Socially Situated AI: What It Means and Why It Matters

Phoebe Sengers
Department of Computer Science and
Program in Literary and Cultural Theory
Carnegie Mellon University
Pittsburgh, PA 15213
phoebe.sengers@cs.cmu.edu

Abstract

Classical AI focuses agent design purely on the agent. Various alternatives to classical AI, such as behavior-based AI and situated action, propose that agent design should include, not just the agent, but also its environment. These alternatives fall short of their own stated goals, however, in that they tend to include only the physical objects with which the agent interacts, leaving out such environmental factors as the audience with which the agent interacts, the people who are judging the agent as a scientific success or failure, and the designer of the agent him- or herself. Here, I argue that ignoring the social and cultural context of the agent actually leads to technical problems in the design of the agent, and propose a model for AI research that includes the full context of the agent. At the same time, I propose that the problems socially situated AI addresses are particularly pressing for applications in AI and entertainment, where the utility of an agent depends not so much on the internal correctness of the agent's actions as on the effect of the agent on the user.

Introduction

Imagine a composer writing a score without knowing which of the notes the audience will actually hear; a writer creating a poem who does not know which of its words the audience will read; or a director making a film without knowing which images the audience will see. We would consider these people to be hopelessly hamstrung in their tasks, fortunate if any of their original conception gets across; but as agent architects we put ourselves in essentially the same situation every day. When building agents for entertainment applications, we are interested in the effect the agent will have on the audience; but most existing agent paradigms simply focus on the agent and its tasks, making the eventual effect on the audience more or less out of our control.

Specifically, if we use classical AI paradigms (e.g. [8]), we focus design purely on the agent itself; if we use alternative AI (i.e., such alternatives to classical AI as situated action, ALife, and behavior-based AI, e.g. [1], [3], [5]), we focus design on the agent and its physical environment. Neither of these paradigms let us build an agent with the design focused on what is essential in entertainment applications, i.e. the information we as designers intend to communicate to the audience. Here, I will propose an alternative paradigm for autonomous agent research called *socially situated AI* which includes the social and cultural context of the agent, thereby allowing designers to program agents with respect to the audience's *perception* of the agent. The relation between these paradigms can be seen in figure 1. This work is inspired by recent work in believable agents such as [6] [4] [9] [2], which focus more and more on the audience's perception of agents, rather than on an agent's correctness per se.

Why Environments Matter

One of the major commitments of alternative AI is that design should be oriented, not towards the agent alone, but towards the dynamics of an agent with respect to its environment. There are two major advantages to this:

1. The agent may be simpler to build, once one can take specific properties of the environment into account.
2. The agent may perform better because it is designed specifically for the environment in which it is being used.

These advantages may also be important for entertainment applications. For these uses it is important to note that it is not enough that the agent perform its tasks well with respect to a physical environment. For instance, it is much more important that an agent's behavior be clear and understandable than that the agent actually do the 'right' thing, where 'right' is defined with respect to some kind of optimal problem-solving behavior in the world. To put it simply, a stupid but understandable agent is more appropriate than a hyperintelligent but completely incomprehensible one.

In order for our agents to be understandable, we have to have some idea of how the audience will react to the agent. This means that for entertainment applications the agent needs to be designed to behave well not only for a physical environment but also with respect to its its *social* environment (the audience) and its *cultural* environment (cultural norms that influence behavior and its inter-

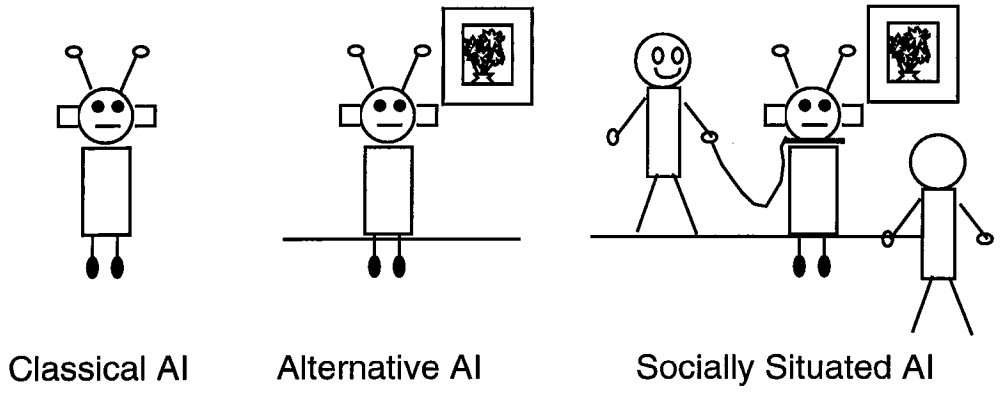

Figure 1: Gradually including more of the agent's context

pretation). To be appropriate for entertainment, the agent needs to be reasonably *comprehensible* to its audience and to *communicate* through its behaviors the sorts of things the designer is trying to get across.

Why Alternative AI Is Not Enough

Knowing what the audience finds understandable and what sorts of things the agent communicates to them is hard when the design of the agent focuses only on the agent itself, as in classical AI. Alternative AI, on the other hand, seems perfect for such an approach because it puts a great emphasis on the fact that an agent can only be thought of in terms of an environment. However, for alternativists the 'environment' of an agent, generally speaking, only includes its physical environment. Often, there are explicit attempts to exclude the social and cultural environment of the agent. Tim Smithers, for example, bewails the fact that the very term 'agent' has cultural connotations. "The term 'agent' is, of course, a favourite of the folk psychological ontology. It consequently carries with it notions of intentionality and purposefulness that we wish to avoid. Here we use the term divested of such associated baggage" [7]. The language Smithers uses reflects a prevailing belief in AI that the social and cultural environment of the agent - unlike its physical environment - is simply so much baggage to be discarded.

At first blush, this is a little strange. If the environment of the agent is so important in alternative AI, why are big chunks of it being left out of agent design altogether? I suspect that this is related to the fact that alternative AI, especially ALife, is often trying to model agents scientifically. Without thinking about it, researchers often fall into a mode I term 'naive objectivity', in which any mention of the social and cultural context is considered to taint the scientificity of the resulting agent. Researchers end up artificially narrowing the context of the agent to exclude the necessarily subjective parts of the environment, i.e. the person who built the agent and the person judging it (see figure 2).

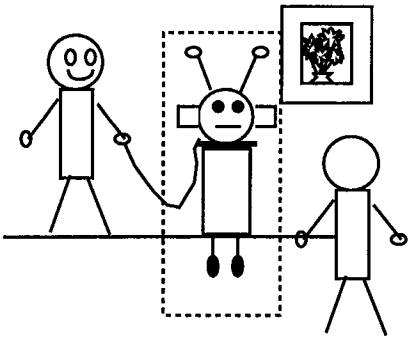

Figure 2: Alternative AI excludes the full environment.

Leaving out the social context becomes problematic for entertainment applications, because in these applications it is precisely the communication between the designer and the audience that is important. Agent architectures that force you to design the agent only with respect to its behaviors, and not with respect to how those behaviors are communicated to and understood by the audience, are not giving you enough power to program the things that really count in entertainment applications.

Excluding the Full Context Causes Confusion

Leaving out the full context is not just bad because you lose power. It actually may make it harder to solve technical problems that come up with the development of agents. This is because simply ignoring the social and cultural context does not make it go away. What ends up happening is that, by insisting that cultural influences are not at work, those influences come back through the back door.

As an example, consider the use of symbolic pro-

gramming. Alternativists often recognize that symbolic programming of the kind classical AI engages in is grounded in culture. Some of them believe that by abandoning symbolic programming they, unlike classicists, have also abandoned the problem of cultural presuppositions creeping into their work. And in fact it is true that many alternative AI programs do use symbols sparingly, if at all, in their internal representations.

Nevertheless, it would be fair to say that the architecture of such agents involves symbols to the extent that the engineer of the agent must think of the world and agent in a symbolic way in order to build the creature. For example, the creature may have more or less continuous sensors of the world, but each of those sensors may be interpreted in a way that yields, once again, symbols - even when those symbols are not represented explicitly in an agent's program. The behaviors the agent is split up into are also fundamentally symbolic ("play fetch," "sleep," "beg," etc.). So while alternative AI has gotten away from symbolic representations within the agent, it has not gotten away from symbolic representations altogether. Once you look at the entire environment of the agent, including its creator, it is clear that despite the rhetoric that surrounds alternative AI symbols still play a large role.

Leaving out the social context is both epistemologically inadequate and obfuscating. By not looking at the subjective aspects of agent design, the very nature of alternative AI programming, as well as the origin of various technical problems, becomes obscured. This is particularly problematic because not being able to see what causes technical problems may make them hard, if not impossible, to solve.

Socially Situated AI Defined

What should alternative AI do instead? In this section, I will lay out the basic premises of socially situated AI. In the next section, I will explain how this can be applied to the problem of building autonomous agents for entertainment applications.

1. *An agent can only be evaluated with respect to its environment, which includes not only the objects with which it interacts, but also the creators and observers of the agent.* Autonomous agents are not 'intelligent' in and of themselves, but rather with reference to a particular system of constitution and evaluation, which includes the explicit and implicit goals of the project creating it, the group dynamics of that project, and the sources of funding which both facilitate and circumscribe the directions in which the project can be taken. An agent's construction is not limited to the lines of code that form its program but involves a whole social network, which must be analyzed in order to get a complete picture of what that agent is, without which agents cannot be meaningfully judged.
2. *An agent's design should focus, not on the agent itself, but on the dynamics of that agent with respect to its physical and social environments.* In classical AI, an agent is designed alone; in alternative AI, it is designed for a physical environment; in socially situated AI, an agent is designed for a physical, cultural, and social environment, which includes the designer of its architecture, the creator of the agent, and the audience that interacts with and judges the agent, including both the people who engage it and the intellectual peers who judge its epistemological status. The goals of all these people must be explicitly taken into account in deciding what kind of agent to build and how to build it.
3. *An agent is a representation.* Artificial agents are a mirror of their creators' understanding of what it means to be at once mechanical and human, intelligent, alive, a subject. Rather than being a pristine testing-ground for theories of mind, agents come overcoded with cultural values, a rich crossroads where culture and technology intersect and reveal their co-articulation.

Socially Situated AI for Entertainment

All these abstract premises are well and good, but what does this actually mean for the design of autonomous agents? Currently, autonomous agents built using the alternative AI approach are often split up into behaviors. These behaviors are then combined in order to allow the agent to achieve various goals in the world. I propose that these behaviors should be designed, not just in terms of fulfilling the internal goals of the agent, but in terms of what the agent is communicating to the audience. Behaviors must be programmed with enough observable actions that the audience for which the work is designed will actually be able to tell that the agent is engaged in that behavior. This has a number of implications.

First of all, behaviors cannot be thought of purely in terms of the agent's goals, but must also be observable to the audience through external signs. It is not enough to just do something; the audience must be able to tell the agent is doing it. This means a behavior, generally speaking, includes the intention to communicate that behavior to the audience. 'Behaviors' in this paradigm therefore become something more like 'meaningful units of action' than the a priori, universal modes of behavior in most behavior-based AI applications.

Also, the behavior must be designed with a concrete audience in mind. For example, an 'insult-

ing' behavior may be very different for a Japanese than an American audience; 'macho' behavior may come across differently for a working class or an academic audience. In general, behaviors do not have meaning in and of themselves; they take on a variety of meanings in a variety of contexts. In order for the agent to be effective in the context in which it is inserted, it is important that the designer consider this context when building the agent.

Socially situated agents therefore incorporate the same kinds of limitations and advantages for social environments as agents situated in a physical environment. Yes, they will only function properly for the audience for which they were designed; on the other hand, in that context they may be simpler to build and may perform better precisely because they are designed specifically for the environment in which they are being used.

Conclusion

Socially situated AI, while inspired by and an outgrowth of alternative AI, ends up proposing quite a different model of what it means to build and be an agent. While alternative AI sees the agent as existing in a cultural vacuum, socially situated AI explicitly puts the agent in its larger social context in order to make agents that can function more effectively in practice. This means agents are designed with *communicable*, rather than a priori, behaviors; and that agents can be designed for the specific contexts in which they will function, rather than being hamstrung by an impossible demand to be culturally universal. Rather than "do the right thing", focusing on a problem-solving approach, we want to "do the thing right," behaving in a way that makes sense to an audience. These changes are especially important for entertainment applications, where designers would like to be able to have reasonably predictable effects on their audiences, rather than taking the shot-in-the-dark approach necessitated by agent architectures that cannot take the audience's perceptions into account.

Acknowledgement

This work was done in the context of Joseph Bates's Oz Project. It was supported by the Office of Naval Research under grant N00014-92-J-1298.

References

[1] Bruce Blumberg. Action-selection in hamsterdam: Lessons from ethology. In *Proceedings of the 3rd International Conference on the Simulation of Adaptive Behavior*, Brighton, 1994.

[2] Bruce Blumberg and Tinsley A. Galyean. Multi-level direction of autonomous creatures for real-time virtual environments. In *Proceedings of SIGGraph*, 1995.

[3] Rodney Brooks. A robust layered control system for a mobile robot. *IEEE Journal of Robotics and Automation*, RA-2:14–23, April 1986.

[4] Bryan A. Loyall and Joseph Bates. Hap: A reactive, adaptive architecture for agents. Technical Report CMU-CS-91-147, Carnegie Mellon University, 1991.

[5] Pattie Maes. How to do the right thing. AI Memo 1180, MIT AI Laboratory, December 1989.

[6] Scott Neal Reilly. Believable social and emotional agents. Forthcoming Ph.D. thesis.

[7] Tim Smithers. Taking eliminative materialism seriously: A methodology for autonomous systems research. In Francisco J. Varela and Paul Bourgine, editors, *Towards a Practice of Autonomous Systems: Proceedings of the First European Conference on Artificial Life*, pages 31–47, Cambridge, MA, 1992. MIT Press.

[8] Steven Vere and Timothy Bickmore. A basic agent. *Computational Intelligence*, 6:41–60, 1990.

[9] Peter Wavish and Michael Graham. A situated action approach to implementing characters in computer games. *AAI*, 10, 1996.