Cognition 130 (2014) 360–379

Contents lists available at ScienceDirect

Cognition

journal homepage: www.elsevier.com/locate/COGNIT

Inferring the intentional states of autonomous virtual agents

Peter C. Pantelis [a,*], Chris L. Baker [b], Steven A. Cholewiak [a], Kevin Sanik [c], Ari Weinstein [c], Chia-Chien Wu [a], Joshua B. Tenenbaum [b], Jacob Feldman [a]

[a] Dept. of Psychology, Center for Cognitive Science, Rutgers University – New Brunswick, United States
[b] Dept. of Brain and Cognitive Sciences, Massachusetts Institute of Technology, United States
[c] Dept. of Computer Science, Rutgers University – New Brunswick, United States

ARTICLE INFO

Article history:
Received 16 February 2013
Revised 16 September 2013
Accepted 13 November 2013

Keywords:
Theory of mind
Action understanding
Intention inference

ABSTRACT

Inferring the mental states of other agents, including their goals and intentions, is a central problem in cognition. A critical aspect of this problem is that one cannot observe mental states directly, but must infer them from observable actions. To study the computational mechanisms underlying this inference, we created a two-dimensional virtual environment populated by autonomous agents with independent cognitive architectures. These agents navigate the environment, collecting "food" and interacting with one another. The agents' behavior is modulated by a small number of distinct goal states: *attacking*, *exploring*, *fleeing*, and *gathering food*. We studied subjects' ability to detect and classify the agents' continually changing goal states on the basis of their motions and interactions. Although the programmed ground truth goal state is not directly observable, subjects' responses showed both high validity (correlation with this ground truth) and high reliability (correlation with one another). We present a Bayesian model of the inference of goal states, and find that it accounts for subjects' responses better than alternative models. Although the model is fit to the actual programmed states of the agents, and not to subjects' responses, its output actually conforms better to subjects' responses than to the ground truth goal state of the agents.

1. Introduction

Comprehension of the goals and intentions of others is an essential aspect of cognition. Motion can be an especially important cue to intention, as vividly illustrated by a famous short film by Heider and Simmel (1944). The "cast" of this film consists only of two triangles and a circle, but the motions of these simple geometrical figures are almost universally interpreted in terms of dramatic narrative. Indeed, it is practically impossible to understand many naturally occurring motions without comprehending the intentions that contribute to them: a person running is interpreted as trying to get somewhere; a hand lifting a Coke can is automatically understood as a person intending to raise the can, not simply as two objects moving upwards together (Mann, Jepson, & Siskind, 1997). Much of the most behaviorally important motion in a natural environment is produced by other agents and reflects unseen mental processes. But the computational mechanisms underlying the inference of mental states, including goals and intentions, are still poorly understood.

Human subjects readily attribute mentality and goal-directedness to moving objects as a function of properties of their motion (Tremoulet & Feldman, 2000), and are particularly influenced by how that motion seems to relate to the motion of other agents and objects in the environment (Blythe, Todd, Miller, & The ABC Research Group, 1999; Barrett, Todd, Miller, & Blythe, 2005; Tremoulet & Feldman, 2006; Zacks, Kumar, Abrams, & Mehta, 2009; Gao, McCarthy, & Scholl, 2010; Pantelis & Feldman, 2012). The broad problem of attributing mentality to others has received a great deal of attention in the philosophical literature (often under the term *mindreading*), and has been most widely

* Corresponding author. Address: Indiana University – Bloomington, Dept. of Psychological and Brain Sciences, 1101 E. 10th Street, Bloomington, IN 47405, United States. Tel.: +1 812 856 7800.
E-mail address: pcpantel@indiana.edu (P.C. Pantelis).

0010-0277/$ - see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.cognition.2013.11.011

studied in infants and children (Gelman, Durgin, & Kaufman, 1995; Gergely, Nádasdy, Csibra, & Bíró, 1995; Johnson, 2000; Kuhlmeier, Wynn, & Bloom, 2003). But the adult capacity to understand animate motion in terms of intelligent behavior has been less studied. Computational approaches to the problem of intention estimation have been scarce historically (for perhaps the earliest example, see Thibadeau, 1986), in part because of the difficulty in specifying the problem in computational terms. But new modeling approaches are emerging from various perspectives and disciplines in this rapidly-developing area of research (Feldman & Tremoulet, 2008; Baker, Saxe, & Tenenbaum, 2009; Crick & Scassellati, 2010; Kerr & Cohen, 2010; Pautler, Koenig, Quek, & Ortony, 2011; Burgos-Artizzu, Dollár, Lin, Anderson, & Perona, 2012).

Experimental stimuli in studies of the interpretation of intentionality from motion have, like the original Heider and Simmel movie, consisted almost exclusively of animations featuring motions crafted by the experimenters or their subjects to convey specific psychological impressions. Traditional psychophysics is then applied to relate attributes of the observed motion to the subjective impression produced (Blythe et al., 1999; McAleer & Pollick, 2008). While this method has yielded important insights, it suffers from certain critical limitations. Apart from the inefficiency of continual reliance on subjective intuition (e.g. via a subject pool) to generate new and varied stimuli scenes, handcrafted stimuli are opaque in that it is unclear exactly *why* the constituent motions convey the particular impressions they do, since they have been designed purely on the basis of the designers' intuitions—intuitions that are, in effect, the object of study. This makes it impossible to explore, for example, the relationship between observers' judgments of the agents' mental states and the true nature of the "mental" processes generating agent behavior. In this case, the independent and dependent variables are both direct reflections of subjective notions of what particular classes of behavior "should" look like.

Other studies have examined the perception of animate motion more systematically, either by varying the velocity and orientation of agents parametrically, or by manipulating parameters of simple programs generating agent behavior (Stewart, 1982; Dittrich & Lea, 1994; Williams, 2000; Tremoulet & Feldman, 2000, 2006; Gao, Newman, & Scholl, 2009; Gao & Scholl, 2011; Pantelis & Feldman, 2012). While this method avoids some of the aforementioned pitfalls of using handcrafted stimuli, our present study represents a substantial departure even from this approach. In the spirit of Dennett (1978)'s suggestion to "build the whole iguana," our goal was to create cognitively autonomous agents whose motions actually were, at least in a limited sense, driven by their own beliefs, intentions, and goals. To this end, we developed a 2D virtual environment populated with autonomous agents—virtual robots—who locomote about the environment under their own autonomous control, interacting with and competing with other agents in the environment. We refer to the agents as IMPs, for *I*ndependent *M*obile *P*ersonalities. Like agents in artificial life environments (e.g. Yaeger, 1994; Shao & Terzopoulos, 2007), IMPs have a complete, albeit severely restricted, cognitive architecture.

The IMPs can be understood to have one overall goal: to obtain "food" and bring it back to a home location. But at each time step, an IMP's behavior is modulated by its continually-updating "goal" state, which determines how it will respond to stimuli in the environment. An IMP can be in one of four discrete goal states: it can **explore** the environment, **gather** food, **attack** another agent, or **flee** from another agent (Fig. 2. These four states were loosely modeled on the "Four Fs" of animal ethology, action categories that are said to drive most animal behavior; see Pribram, 1960).

The agents obtain information about their environment via on-board perception, consisting of a simple visual module with a 1D retina (a perceptual ability reminiscent of that of the 2D characters in Abbott's (1884) novella *Flatland*). The agents progressively learn a map of their environment as they move about the environment. Lastly, the agents have a limited capacity to reason about how to accomplish their goals (for example, they can calculate the shortest path through the environment between their current location and a goal location). Thus the IMPs are complete, though crude, cognitive agents. Their observable actions are based entirely on what they "want", "know", and "think" about their environment.

The subjective appearance of IMP behaviors corresponding to their respective goal states are necessarily connected to the subjective intuitions of the programmers, but this connection is far more indirect than in the case of stimuli created via handcrafted animation. We can hardly predict how stimulus scenes will appear with any precision, given that the IMP subroutines connected with respective goal states execute within the complicated context of other modules in the IMP programming, and that these scenes are dynamically generated within multi-agent environments which are explicitly probabilistic. Manipulation of the parameters of particular IMP modules may have inherently unpredictable effects; for example, we have no strong and precise intuition about what it would "look like" if the resolution of an agent's vision were changed. Finally, whereas the semantics we attach to the various IMP goal states may be arbitrary and subjective (why "attack" and not "chase"?[1]), there is an objective ground truth to the existence of behavioral states contained in the IMP program, among which the IMP actually transitions, and each of which predisposes the IMP toward particular actions. There is therefore a ground truth basis for assessing subject's accuracy when they attempt to infer these underlying states.

In the studies below, we ask what human subjects can infer about the IMPs' goal states on the basis of observing them move about and interact within a sparse environment, and model how they might go about performing this inference. The appearance and behavioral repertoire of the IMPs are quite simple; they are rigid triangles which may only translate or rotate. This does not mean to imply that the perceptual features of these IMP stimuli exhaust the possibly important cues subjects may use to make

[1] For more on the semantics subjects attach to IMP behavior without being first supplied with our labels, see Pantelis et al. (2011).