Giving AI a Theory of Mind

Bobbie Lynn Eicher
Design & Intelligence Laboratory, School of Interactive Computing
Georgia Institute of Technology, Atlanta, GA 30332
beicher3@gatech.edu

Abstract

Effective collaboration between humans and artificially intelligent agents will require that the two are equipped to build a sense of mutual understanding with each other. When humans have an intuitive understanding of the motives and intentions of other humans, it is known as Theory of Mind. My work revolves around designing artificial intelligence to leverage this capacity to improve human collaborations with artificial agents.

Area Background and Central Question

My work sits at the intersection of cognitive science and artificial intelligence. The central question of my work is how we can go about building AI systems that are able to understand how the human mind works, and to use this knowledge to better collaborate with human uers.

Theory of Mind

In essence, Theory of Mind represents the capacity that neurotypical adult humans, and some other animals, have that allows them to develop accurate mental models of what others intend and believe at any given time (Premack and Woodruf 1978). It's possible to measure this capacity for a given individual by testing their ability to interpret the emotion of a human in a photograph that shows only the eyes (Baron-Cohen et al. 2001).

Among humans, this capacity is critical to effective collaboration within groups to complete tasks (Woolley et al. 2010). This is true even when the participants cannot rely on seeing each other's faces, because the collaboration is purely virtual (Engel et al. 2014).

Existing Work

The work that I have done so far focuses on a mixture of identifying and coping with human misconceptions about how computer systems work, as well as examining the ethical implications of this work.

Virtual Teaching Assistant: Jill Watson

The idea behind the Virtual Teaching Assistant (VTA) is to address the difficulties of providing support for very large courses, particularly those offered online. We found that larger online courses members of our lab were involved with running could generate six times as much traffic as the typical course run at Georgia Tech's Atlanta campus (Goel and Joyner 2016).

The VTA project aims to create an artificially intelligent teacher that can monitor the course forums, recognize common questions raised by the students, and provide answers both accurately and quickly (Goel and Polepeddi 2017). I joined the project in 2015 and took over leadership on building and operating it in 2017. My work with it has focused on analyzing the ethical issues of operating an experiment in artificial intelligence in a real university course, while also working to make the behavior of the system more nuanced and understandable by the human students who interact with it.

International Conference on Computational Creativity 2017

For the ICCC conference in 2017, I presented a paper titled "Toward Mutual Theory of Mind as a Foundation for Co-Creation" to the workshop on co-creation and a poster titled "Modeling Student Misunderstandings: A Tool for Human-Computer Collaborative Learning of Introductory Programming" for the conference poster session. Both of these presentations were based on a group class project for an Intro to Cognitive Science course in Spring 2017, and

were co-authored with Kathryn Cunningham, Marissa Gonzales, Sydni Peterson, and Ashok Goel.

The project and poster resolved around taking existing research that catalogued the kinds of misconceptions students in introductory programming courses have about the way that assignment statements work (Ma 2007). We framed this in terms of the notional machine, which is the mental model of the way a computer functions that students need to construct in their own minds as part of the learning process to properly predict what program code will do (Sorva 2013).

Our project was designed to take the misconceptions about assignment statements that had already been noted to be common, and to simulate the behavior of code as if that misconception was correct. This way, we had a model of the student's flawed notional machine. We also built a version that executed the correct notional machine. Then, we created a tool that could execute the two side-by-side on small chunks of program code with the goal of being able to help students correct their own flawed notional machine by observing exactly how and where their expectations differed from the correct behavior. The prototype also had a diagnostic mode where it could take in the values that a student expected from a program, predict which common misconception the student had, and use that information to show them where they were going wrong.

In the paper I presented to the workshop on co-creation, I extended this idea into a larger argument about how Theory of Mind should actually be thought of as a mutual process rather than one that two entities engage in separately. This was exemplified by the way that our software wouldn't just tell a student they were wrong, but would actually attempt to understand what their thought process must look like, and then provide them with the information necessary to make corrections (Eicher et al. 2017). This argument that Theory of Mind as a collaborative process between two entities, and that this applies even when one of the entities is artificial, is the primary contribution of the paper and I believe that this framework makes it easier to think about the ways that agents need to account for human cognition as they perform their tasks.

AIES 2018

My submission to AIES 2018, "Jill Watson Doesn't Care if You're Pregnant: Grounding AI Ethics in Empirical Studies", is both an examination of the ethical implications of building a virtual teaching assistant for courses, as well as an argument that empirical observations should be given more attention in the larger ethical issues of developing AI. The paper was co-authored by Lalith Polepeddi and Ashok Goel. We were able to leverage our experiences with doing AI research with actual students to contribute a meaningful exploration of what it means to build artificial intelligence in an educational setting for use in real tasks.

Future Work

My research goal is to continue using the findings of cognitive science to inform the way that artificial agents relate to humans, as well as how they understand the intentions of humans who are attempting to collaborate with them. I want the agents that help us with daily tasks to be able to indicate whether and why it is confused with a request, and at the same time to be able to make reasonable inferences about what the user most likely intended so that it can provide coaching on how to improve mutual understanding and communication.

I also have hopes of continuing my work in the intelligent tutor area as presented at ICCC. I foresee systems that not only help us to understand themselves, but also guide us in better understanding ourselves by monitoring what we're doing and giving us advice on how to make better use of our time or to adjust our approach to account for the way the human mind works and learns.

References

Baron-Cohen, S.; Wheelwright, S.; Hill, J.; Raste, Y.; and Plumb, I. 2001. The reading the mind in the eyes test revised version: A study with normal adults, and adults with asperger syndrome or high-functioning autism. Journal of Child Psychology and Psychiatry 42(2):241–251.

Eicher, B.; Cunningham, K.; Marissa Gonzales, S. P.; and Goel, A. 2017. Toward mutual theory of mind as a foundation for co-creation. Presented to the International Conference on Computational Creativity, Co-Creation Workshop, June 2017.

Engel, D.; Woolley, A. W.; Jing, L. X.; Chabris, C. F.; and Malone, T. W. 2014. Reading the mind in the eyes or reading between the lines? theory of mind predicts collective intelligence equally well online and face-to-face. PLoS ONE 9(12):1 – 16.

Goel, A., and Polepeddi, L. 2017. Jill Watson: A virtual teaching assisant for online education. Presented to the Learning Engineering for Online Learning Workshop, Harvard University, June 2017. To appear as a chapter in Dede, C.,Richards,J.,&Saxberg,B.,Editors(inpreparation),Education at scale: Engineering online teaching and learning. NY: Routledge.

Goel, A., and Joyner, D. 2016. An experiment in teaching artificial intelligence online. Journal for Scholarship of Technology-Enhanced Learning 1(1).

Ma, L. 2007. Investigating and improving novice programmers' mental models of programming concepts. Ph.D. Dissertation, University of Strathclyde.

Premack, D., and Woodruff, G. 1978. Does the chimpanzee have a theory of mind? Behavioral and Brain Sciences 4(4):515–629.

Sorva, J. 2013. Notional machines and introductory programmingeducation. Trans.Comput.Educ.13(2):8:1–8:31.

Robinson, A. L. 1980a. New Ways to Make Microcircuits Smaller. *Science* 208:1019-1026.

Woolley, A. W.; Chabris, C. F.; Pentland, A.; Hashmi, N.; and Malone, T. W. 2010. Evidence for a collective intelligence factor in the performance of human groups. Science 330(6004):686–688.