"Quis custodiet ipsos custodes?"[1] Artificial Intelligence and the Interactionist Stance

Nick DePalma
Personal Robots Group, MIT Media Lab
20 Ames Str. E15-468
Cambridge, MA 02139
ndepalma@mit.edu

Abstract

The lure of understanding biological intelligence has long occupied researchers. Success has always been measured in peer review, number of citations, or how influential some piece of work is in inspiring the next generation of researchers. What human-robot interaction (HRI) and artificial intelligence (AI) promises is a metric of believability that is not intrinsic to the values of the researcher or community of practice but to the utility and successful function of the robotic artifact within a larger society. This paper is a reflection and response to the hypothesis that HRI is a pure, fundamental art of artificial intelligence and the last great successor to a domain fraught with the trappings of an art that lost its way.

Introduction

What is intelligence? A high score on the SATs or GREs? Could you identify another person who is intelligent? Who is more intelligent than you? Cue the academics taking complex stances regarding cognitive architectures, probabilistic methods, learning, logic, knowledge modeling, and other such esoterica. Much of popular, present day, AI research places the core values of the academic community around the quantitative performance of the agent's planning or learning algorithms, pitting algorithm against algorithm in an artifactual quest of champions. Rather than placing the onus of success on recall, precision, or optimality, our community, in its purest, places measures of success on the fluency and adaptability of a socially embedded agent, perhaps the only way of quantifying subjective perspectives of intelligence. Brooks' work on behavior based AI (1991) was key in this evolution. It was some of the first work that attempts to redefine artificial intelligence, rebelling against the status quo, as an emergent property that is not well understood. Rather than a single, centralized, saturated decision making routine, emergent behavior was a property of the confluence of a number of internal competing systems. When your robot is interacting with a passive environment like a manufacturing floor, you value the robot that does a specific task most efficiently and optimally, saving time and money. Instead, when your robot must interact with an everyday user, a robots' ability to adapt and improvise in new situations behaviorally becomes one of the most important aspects of its response to the user. Regardless of how optimal or quantitatively superior your results may be, you still must contend with a finite state machine that produces a broader range of behaviors that the user cannot perceive the difference between. This is a common challenge in agent literature as well. Systems like Façade (Mateas and Stern 2003) provide a superior interaction with the agent at the cost of brute force programmer time.

I focus this paper on a number of arguments to set the record straight and inspire this community to take a step out on its own wholly separate from AI and robotics. The arguments I intend to contest are the following:

- Innovation comes by borrowing techniques from other domains such as learning, planning, vision, or the like. We are merely system integrators.
- Humans dont provide us with the fidelity of response we need to make reasonable modifications to our intelligent agent architectures requiring us to perform large human studies.

I will conclude with a list of challenges to the community that could have a large impact on both the AI and the HRI community.

(Un)Bounding Artificial Intelligence: AI-HRI as a community of practice

Lave and Wenger (1991; 1998) roughly define communities of practice as communities that have their own sets of values, knowledge, and practices. Some researchers bridge multiple communities of practice, seamlessly transitioning between each community (say HRI and AI), holding different power relationships between each community of practice. My suggestion is to define a set of goals that we as a community must accomplish on our own, separate from the other communities whose values are motivated by other objectives. Presently, we as researchers have focused on measuring qualitative as well as quantitative factors during an interaction, in essence having multiple masters. This is untenable. The behaviors that emerge from our systems and algorithms become the key to keeping the interaction moving.

[1] "Quis custodiet ipsos custodes?" translated is "Who watches the watchmen?"

My proposal is to leverage this advantage of our domain to the fullest and focus primarily on interaction factors, leaving the details of the algorithms to be debated philosophically instead of quantitatively.

This community has many objectives but they frequently lead to one vision of the future: that robots should benevolently exist among us, cooperating, understanding our intentions, reasoning effectively about socio-cultural protocols, and communicating effectively with us. This opportunity allows us to focus on architectures that support short term and long term interactions, that must remember previous interactions, that can communicate effectively about mixed representations like events, states, actions, objects, affordances, etc. and that can communicate using verbal *and* nonverbal cues. This community is, at its purest, a community beholden to the user, not a community of practice. We shouldnt be afraid to dip our feet into the various fragments of AI that arbitrarily came apart at its inception (for a review, see Brooks 1991) to create a more believable agent.

A truly believable agent must succeed on an incredibly large number of metrics. This is more than system integration but a tightly controlled architecture that is able to reason about complex relationships between the subtle actions of the user and the intention of what is or should be communicated. Our job as believable agent designers should be synonymous with building an artificially intelligent agent that is able to understand how we intend to communicate with it.

Introducing the Layman: Doing away with going "off script"

A layman is the person who ruins your results, your expectations, and forces you to explain that "no, the robot wasnt designed to handle that." There is a fine line in our community between science and engineering. Standards in modern HRI science revolve around defining an experiment, engineering a solution that you are (reasonably) sure will work, and then showing that it does or doesnt work for some set of behaviors. The challenge that those who bridge the AI and HRI must accept is that these pieces of the AI puzzle must speak for themselves in how they fit together for us to make long-term progress. We must not accept other communities narratives of how their pieces may fit with our goals unless they truly do.

AI-HRI researchers must fundamentally accept this fine line they play between engineering and science. We are used to engineering a solution to a problem we define ourselves instead of allowing the users to speak for themselves. Scripts, acceptable key phrases, permissible and impermissible behaviors bely the need for a truly adaptable behavior system situated within true, unstructured social interaction. We must listen to the users that interact with our robots and define what it means to be a believable agent, even if it means redrawing the AI landscape for our needs. There is a lot of value in both the AI and the HRI community separately but AI and HRI coming together truly means is to define a whole new academic sport. A sport true to its origin; one that understands we are engineers *and* scientists, that we should be more like anthropologists than psychologists, and that the user is always right regardless of how much we want to validate our piece of the puzzle.

The key challenge we have as a community is to define our vision of socially adept robotics going forward. We have become increasingly adept in the HRI community at "setting expectations," "bounding the problem," and, in general, getting the results we want or know how to get. We must eschew these practices and we must listen to the user. Negative results and unexpected interactions should be a point of pride rather than a Scarlet Letter. As such, we should reward ideas and perspectives, not results. After all, we are pushing the envelope and making new observations, arent we?

For us to define AI and HRI as a subject in its own right, we must define algorithms that reason about affect, that understand nonverbal behavior, that reason about anothers mind, that understands implicature, backchanneling and intentionality, and that can navigate social norms and values. These challenges will require fundamentally creative thinking, fresh reasoning models, new forms of action selection and inference and, more than anything, *an open mind* within the community.

To structure more architectural questions, we must face the idea of *long-term interaction* and *open domain interactions*. Robots that face people in everyday situations remind us how the pieces fit together to create autonomous systems and remind us how far we have left to go, not how much we should control through expectation setting.

Who should watch over the researchers in this community? The users should. And that is what will set us apart.

Acknowledgements

This research was supported by the National Science Foundation (NSF) under Grants CCF-1138986 and the MIT Media Lab Consortia. Any opinions, findings and conclusions, or recommendations expressed in this paper are those of the authors and do not represent the views of the NSF or the MIT Media Lab.

References

Brooks, R. A. 1991. Intelligence without representation. *Artificial intelligence* 47(1):139–159.

Lave, J., and Wenger, E. 1991. *Situated learning: Legitimate peripheral participation*. Cambridge university press.

Lave, J., and Wenger, E. 1998. Communities of practice. *Retrieved June* 9:2008.

Mateas, M., and Stern, A. 2003. Façade: An experiment in building a fully-realized interactive drama. In *Game Developers Conference*, 4–8.