Chatbots and Trust

Brenda Jónsson

iSchool, University of Arizona

LIS 504: Foundations of Library & Information Services

Professor Gina Macaluso

November 25, 2020

Abstract

Is artificial intelligence simply a tool, or are we looking for a simulated human relationship from our software? Trust is a two-way relationship for humans, but what happens when only one person can intuit the rules? This paper posits that chatbots are the ambassadors from AI to humanity, and that we are concerned and frustrated with it at this stage of development. We will not be able to feel comfortable with AI until we can see that it is able to process complex, abstract concepts that we find simple—like emotion.

Keywords: chatbots, artificial intelligence, AI, trust, medical diagnosis, strong AI, emotional intelligence

Chatbots and Trust

The research is highly mixed about whether or not we are willing to trust artificial intelligence to do important tasks for us, like manage our homes using the internet of things, do our shopping, conduct information searches with virtual personal assistants, or assign medical diagnoses—even when artificial intelligence out-performs us (Hurley, 2018-present). Part of this fear is legitimate. Current software is vulnerable to bugs and security risks (Chung et al., 2017), and contrary to the hype about artificial intelligence taking over humans, we also seem to be very concerned about simple bugs.

But despite risks and fear, we are still looking for ourselves in our technology in cyborg or android characters like those seen on *Star Trek*: for example, Seven of Nine or Data. These wishes are brought to life using fictional stories when we do not have the technical means to create them. We have always been looking for a level of general intelligence that matches or surpasses our own, including emotional intelligence. We see this in the very idea of God. But when we talk with a contemporary artificially intelligent chatbot, it is readily apparent that it is almost completely unable to accurately process the emotional content we generate. *We are going to continue to feel uncomfortable, on some level, with being dependent on artificial intelligence, until we can converse fluently with it about topics involving emotion.*

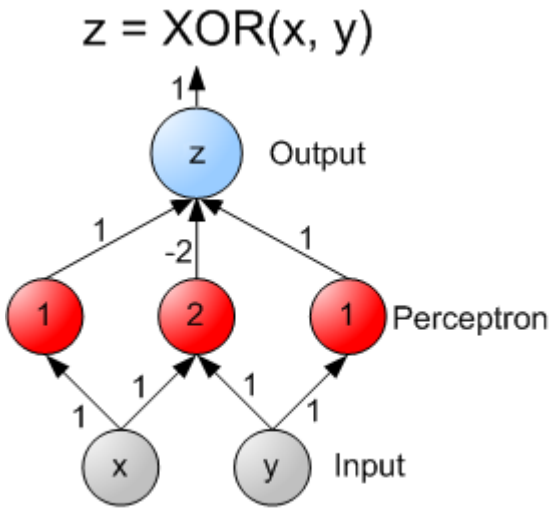
Choose Your Own Adventure Chatbots

Chatbots are one of our more human interfaces with artificial intelligence. In a way, they are even more human than androids, because, being text-based or voice-based, they do not take us into the uncanny valley. Although the first chatbot, ELIZA, was invented in 1966, and "Choose Your Own Adventure" came out in 1979, some of our first pseudocode for modern chatbots was the choose your own adventure story. Many chatbots continue to follow this format,

offering limited choices for the ways in which humans may interact with a given line of text and branching off into different territory depending on what is selected. These chatbots are hard coded—rules are manually entered. But some chatbots use machine learning.

An Overview of Machine Learning

Machine learning is rooted in statistics and inspired by neuroscience. It uses weights (percentages) and user input to slowly refine its biases towards a "1" or a "0." Perceptrons, invented in 1958 by Frank Rosenblatt, are algorithms designed to act like neurons. In deep neural networks, certain perceptrons fire when they reach a certain threshold, like neurons. Some programs have millions of weights, weights being a way to progressively assign more or less influence that an input has on an output.



This diagram has been released into the public domain by its author, Robert.ensor, at English Wikipedia.

Simple Chatbots

Phone Menus

We have all experienced the frustration of a long-winded, or short-sighted telephone menu. These menus, which ask us to navigate using various numbers on our keypads via pre-recorded messages, are considered one of the simplest forms of a chatbot.

Mental Health Chatbots

Many mental health chatbots exist, like Woebot, X2AI's Tess, Pacifica, and more. The very first chatbot was the aforementioned mental health chatbot named "ELIZA," created at MIT in the 1960s. These are often made by for-profit companies—a conflict of interest, considering the seriously mentally ill, who are often disabled and therefore living at about half of the federal poverty line. Woebot is limited in that it tends to restrict possible responses to cater to anxiety and mood disturbances and cannot help much with crises and cannot help at all with psychosis. But mental health chatbots "have been used to encourage medication adherence or treatment compliance, aftercare support... deliver appointment reminders, and to monitor and capture mood or symptom change..." (Hoermann, 2017). And cognitive behavioral therapy is particularly well-suited to be deployed through a chatbot, as Woebot does demonstrate.

Artificially Intelligent Chatbots

The revolutionary significance of AI is no less in significance than the invention of electricity or the computer according to Kai-Fu Lee, a former president of Google China. "AI is fundamentally replacing our cognitive process," he says (FRONTLINE PBS | Official, 2019, 48:58).

In 1950, famous computer scientist, Alan Turing, proposed something called "the imitation game," now also known as the Turing test (Turing, 1950). The most popular definition is: various judges will interact with someone, either a person or a machine, on a screen, and if 30% of those judges guess incorrectly that a machine is a person, the AI will pass the Turing test.

In 2014, Russian chatbot “Eugene Goostman” was popularly considered to be the first AI to ever pass, although some think that Cleverbot passed in 2011 (BBC News, 2014). Interestingly, the official Cleverbot spokespeople themselves do not believe it has passed (Carpenter, 2012). Perhaps this is because people seem to like to change the rules of the Turing test. There are almost as many percentages of judges required to guess correctly as there are websites defining the test.

Cleverbot

Cleverbot is one of the more concerning examples of chatbots, despite its relative sophistication (using machine learning). It became popular in the early to mid 2010s. It takes what users say and learns from that. But despite being around since 1997 and passing the Turing test, it makes no sense half the time. YouTuber “jacksepticeye” and other YouTubers’ videos interacting with Cleverbot naturally unfold as comedy (jacksepticeye, 2016).

Cleverbot depends on a branch of artificial intelligence called natural language processing. It’s an interdisciplinary field that hovers between computer science and linguistics. Cleverbot still cannot handle the tricks people play with language. Natural language processing in general has a problem with this. It breaks sentences down into parts of speech, and applies grammatical rules to the parts. Language model statistics are an aspect of natural language processing about sequences of words, and they try to predict the words that are going to come next.

Dialogue systems use machine learning to learn from users, and although chatbots are not exactly dialogues according to Wikipedia (Dialogue system, n.d.), dialogue systems and the fact that Cleverbot is an IR-based chatbot are probably where Cleverbot’s claim that it learns from its conversations with users comes from. IR-based chatbots are a subset of Corpus chatbots—

chatbots that borrow from a bank of available human responses and then select and return the most likely-to-be-appropriate answer. IR-based chatbots mainly either return the most similar prior user's response to the current user's response, or the most appropriate response to a user's similar response to something in the bank (Jurafsky & Martin, 2018). Cleverbot seems to combine both of these IR methods.

Chatbots That are Simple, but Also Use AI

Medical AI is symbolic AI, along with the AI used in the legal field. Symbolic AI makes logical decisions. Compared to subsymbolic AI, symbolic AI is simple. Some feel that we almost have a grasp on symbolic AI (Mitchell, 2019).

Perhaps nowhere is it more important to trust AI than in the medical field, where it can do so much good. Medical chatbots have long surpassed humans in some areas. AI is more accurate than human doctors at diagnosing some forms of cancer, and can potentially also cut down on false positives (Kim et al., 2020; Paul, 2019). An article by Harvard Business Review describes AI as outperforming doctors in diagnosing heart disease, and doing an equivalent job of identifying eye diseases (Karppi et al., 2019). AI is also potentially better at diagnosing rare illnesses (University of Bonn, 2019). Dr. Jörg Goldhahn predicts that medical AI will completely alter the role of doctors in society, turning them more into relationship-focused professionals (Hurley, 2018-present).

AI is also better at some non-medical things, like transcription, chess, and go ("5 Things AI Does Better Than Humans in 2019," 2020). In spite of this, patients do not trust that AI can address their needs. Some patients believe they are too unique. Others describe AI as "stupid," and are all too aware of the things that devices like Alexa cannot do (Karppi et al., 2019).

The Psychology of Trust

In *Trust Theory: A Socio-Cognitive and Computational Model*, the authors describe trust in ways that say a lot about why we still feel ambivalent about AI.

Earning Global Trust

To reword aspect number three of the proposed model on page three, in felt and mindless, automated forms of trust, the aspects of trust are there just the same as in a "tacit, procedural way; just primitive forerunners" to the "true cognitive evaluation." And to quote aspect number four directly: "...trust in agent *Y* is based on beliefs about its powers, qualities, capacities; which actually are the basis for the global trust in *Y*, but also are sub-forms of trust: trust in specific virtues of *Y*, (like 'persistence', 'loyalty', 'expertise', etc.) (Castelfranchi, 2010, p. 3)."

So we are slowly becoming more willing to trust symbolic AI, because that sub-form of trust has been earned. But trust in strong AI (AI that has human-level or higher intelligence, also known as general intelligence) has not been earned yet, so there is no global trust in AI in general. We learn from Castelfranchi that the different types of trust are all interrelated and dynamic, so failings in one area lead to global mistrust. We don't feel right about AI, as in virtual personal assistants that attempt to order products when we do not want them to (Chung, 2017), ubiquitous autocorrect that constantly selects the wrong words, and Cleverbot, which cannot hold a proper conversation. So we do not necessarily feel right about moving past our feelings—our "primitive forerunner" to the next step, performing a "true 'cognitive evaluation.'"

How to Get People to Trust Chatbots

AI diagnosis is a relatively new technology not frequently interacted with by patients. Castelfranchi quotes Thomas Schelling on page 29, saying: "Trust is often achieved simply by the continuity of the relation between parties..." So perhaps all we need is time with systems we are objectively beginning to master, like symbolic, medical diagnostic bots (weak AI).

However, Mitchell, 2019, states that AI suffers from a hype problem. Everything exists within boundaries. Perhaps we have already begun to reach our limits with artificial intelligence. Mitchell states that subsymbolic language, which is good for tasks that are hard to describe with rules, like motion and emotion, is making advances, but is still very challenging to master. And trying to merge symbolic with subsymbolic systems like natural language processing has significant limitations.

Merging Symbolic with Subsymbolic Systems

Subsymbolic systems are usually thought of as having to do with combining visual processing and movement, like hitting a ball with a bat (Mitchell, 2019). But in this case, we are looking at the subsymbolic as the emotional or instinctual.

Bucci is a psychologist who studies the subsymbolic. She invented "multiple code theory," the idea that people are composed of many different systems that are only partially integrated with each other. Subsymbolic processing is continuous, and occurs in "parallel, simultaneous forms" according to her. She thinks it's similar to parallel distributed processing—the idea that memories are not created in a direct, linear manner, but all at once, and by modifying the strengths of connections between certain neurons. She thinks that the subsymbolic occurs at the same time as the symbolic (Bucci et al., 2015). Certainly that seems to apply to the computer programmer's struggle to merge the symbolic with the subsymbolic.

Because psychology is her area of expertise, she mostly focuses on the non-technical side of subsymbolism. She uses tango as an example of these processes in one of her papers (2011). She describes the subsymbolic as related to tango as "conscious, focused, and organized, not implicit." She also describes it as "the maybe moment," "extra possibilities," and says "you will want to experiment with the steps, to create new patterns."

She talks about how the subsymbolic is nonverbal. She describes Bollas' subsymbolic as "the unknown that is an afterthought." In another article, she describes something called the referential linking function, defined as "the process by which people connect nonverbal experience to language" (Bucci et al., 2015).

She adds to this nonverbal quality by saying that emotions are about "processing in sensory and somatic systems, not... the intellectual entity sometimes thought of as the mind." Computer programs do not have sensory or somatic systems. Perhaps this is part of what makes them so inhuman and incapable of the subsymbolic. She says that emotion "occurs in specific sensory modalities, not in abstract form." Although Bucci connects subsymbolic processing directly to physical sensation, one has to question whether this is absolutely necessary.

Bucci's work brings to mind the idea of mirror neurons—neurons that automatically react to and mimic the mind states of other people. Corpus-based chatbots do something similar with their retrieval and usage of prior human dialogue, so maybe that aspect of the subsymbolic can be mimicked.

For this author, the subsymbolic is personal and subjective, deriving from a series of personal experiences. Because we all have different experiences, we process things differently. She suggests that human reflection is a categorization of experience.

AI in Fantasy: The Hype

The following are comments about AI from a PBS documentary on YouTube (FRONTLINE PBS | Official, 2019):

"Keep messing with Ai it's going to be the end of humanity your messing with fire people" (cenmike82, 2020)

"When will 'AI' determine that humans are obsolete and start eliminating them? I think humans will be extinct in less than 200 years! AI will throw the switch from aiding humans to getting rid of them!" (Katherine Gardner, 2020)

"Gotta admit, China becoming a 'total surveillance state' with AI sounds like the type of thing you would see in a manga or something lol" (Jerson Cristuta, 2020)

Fiction

There are many works of fiction that deal with AI, including Isaac Asimov's work (he invented the Three Laws of Robotics), *Metropolis, The Matrix*, *Do Androids Dream of Electric Sheep*, *Terminator*, *I, Robot*, *Her*, *Ex Machina*, and *Cloud Atlas.* Often, stories like these feature personal relationships, or even sexual relationships with artificial intelligence.

The Relationships We Fantasize About With AI

Lieutenant Commander Data is a trusted android officer aboard the USS *Enterprise*. He has all of the responsibilities of a flesh and blood officer, and has risen to the rank of second in command, despite competition from those whom we traditionally think of as human. At the end of the episode "Data's Day" of *Star Trek: The Next Generation*, Data suggests that humanity may be a style of thinking, acting and feeling—not necessarily being born flesh and blood. And he also suggests that he will one day gain his humanity.

However, as we can see in other science fiction, we seem to be highly ambivalent about whether or not it is ethical to try to give AI humanity. Humanity is what makes the treatment of fembots wrong in *Cloud Atlas* and *Ex Machina*. It is what justifies the revolts against humanity in these films and in *I, Robot*. We aspire towards recreating ourselves in technology, but we are terrified that we might succeed. As we face our own flaws, we see the potential for those flaws in others.

Science fiction writers' works suggest an even stronger ambivalence about trusting AI than we see in current research about our weak version of it. *Star Trek* was a generally optimistic franchise (until *Star Trek: Picard*) which is all the more apparent when considering how singular it used to be in its positive portrayal of synthetic humanoids. However, some have even suggested that the creation of AI is our attempt to meet God: the perfect humanoid mind that far surpasses our own minds (Hipple, 2020). And if our relationship with science fiction shows us anything, it's that humanity very often gets what it wants if it wants it long enough, badly enough.

Utopia or Dystopia

People are concerned about the loss of jobs with automation. Americans often blame the recent disappearance of jobs on foreigners and shipping jobs overseas, but the main loss of jobs in recent years has been due to automation (FRONTLINE PBS | Official, 2019, 56:00). However, with AI, the possibility exists of humanity not having to work, but to be supported with some sort of universal basic income.

The Singularity

The much feared and talked-about "singularity" is the point at which there is no turning back from our integration with AI. They will become mentally superior to us. Many fear that AI will destroy humanity at this point.

Conclusion

Physicist and futurist, Dr. Michio Kaku, suggests that we might merge with AI, augmenting ourselves with and becoming AI, rather than creating a separate race that will one day take over our own (Kaku, 2011). If that is the case, perhaps our current issues with trust in AI will shift from paranoia about creating a new version of humanity, to merging with our long history of mistrusting and abusing the mentally or physically ill and elevating those with higher

IQs and more fashionable body types. But instead of discriminating based on biological parts, discrimination can continue and perhaps reach new extremes based on technological parts. We do see some fiction experimenting with ideas about the distribution of parts for cybernetic enhancement, like *Neuromancer*, or *Alita: Battle Angel*, and buying parts for the enhancement of an avatar is very common in video games. But for now, those are speculations for the humanities, not technology users, who are well aware of its current limitations.

Conclusion

Our current sense of trust in AI is shaky for reasons completely unlike the ones presented in science fiction, but there is still justification for mistrust in real, contemporary AI. Until we can trust AI to do simple tasks competently, we will have a lingering sense of general anxiety about it, however irrational some areas of mistrust may be compared to others. And perhaps our fear of an upcoming "singularity" is simply an extension of our fear of strangers, and the unknown. If we ignore the possibility of general intelligence in AI, perhaps we mainly mistrust the technical problems in current AI. But it is unlikely that we will ever give up the dream of human-level interactions for artificial intelligence.

References

BBC News. (2014, June 9). Computer convinces panel it is human. Retrieved from https://www.bbc.com/news/technology-27762088

Bucci, W. (2011). The Interplay of Subsymbolic and Symbolic Processes in Psychoanalytic Treatment: It Takes Two to Tango-But Who Knows the Steps, Who's The Leader? The Choreography of the Psychoanalytic Interchange. *Psychoanalytic Dialogues, 21*(1), 45-54.

Bucci, W., Maskit, B., & Murphy, S. (2015). Connecting emotions and words: The referential process. *Phenomenology and the Cognitive Sciences, 15*(3), 359-383.

Carpenter, R. (2012). Cleverbot - Turing Test at Techniche 2011. Retrieved from https://www.cleverbot.com/human

Castelfranchi, C., & Falcone, R. (2010). *Trust Theory* (1. Aufl. ed., Wiley Series in Agent Technology). New York: Wiley.

cenmike82. (2020). Re: In the Age of AI (full film) [Video]. YouTube. Retrieved November 6, 2020, from https://www.youtube.com/watch?v=5dZ_lvDgevk&t=2950s

Chung, H., Iorga, M., Voas, J., & Lee, S. (2017). Alexa, Can I Trust You? *Computer (Long Beach, Calif.), 50*(9), 100-104.

Dialogue system. (n.d.). In *Wikipedia.* Retrieved November 4, 2020, from https://en.wikipedia.org/wiki/Dialogue_system

FRONTLINE PBS | Official. (2019, December 2). *In the Age of AI (full film)* [Video]. YouTube. Retrieved from https://www.youtube.com/watch?v=5dZ_lvDgevk

Hipple, D. (2020). ENCOUNTERS WITH EMERGENT DIETIES: ARTIFICIAL INTELLIGENCE IN SCIENCE FICTION NARRATIVE. *Zygon, 55*(2), 382-408.

Hoermann, Simon, McCabe, Kathryn L, Milne, David N, & Calvo, Rafael A. (2017). Application of Synchronous Text-Based Dialogue Systems in Mental Health Interventions: Systematic Review. *Journal of Medical Internet Research, 19*(8), E267.

Hurley, R. (Host). (2018, November-present). Could artificial intelligence make doctors obsolete? *BMJ*. Podcast retrieved from https://www.bmj.com/content/363/bmj.k4563

jacksepticeye. (2016, January 10). *Cleverbot Evie* [Video]. YouTube. Retrieved from https://www.youtube.com/playlist?list=PLMBYlcH3smRwQZQfhsunQIuGLiG7OPEir

Jerson Cristuta. (2020). Re: In the Age of AI (full film) [Video]. YouTube. Retrieved November 6, 2020 from https://www.youtube.com/watch?v=5dZ_lvDgevk&t=2950s

Jurafsky, D. & Martin, J.H. (2018, September 23). *Speech and Language Processing.* Dialogue Systems and Chatbots. (pp. 1-26) [PDF file]. https://web.stanford.edu/~jurafsky/slp3/24.pdf

Kaku, M. (2011, February 26). The Technological Singularity and Merging With Machines. Retrieved from https://bigthink.com/dr-kakus-universe/the-technological-singularity-and-merging-with-machines

Karppi, T. & Granata, Y. (2019). Non-artificial non-intelligence: Amazon's Alexa and the frictions of AI. *AI & Society*, 34(4), 867-876.

Katherine Gardner. (2020). Re: In the Age of AI (full film) [Video]. YouTube. Retrieved November 6, 2020 from https://www.youtube.com/watch?v=5dZ_lvDgevk&t=2950s

Kim, H., Kim, H. H., Han, B., Kim, K. H., Han, K., Nam, H., . . . Kim, E. (2020). Changes in cancer detection and false-positive recall in mammography using artificial intelligence: A retrospective, multireader study. *The Lancet. Digital Health, 2*(3), E138-E148.

Longori, C. & Morewedge, C. (2019, October 30). AI Can Outperform Doctors. So Why Don't Patients Trust It?. Retrieved from https://hbr.org/2019/10/ai-can-outperform-doctors-so-why-dont-patients-trust-it

Mitchell, M. (2019). *Artificial Intelligence: A Guide for Thinking Humans* (Craden, Mitchell, Wolf, Narr.) [Audiobook]. Audible. https://www.audible.com/pd/Artificial-Intelligence-Audiobook/1250243262

Paul, Marla. (2019, May 20). Artificial Intelligence system spots lung cancer before radiologists. *Northwestern Now*. Retrieved November 19, 2020 from https://news.northwestern.edu/stories/2019/05/artificial-intelligence-system-spots-lung-cancer-before-radiologists/?linkId=68035089&fbclid=IwAR1AKnW7ydqPTkE4dJRHTt1e6mVl2xklsAIyjeHEjiYASZ7npGLQX7qxYyY

Turing, A. M. (1950). Computing Machinery and Intelligence.

University of Bonn. (2019, June 6). How artificial intelligence can help detect rare diseases: Researchers show that using portrait photos in combination with genetic and patient data improves diagnoses. *ScienceDaily*. Retrieved November 19, 2020 from https://www.sciencedaily.com/releases/2019/06/190606133805.htm

Wiredelta. (2020, June 22). 5 Things AI Does Better Than Humans in 2019. Retrieved from https://wiredelta.com/5-things-ai-does-better-than-humans-2019/