactly the same way that Ullman wrote down a geometric condition on the third view for it to be consistent with a rigid transformation of the first two. After observing the agent's first action, one can't say if it is intentional or not. But if and only if its second action satisfies Eq. 19, then we *can* conclude that the Intentionality Condition holds; the final states of the agent after the two actions (respectively, $(x_1 + v_1, e_1)$ and $(x_2 + v_2, e_2)$) are utility-equivalent and each is utility-maximal; the agent chooses its actions via utility maximization; the agent's mental architecture can be simplified as in Fig. 10; and the agent is intentional.

Like SFM, this is the "competence theory"—the ideal—and a real observer may not be able to recognize the rigid (utility-preserving) relationship after minimal set of views (actions). In the case of SFM, subsequent to Ullman's theorem it was found that human observers in many cases cannot detect rigidity after the theoretically minimum number of points and views, or even could exceed his theoretical upper limit of performance by taking advantage of other constraints not reflected in the theorem. Similarly, real observers may not be able to detect the intentional interpretation after only two actions if, for example, they don't postulate the correct utility-invariant transformation. Conversely if the observer is able to correctly guess the target's utility function—perhaps by projecting a plausible or familiar utility function—then, likewise, intentional behavior can be detected ahead of the theoretical minimum of two actions given by the theorem. All these possibilities though involve attributions and mechanisms outside the formal assumptions of our theorem, that we freely admit might be part of a real observer's implementation of intentionality detection.

Conclusions

In the above, we have speculated that the inference of mental qualities involves (a) attributing to the target a particular mental architecture, in the form of an automaton, that is as simple as possible while still sufficing to account for the target's observable behavior, and then (b) evaluating the mental qualities of the target by reference to the formal properties of the attributed automaton. Although we have focused on the moving-dot world exemplified by our experiments, we have tried to keep the formalism general enough to embody arbitrary action spaces and utility functions. Thus, we feel that the broad outlines of our argument apply to the more challenging stimuli faced by the infant deciphering its new environment, the prey animal attempting to predict the behavior of a predator (or vice-versa), or to adult human's understandings of the actions of others.

One might object that our notion of computational architecture (finite automata) is too simple or simply wrong; that our notion of simplicity (subset inclusion over state diagrams) is inappropriate; or that our formal definitions of mental qualities such as intentionality are inapt or overly broad. We agree with all these objections. Each of these formal choices is, we feel, simply a place-holder for some more sophisticated conception that ought to replace it. We welcome suggestions. What *is* important in what we propose is the larger conceptualization of what is entailed in the attribution of mentality: (a) attribution of a formal computational architecture (whatever specific form it may take), via (b) some well-defined inductive inference scheme (such as a simplicity metric), and (c) evaluation of mental properties with respect to, and only with respect to, the attributed mental architecture. Thus we have attempted to realize computationally Gelman et al's notion of "storytelling:" we attribute intentionally when the best story we can tell about the agent (minimal automaton) is intentional (maximizes utility).

It might more specifically be objected that our automata models of mental architecture are absurdly oversimplified renditions of the capacities exhibited by even the simplest living things. But in a sense, this oversimplification is the whole point: entities are regarded as intentional, and hence animate, when even a "minimalist" model of their behavior necessarily includes mental states and goals. Indeed, even with highly impoverished internal architectures, agents can produce actions with a surprisingly high degree of *apparent* meaningfulness, a point famously made by Braitenberg in his book *Vehicles* (Braitenberg, 1984). Our point is the inverse of Braitenberg's: a useful interpretation of the mental capacities of moving targets does not require that they be completely simulated, but can be based on an inference of the *essential* mental-architectural components without which the target's motion cannot be explained.

The search for intentional life

The well-known Search for Extra-Terrestrial Intelligence (SETI) program is based on the idea that we can detect intelligence via patterns in the observable (electromagnetic) output of other civilizations, without knowing any of the details or particulars of the communication systems those civilizations might use. The problem is closely analogous to the one we have posed here. How can you distinguish intelligent behavior from random noise or natural phenomena? How can you detect an intentional agent without knowing what its intentions are?

Notice that there's a subtle non-monotonicity in the *complexity* of patterns with respect to the inference of intelligence. Too simple—a simple periodic noise burst, say—and it's an inanimate source, say, a rotating pulsar. Too complex—a totally patternless sequence—and it's just random electromagnetic interference. To seem intelligent it has to be somewhere in between: patterned, but neither perfectly periodic nor completely chaotic.

Our model of intentionality attribution runs along similar lines. Our model of intentional behavior assumes an agent complex enough to maximize its subjective utility, and thus to evaluate its environment in terms rich enough to estimate the relative benefit of various actions. (Recall that automata with too few states can't even be perceptual or categorial, much less intentional.) Yet following Thm. 3 we can't *attribute* such utility-seeking behavior unless attributing it simplifies our model of the agent—or, putting this another way, unless then agent's behavior is capable of being simplified by an intentionality attribution. (Eq. 19 is the condition it has to satisfy for this to be true.) Thus intentional behavior has to be both reasonably complex *and* reasonably regular, in this specific sense. Otherwise, it's better explained as something either more *random* or more *boring* than intelligence.

References

Baker, C. L., Tenenbaum, J. B., & Saxe, R. R. (2006). Bayesian models of human action understanding. In Y. Weiss, B. Schölkopf, & J. Platt (Eds.), *Advances in neural information processing systems 18.* Cambridge, MA: M.I.T. Press.

Bingham, G. P., Schmidt, R. C., & Rosenblum, L. D. (1995). Dynamics and the orientation of kinematic forms in visual event recognition. *Journal of Experimental Psychology: Human Perception and Performance*, *21*(6), 1473–1493.

Blythe, P. W., Todd, P. M., & Miller, G. F. (1999). How motion reveals intention: categorizing social interactions. In *Simple heuristics that make us smart* (pp. 257–285). New York: Oxford University Press.

Braitenberg, V. (1984). *Vehicles: Experiments in synthetic psychology*. Cambridge: MIT Press.

Carey, S. (1985). *Conceptual change in childhood*. Cambridge: M.I.T. Press.

Chatterjee, S. H., Freyd, J. J., & Shiffrar, M. (1996). Configural processing in the perception of apparent biological motion. *Journal of Experimental Psychology: Human Perception and Performance*, *22*(4), 916–929.

Csibra, G., Bíró, S., Koós, O., & Gergely, G. (2003). One-year-old infants use teleological representations of actions productively. *Cognitive Science*, *27*(1), 111–133.

Dasser, V., Ulbaek, I., & Premack, D. (1989). The perception of intention. *Science*, *243*, 365–367.

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

Gallese, V., & Goldman, A. (1998). Mirror neurons and the simulation theory of mind-reading. *Trends in Cognitive Sciences*, *2*(12), 493–501.

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.* New York: Oxford University Press.

Gelman, R., & Spelke, E. (1981). The development of thoughts about animate and inanimate objects: Implications for research on social cognition. In J. H. Flavell & L. Ross (Eds.), *Social cognitive development: Frontiers and possible futures* (pp. 43–66). Cambridge: Cambridge University Press.

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

Goldman, A. (1992). In defense of simulation theory. *Mind and language*, *7*, 104–119.

Heider, F., & Simmel, M. (1944). An experimental study of apparent behavior. *American journal of psychology*, *57*, 243–259.

Jepson, A., & Feldman, J. (1996). A biased view of perceivers: commentary on Bennett, B. & Hoffman, D., *Observer theory, Bayes theory, and psychophysics*. In D. Knill & W. A. Richards (Eds.), *Perception as Bayesian inference* (pp. 229–235). Cambridge: Cambridge University Press.

Johansson, G. (1973). Visual perception of biological motion and a model for its analysis. *Perception & Psychophysics*, *14*(2), 201–211.

Johnson, S. C. (2000). The recognition of mentalistic agents in infancy. *Trends in Cognitive Sciences*, *4*(1), 22–28.

Johnson, S. C., Booth, A., & O'Hearn, K. (2001). Inferring the goals fo a nonhuman agent. *Cognitive Development*, *16*, 637–656.

Kaiser, M. K., & Proffitt, D. R. (1987). Observers' sensitivity to dynamic anomalies in collisions. *Perception & Psychophysics*, *42*(3), 275–280.

Kleene, S. C. (1956). Representation of events in nerve nets and finite automata. In C. Shannon & J. Mc-

Carthy (Eds.), *Automata studies.* Princeton: Princeton U. Press.

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

Lettvin, J. Y., Maturana, H. R., McCulloch, W. S., & Pitts, W. H. (1959). What the frog's eye tells the frog's brain. *Proceedings of the Institute of Radio Engineers, 47*(4), 1940–1951.

Lewis, H. R., & Papadimitriou, C. H. (1981). *Elements of the theory of computation*. Englewood Cliffs, New Jersey: Prentice-Hall, Inc.

Mann, R., Jepson, A., & Siskind, J. M. (1997). The computational perception of scene dynamics. *Computer Vision and image understanding, 65*(2), 113–128.

McCulloch, W. S., & Pitts, W. H. (1943). A logical calculus of the ideas immanent in nervous activity. *Bulletin of Mathematical Biophysics, 5*, 89–93. (Reprinted in W. S. McCulloch, "Embodiments of Mind"; Cambridge, MIT Press, 1965)

Pylyshyn, Z. (1999). Is vision continuous with cognition? the case for cognitive impenetrability of visual perception. *Behavioral and Brain sciences, 22*(3), 341–423.

Rock, I. (1983). *The logic of perception*. Cambridge: MIT Press.

Schmidt, C. F. (1976). Understanding human action: recognizing the plans and motives of other persons. In J. S. Carroll & J. W. Payne (Eds.), *Cognition and social behavior* (pp. 47–66). Hillsdale, New Jersey: Lawrence Erlbaum Associates.

Scholl, B. J., & Tremoulet, P. D. (2000). Perceptual causality and animacy. *Trends in Cognitive Science, 4*(8), 299–309.

Shiffrar, M., Lichtey, L., & Chatterjee, S. H. (1997). The perception of biological motion across apertures. *Perception & Psychophysics, 59*(1), 51–59.

Stewart, J. A. (1982). *Perception of animacy*. Unpublished master's thesis, University of Pennsylvania.

Stich, S., & Nichols, S. (1992). Folk psychology: Simulation or tacit theory? *Mind & Language, 7*(1), 35–71.

Stich, S., & Nichols, S. (1995). Second thoughts on simulation. In M. Davies & T. Stone (Eds.), *In mental simulation: Evaluations and applications* (pp. 87–108). Oxford: Basil Blackwell.

Thorpe, S., Fize, D., & Marlot, C. (1996). Speed of processing in the human visual system. *Nature, 381*, 3.

Tremoulet, P. D. (2000). *Inferring animacy from motion, form, and context cues*. Unpublished doctoral dissertation, Rutgers University.

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*(6), 1047–1058.

Ullman, S. (1979). The interpretation of structure from motion. *Proc. R. Soc. Lond. B., 203*, 405–426.