AgentSmith: Exploring Agentic Systems

David Miller
Stanford University
Stanford, CA 94305, USA
davebmiller@stanford.edu

CHI'16 Extended Abstracts, May 07-12, 2016, San Jose, CA, USA
ACM 978-1-4503-4082-3/16/05.
http://dx.doi.org/10.1145/2851581.2859025

Abstract
The design of systems with independent agency to act on the environment or which can act as persuasive agents requires consideration of not only the technical aspects of design, but of the psychological, sociological, and philosophical aspects as well. Creating usable, safe, and ethical systems will require research into human-computer communication, in order to design systems that can create and maintain a relationship with users, explain their workings, and act in the best interests of both users and of the larger society.

Author Keywords
Agentic Systems; Automation; Ethics; Communication; HCI

ACM Classification Keywords
H.1.2. Human Factors, Software Psychology.
I.2.0. Artificial Intelligence: Philosophical Foundations.
I.2.9 Robotics: Autonomous vehicles. K.4.1. Public Policy Issues: ethics, human safety, privacy, use/abuse of power.

Context and Motivation for Research
The theme of CHI'16 is "CHI for Good." My research is motivated by an underlying concern for sustainability, safety, and health. The proliferation of agentic systems

Figure 1. Agentic systems vary in communicativity and physical agency. These dimensions can be considered in a two-dimensional space to classify agentic systems.

that have the ability to act on the environment directly, and to exert influence over humans, portends great promise but these systems also present significant risks of misuse both by users and by other entities such as designers, manufacturers, and regulators. My goal is to explore the psychology of interaction with agentic systems, as well as the sociological and philosophical issues inherent in work in this area.

Background

Agentic systems can be described as having independent agency to influence the world directly or which interact with humans and can thus act as persuasive technologies [3]. I derive the term "agentic system" from the work of Bandura [1] and Mischel [10], where the (human) individual is described as an autonomous, agentic entity. Computers, robots, and virtual agents can exert agency similar to humans, even if at this time their abilities are very different. Many of the differences in the way humans consider interaction with computers, on both an interaction level and a legal-regulatory level inform design, even when the ways we nonconsciously interact with interactive systems is akin to interaction with other humans [13]. Designing systems with independent agency that can act in the stead of humans (for example, automated driving or automotive safety systems, aviation systems), or in a persuasive interactive capacity (such as search or conversational agents), will require a consideration of how these systems communicate, as well as how they act directly on the physical world, see Figure 1.

Communication design is a critical issue for the design of agentic systems; it is important for systems to create, maintain, and repair the relationship with a human user, so that the two can act as an effective team, and so that humans can form an accurate and usable mental model [11] of the system, this will allow users to employ automation properly, especially when the state of the system changes or its capabilities are degraded.

This human-computer relationship exists over time, and as a result, the two entities will build an understanding of each other, the system taking in information through communication with the user(s), through observation of users using biosensing technologies, and by sensing the ambient environment. This relationship must be maintained and strengthened, and at times repaired if the system breaks the user's trust, even if for necessary reasons. As the relationship can extend over a long period, this relationship management will be of high importance, so that the user appropriately trusts [8] the system, knows its limits and capabilities, and avoids misuse [12]. The computer must also know the limits of the human's performance, so that the computer will not demand more of the user than he or she can provide.

Competing Interests

How should an agentic system incorporate various inputs from direct user communication, sensing of the user, sensing of the environment, and signals from other systems, in order to form an optimal model for determining actions (see Figure 2)? This is a technical problem, but beneath it lies another set of issues: how should the system value the desires of the user, those of regulators, those of the manufacturers, and local and contemporary social norms? Does the relationship between a human and an agent become a special and privileged relationship—in terms of the allegiance of the

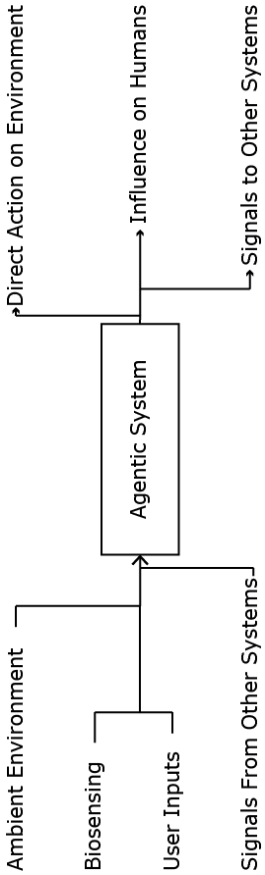

Figure 2. Agentic systems take inputs from human users and from the environment, and can communicate with other systems, influence humans, and can act directly on the environment

system? In the case of a highly automated vehicle, should the system prefer to advance the interests of the driver or occupants over those of other road users? What would a human do in this case, and should machines emulate human behavior?

Considering the balance of potentially competing interests (e.g. the interests of the user, the interests of designers and manufacturers, the interests of regulators, the interests of other members of the public), the role of explanation expands in importance. The agent must explain how these are balanced, and how these interests can be and should be balanced, if it is possible for the parameters to be adjusted (see Figure 3).

Persuasion

Persuasive technologies, defined by Fogg [3] as systems designed specifically to influence beliefs and behaviors, raise specific ethical concerns. As these systems can influence one's behavior, desires, and beliefs, the ethics espoused by the system (set in part by the designers, regulators, and manufacturers) will act with potentially great power. Ethical issues have been discussed by Fogg, among many others; and to some degree the ethical considerations raised by persuasive systems fall under the umbrella of value-sensitive design [4]. With an environmentally reactive and user-configurable system, the issues extend beyond classical value-sensitive design.

Whose values should the system champion? Those of the system designers? Those of society at large? Those of the user? How should these different positions be balanced, especially as the human-computer relationship changes over time and the environment varies? Experimental inquiry to determine what the issues are, and how to incorporate them into design will be necessary steps to take to understand this area, and to promote successful and ethical design of agentic systems in multiple areas.

Maintenance and Repair of the Relationship

The human-computer relationship may be a strong one in many cases (consider the bond between Luke Skywalker and R2-D2, or between most people and their phone), but as systems increase in agency and improve their communication abilities, this relationship will necessarily change. Where now computers are often insufficiently communicative, systems that exert significant independent agency and may act against a user's expectations or desires will have to communicate to explain their actions, and to describe the inputs and the computer's interpretation of them such that the user can form an accurate mental model, and thus predict what the system can do and is likely to do in various circumstances.

Explanation

Say you are driving in a highly-automated vehicle and the computer determines the best way to evade a crash is to steer around the obstacle, while you choose to brake (not realizing there is a large truck right behind you). The computer has disregarded your input in that situation, and subsequently it must explain why it performed the actions it did, effectively relieving you of command authority, if for good reason.

In the case of persuasive agents, it may be that the system must obscure or conceal information in order to help you achieve a goal which you have explicitly set. For example, if you have the goal of avoiding

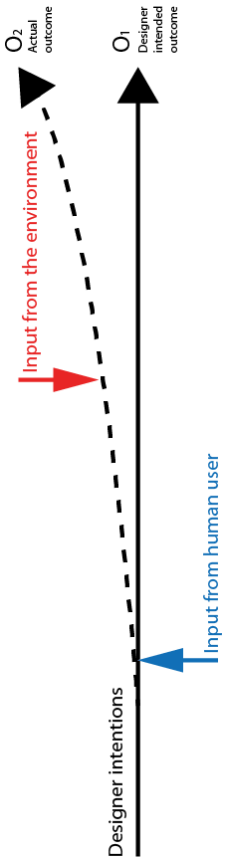

Figure 3. The actions of reactive systems are a function of designer intents, user behavior, and inputs from the environment.

temptation by ice cream, the system may elect to guide you on a less efficient route to avoid the weakness you previously set the system to protect you from, and it may not be able to ask you immediately ahead of time, as that would prime you and potentially sway you towards your weakness. The system would then have to explain, post-hoc, why it took a longer route, and that it acted in your best interests, in the service of your own superordinate desires, and in accordance with your prior wishes.

Research Questions

1. How should an agentic system communicate its knowledge and intents to humans?
2. How should systems integrate various data streams to create proper basis for actions directly on the environment and actions influencing humans?
3. How should an agentic system share control of a system such as a car—how should the system take control from a human, and how should it transfer control to a human operator? How can such a system negotiate, especially in a time pressured situation?

Research Approaches

Virtual reality simulation can be used as a tool for psychological research [2], with validity dependent on the presence level of the simulation [9] and on the specifics of the application. High presence VR provides the ability to perform research that cannot be effectively conducted in reality, is too dangerous to conduct in a real environment, or where only a virtual environment can afford appropriate repeatability.

Research Methods and Contributions

To this point, my research has explored critical issues in automated driving, transitions between automated and driver control, as well as studying trust in automation, mental models of automated systems, and situation awareness with partial automation.

My future research agenda includes experimental research on human ethical decisionmaking in simulation, and the study of how to design communication models for agentic systems that work collaboratively with humans. This knowledge can form a foundation from which to build arguments about what agentic systems should do, in terms of incorporating human inputs into the decision making process, both in the design and creation phases, through the interactions with users, and in the moment. This research extends experimental ethics and moral psychology (e.g. [5,6]).

Acknowledgments

I would like to thank my advisers Byron Reeves and Wendy Ju; Jeremy Bailenson, Fred Turner, Göte Nyman and Kiisa Hulko-Nyman for their assistance in developing my research agenda. I also owe a great deal to my colleagues at the Center for Design research and in the Department of Communication for helping operationalize my research. My research is supported by a Stanford Interdisciplinary Graduate Fellowship, funded by Coleman F. Fung.

References

1. Albert Bandura. 2001. Social Cognitive Theory: An Agentic Perspective. *Annual Review of Psychology* 52, 1: 1–26. http://doi.org/10.1146/annurev.psych.52.1.1

2. Jim Blascovich, Jack Loomis, Andrew C. Beall, Kimberly R. Swinth, Crystal L. Hoyt, and Jeremy N. Bailenson. 2002. Immersive Virtual Environment Technology as a Methodological Tool for Social Psychology. *Psychological Inquiry* 13, 2: 103–124. http://doi.org/10.1207/S15327965PLI1302_01

3. B. J. Fogg. 2003. *Persuasive technology: using computers to change what we think and do*. Morgan Kaufmann Publishers, Amsterdam; Boston.

4. Batya Friedman. 1996. Value-sensitive Design. *interactions* 3, 6: 16–23. http://doi.org/10.1145/242485.242493

5. Jonathan Haidt. 2001. The emotional dog and its rational tail: a social intuitionist approach to moral judgment. *Psychological review* 108, 4: 814.

6. Guy Kahane. 2011. The armchair and the trolley: an argument for experimental ethics. *Philosophical Studies* 162, 2: 421–445. http://doi.org/10.1007/s11098-011-9775-5

7. Jeamin Koo, Jungsuk Kwac, Wendy Ju, Martin Steinert, Larry Leifer, and Clifford Nass. 2014. Why did my car just do that? Explaining semi-autonomous driving actions to improve driver understanding, trust, and performance. *International Journal on Interactive Design and Manufacturing (IJIDeM)* 9, 4: 269–275. http://doi.org/10.1007/s12008-014-0227-2

8. John D. Lee and Katrina A. See. 2004. Trust in Automation: Designing for Appropriate Reliance. *Human Factors: The Journal of the Human Factors and Ergonomics Society* 46, 1: 50–80. http://doi.org/10.1518/hfes.46.1.50_30392

9. Kwan Min Lee. 2004. Presence, Explicated. *Communication Theory* 14, 1: 27–50. http://doi.org/10.1111/j.1468-2885.2004.tb00302.x

10. Walter Mischel. 2004. Toward an Integrative Science of the Person. *Annual Review of Psychology* 55, 1: 1–22. http://doi.org/10.1146/annurev.psych.55.042902.130709

11. Donald A. Norman. 1983. Some observations on mental models. *Mental models* 7, 112: 7–14.

12. Raja Parasuraman and Victor Riley. 1997. Humans and Automation: Use, Misuse, Disuse, Abuse. *Human Factors: The Journal of the Human Factors and Ergonomics Society* 39, 2: 230–253. http://doi.org/10.1518/001872097778543886

13. Byron Reeves and Clifford Ivar Nass. 1996. *The media equation: how people treat computers, television, and new media like real people and places*. CSLI Publications; Cambridge University Press, Stanford, Calif.; New York.