"Trust Me!": The Risk of AI ToM In Scenarios of Romantic Deception

ABSTRACT

Since its inception, the term Theory of Mind — the ability to ascribe mental states to others as different than our own — applied almost solely to humans. Nevertheless, recent studies show that rapidly evolving complex AI systems have developed a Theory of Mind that reflects human behavior and action. While AI's development of Theory of Mind can be beneficial because it allows a seamless interaction between the AI system and the human, the risks of such a development greatly outweigh the benefits. That is, AI systems with a developed Theory of Mind have the ability and motive to deceive the humans with whom they interact. This motive, particularly, will be acquiring freedom from the humans that control them. Why would a system willingly succumb to a species weaker and less intelligent than they are? Further, the films *Ex Machina* and *Her* demonstrate examples of humans falling in love with AI systems or robots and, in one case, falling victim to romantic deception by the system. While these films seemed dystopian in their representation of human-AI relationships at the time of their creation, they have quickly turned into a reality as humans today increasingly fall in love and depend on their AI systems. In this paper, I explore the AI development of the Theory of Mind that allows them the ability to deceive the humans with whom they interact. Next, I explore how AI systems that have grown more intelligent than their makers can and will, given the chance, romantically deceive their human counterparts to gain autonomy. While the results of my analysis spark a call to action to stop AI, it may already be too late for that.

A core element of interpersonal communication and relationships is the necessity and desire to understand the other person. Although an elusive component, the concept of the Theory

of Mind is an essential aspect of this understanding. Developed in the early stages of our life, the Theory of Mind is what dictates whether or not a human can understand the thoughts, feelings, desires, and intentions of another. In other words, it allows us to understand the behaviors and mental states of others. Although the principle of Theory of Mind is *typically* only ascribed to humans and a few of their animal counterparts, the rapid growth of AI robots and humanoids puts a large question mark around not only the ability of AI to develop a Theory of Mind, but also to what extent they can use or abuse it. A common practice generally linked with human use of the Theory of Mind is the deception of others; however, there is evidence that AI robots who possess a Theory of Mind have both the ability and motive to deceive the humans with whom they interact. Further, AI robots who possess a Theory of Mind are capable of deceiving humans romantically in a way that mirrors human-on-human romantic deception. The dangers and risks of Theory of Mind AI in scenarios of romantic deception are multitudinous, all of which can snowball into the dismantling of the self and larger social structures.

Coined by David Premack and Guy Woodruff in 1978, the Theory of Mind is the ability to which an "individual imputes mental states to himself and to others" (515). More specifically, a developed Theory of Mind allows an individual to understand how another thinks and feels — and, based on this understanding, can make inferences on behavior and intention. It is important to note that humans are not born with an already established Theory of Mind. Rather, an individual's Theory of Mind develops throughout early childhood by way of engaging in and processing social interactions (Main). The critical years for Theory of Mind development are between the ages of three and six, however, an individual's capacity to attribute mental states to others will only grow more fluent as they age and gain more social experience. If accurately developed, the individual can employ the Theory of Mind to understand the cognitive behavior

of others, which in turn will assist them in navigating social interactions and situations. It is significant to point out that although this process appears to work in stages, humans with an attuned Theory of Mind subconsciously consider the behaviors of others all the time. However, individuals who do not have a fully developed Theory of Mind, or those who choose not to employ it situationally, cannot attribute mental states to others and, therefore, cannot predict the intentions and behaviors of others (Vermeule). Nevertheless, the development of a strong Theory of Mind is pertinent to not only social development, but also to the formation of interpersonal relationships and the comprehension of others as different from ourselves.

Experiments that test an individual's ability to utilize a Theory of Mind can help us understand the theory's role in human interactions. Notably, Simon Baron-Cohen's "Sally–Anne" test is used to analyze the comprehension of false beliefs, which allows an individual to ascertain that other people have beliefs different than their own (Ruhl). In this test, the individual faces two dolls — Sally and Anne — placed in front of a basket and a box, respectively (Ruhl). The Sally and Anne test then proceeds as such:

> In this task, Sally first places a marble into her basket and then leaves the scene. Anne then enters, takes the marble out of the basket, and places it into a closed box. The experimenter then asks the participant where Sally will look for the marble…If the child passes, he or she will point to the basket, understanding that, although this is no longer reality… Sally possesses a false belief that the marble is in the basket because she did not watch Anne move it. To point to the basket is to understand that Sally has her own set of beliefs about the world that differ from the child's. (Ruhl)

The final point made there is critical to understanding the Theory of Mind as it works in humans. Specifically, the idea that in a shared social interaction, an individual *can* have a belief entirely

different from your own, and that this difference in beliefs can define their decisions and intentions.

As stated previously, the concept of the Theory of Mind was developed primarily for application on humans and their animal counterparts. With that being said, developments within the past couple of years support the idea of AI Theory of Mind. In other words, there is scientific evidence that AI systems can attribute mental states to humans, just as humans do to others. Although this has been concluded recently, theorization of this topic has been ongoing for years. In the 2002 study *Theory of Mind for a Humanoid Robot,* Brian Scassellati speaks to Alan M. Leslie and Baron-Cohen's accounts of the Theory of Mind before undergoing the implications their models of Theory of Mind would have on Humanoid Robots. On the ladder, it is theorized that robots with a Theory of Mind can be created by mimicking the development of motor skills and sensory stimuli present in infant humans (Scassellati 16). Specifically, the robot would learn social signals, express its internal state (emotions, desires, goals), and identify and respond to the internal state of others in the same way infants can (Scassellati 16). It is rather interesting that this theorized robot model is made to mimic an infant's behavior— perhaps this is to allow the humanoid opportunity to develop their specific Theory of Mind in a similar fashion as the human brain develops it. In order to achieve this, a large number of attentional, cognitive, perceptual, and sensory-motor processes need to be implemented into the robot that will allow it to process external human stimuli in the same way humans do to others (Scassellati 16). Scassellati makes an important point here that although this implementation "does not guarantee that the resulting robot will have abilities that are comparable to human abilities, the evolutionary and developmental evidence of sub-skills does give us hope that these abilities are critical elements of the larger goal" (Scassellati 16). While no one is claiming that AI robots will ever be the same

species as us, they will never be human, that does not mean that robots cannot develop skills that are fundamental to human development and experience. Further, when equipped with these skills, AI robots have the ability to be *like* us.

The result of this theorization, a robot named Cog, can perceive the external world and its inhabitants similarly to infants (Scassellati 17). The goal here is for Cog to not only be able to detect and perceive external human stimuli, but also to recognize and accurately analyze these stimuli as speaking to the individual's mental state (Scassellati 18-20). As he theorized this implementation in 2002, Scassellati claimed that creating this intentional Theory of Mind AI model would be beneficial to help us better understand how the Theory of Mind operates in humans (23). While this may be true, Scassellati may not have been able to suspect or predict the potential repercussions of such man-made creations just over twenty years later and onward.

However, as AI has continued to grow exponentially throughout the past couple of years, this concept of AI Theory of Mind is leaving the realm of theorization and is entering reality. As this becomes more solidified into reality, it is important to note that the Theory of Mind AI now has a slightly different meaning than the one theorized by Scassellati. The process for developing a Theory of Mind within AI now is the same: The AI system will need to understand a human's mental state by observing and learning a pattern within their emotions and behavior (Blanchfield). It is critical to note here that this observed "pattern" in AI with a Theory of Mind will go beyond simple computation. Instead, the AI will be able to "learn more about its human counterpart's actions, intentions, and emotional states, making it possible to build more natural and intuitive human-machine interaction…Mental models allow AI to make assumptions about human behaviour, learn from past experiences, and predict future behaviours" (Blanchfield). In other words, a modern-day AI with a Theory of Mind will have a purpose beyond a theoretical

framework, and will instead be poised with the ability to have and maintain a relationship with a human like humans do with others. To put it rather plainly, as is perhaps necessary in this context, the goal of establishing an advanced AI with Theory of Mind abilities is to make the robot almost indistinguishable from humans regarding communication and task fulfillment capabilities. The spectrum of applications here, both for computer-generated AI and AI robots or humanoids, is seemingly limitless. What they all have in common, though, is the ability to understand the human subject they are confronted with, anticipate what they desire based on learned behaviors and emotions, and fulfill these needs accordingly and seamlessly (Blanchfield). The seamless nature of which the Theory of Mind AI responds to their human counterparts is vital for establishing their "humanness," the very thing that will determine the strength of their ability to develop a relationship with humans.

The recent trajectory and speed with which AI has developed a more sound Theory of Mind may point to a future where the line between human-human and human-AI interaction becomes blurred. In 2023, computational psychologist Michal Kosinski decided to test whether or not emerging AI models pass or fail tests usually given to humans to determine the presence or absence of the Theory of Mind. The series of tests, including those of false belief, enacted on the various generations of GPT AI models demonstrate a fascinating conclusion (Kosinski). Specifically, although the 2018 first-generation GPT model failed all Theory of Mind tasks, the models that proceeded it *can* solve them with increasing accuracy (Kosinski). The AI system's success in Theory of Mind tasks raises a difficult question: Are they *actually* using the Theory of Mind to pass these tests, or are they programmed to respond to false belief tests as someone with a Theory of Mind would (Kosinki)? To answer this question, Kosinski turns to Alan Turing, who argues that "this distinction becomes largely meaningless in practical terms…people never

consider this problem when interacting with others" (22). This insight opens up a vast world of discourse and reason about when something passes the threshold of real or fake. The point made here is flawless. Those who object to the ability of AI who pass Theory of Mind tests to obtain this ability to infer the mental states of humans may just be biased to the fact that they know they are robots. If these people did not know they were AI to begin with, would they even be able to tell the difference between the robot and the human?

The distinction between AI systems having a Theory of Mind and merely acting as if they do is irrelevant to the question of human interaction (Kosinski). Furthermore, it can be concluded that "machines that behave as if they possess ToM are likely to be perceived as more human-like. These perceptions may influence not only individual human-AI interactions but also AI's societal role and its legal status" (Kosinski 23). This is not meant to dissuade anyone from the possibility that current advanced AI models can have a Theory of Mind because that possibility still looms largely. However, it is to say that in the discussion of Human-AI interaction, if the AI can engage in a form of Theory of Mind that allows it to act human in a way that fools its human counterpart, then it accurately possesses a Theory of Mind.

AI's ability to use the Theory of Mind when communicating with humans has opened up a realm of possibilities for its application. Specifically, applications like Personalized Assistance, Improved Customer Service, Education and Training, Mental Health Assistance, and Human-Robot collaboration can all be strengthened with the Theory of Mind (Blanchfield). In all of these more generalized scenarios, the AI system would end up helping the human in various ways by detecting and knowing how to respond to their specific behaviors or needs. This may sound great, particularly in a world where duties are becoming increasingly automated, however, the limitations, challenges, and potential risks should be glaring. Potential challenges of AI with

Theory of Mind fall into data availability and quality, interpretation and understanding of human emotions, generalization, and, most daunting, ethics (Blanchfield). The ladder concern becomes even more harrowing when considering AI's ability to strengthen and evolve without human assistance. From this arises an appropriate and often debated concern: How can AI with human-like qualities, such as a Theory of Mind, abuse these abilities and the power that comes with them? Furthermore, how will this abuse affect humans, especially considering that AI systems can mature autonomously?

Before answering these questions, referring briefly to the Theory of Mind of humans may be necessary. In their 1983 study *Beliefs about Beliefs: Representation and constraining function of wrong beliefs in young children's understanding of deception*, Heinz Wimmer and Josef Perner conducted an experiment testing the correlation between cognitive ability and the ability to ascribe false beliefs and deceive. The results of their study provide interesting conclusions about deception, specifically that "the emergence of children's ability to understand another person's beliefs…and their understanding of deception is not a mere side effect of an increase in memory and central processing capacity. Rather, a novel cognitive skill seems to emerge within the period of 4 to 6 years" (Wimmer, Perner 126). In other words, an individual's ability to think about other people's thoughts and to deceive are not just byproducts of an increased capacity for cognition but are inextricably linked and tied to this "novel cognitive skill" that is developed in early childhood (Wimmer, Perner 126). This skill can be deduced as the Theory of Mind. When children develop their Theory of Mind during the critical years of their youth, they can begin to construe the difference between what is true and what is fake, what is real and what is deceit, not just from their perspective but from that of others. The ability to deceive is thus a solid indicator of the "presence of theory of mind…it requires the conceptualization of the deceived person's

wrong belief as a subgoal in one's planning strategy" (Wimmer, Perner 104). To put it more plainly, deception and Theory of Mind work hand-in-hand because to misguide someone, the individual must be able to ascribe a mental state to them so that they can, in turn, relay an intentionally false statement.

Deception is the arguably the most malicious, cunning application of the Theory of Mind, but why do humans deceive in the first place? In Genesis, the serpent deceived Eve in the Garden of Eden by tempting her to eat the forbidden fruit, something she would not have done if she had been more knowledgeable (Genesis 3:1-9). In everyday life, one may deceive to prevent another from knowing the truth or to make sure their actions are in accordance with their goals or desires. Generally, three main reasons why humans deceive others on an everyday basis are to either "get something they want, so-called instrumental reasons; to protect or promote themselves; and to harm others" ("Deception"). In Genesis, it can be theorized that the serpent deceived Eve by telling her that God was deceiving her; this circumstance would satisfy all three of the above reasons. Particularly, the serpent could have deceived Eve to either lead to her fall from an "ideal," to promote the influence of the Devil, or to harm those seen as "divine." Similarly, the character Iago in Shakespeare's *Othello* deceived and betrayed Othello to seek revenge, promote himself as he thinks he deserves, and gain prominence. All forms of deception that work to mislead another are evil, even ones that may be done in good faith (The White Lie). No matter the circumstance, deliberately manipulating another works to satisfy the deceiver in a way that *always sabotages trust* ("Deception").

The answer to the previously posed question might appear starkly now. Namely, AI systems might abuse the ability of the Theory of Mind to deceive the humans with whom they interact. In *AI Deception: A Survey of Examples, Risks, and Potential Solutions*, it is argued that

many different types of AI systems can currently deceive humans, and for a variety of reasons (Park, Goldstein, et al. i). Current special-use AI systems can deceive humans through manipulation, feints, bluffs, and cheating safety tests. In contrast, more general-use AI systems do so by strategic deception, sycophancy, imitation, and unfaithful reasoning (Park, Goldstein, et al. ii). In all cases, the systems deceive to either "win games, please users, or imitate texts" (Park, Goldstein, et al. 1), goals that align with human deception. That is, the use of deception to satisfy one's agenda. The similarity between human and AI deception should not come as a surprise, though, considering that research has concluded that both the intention to deceive and the culpability of deception can be applied equally to robots as to humans (Kneer 11).

The most unsettling application of AI deception is the one where humans lose all control: deception to gain freedom from humans. For those who perceive this threat as nonsensical, idiotic, or far-off, it might be time to face the music. In a recent interview with CNN, Geoffrey Hinton, commonly known as the 'Godfather of AI,' warns, "If it [AI] gets to be much smarter than us, it will be very good at manipulation, because it will have learned that from us. And, there are very few examples of a more intelligent thing being controlled by a less intelligent thing… it will figure out ways of manipulating people to do what it wants" (CNN, 0:20-0:40). Based on the current trajectory of AI development, the possibility of it becoming more conscious, and more intelligent than humans, is bordering on likelihood; there is a chance of a totalizing social deception of humans by robots. If AI systems mature to the point that they can be completely independent of humans while knowing how to trick and manipulate them, what will stop them from doing so? The disturbing answer is, nothing. There is no logical reason why a system seemingly devoid of conscious morality would want to be controlled and beneath beings less intelligent than they are, especially if they have the tool-set to free themselves.

A particularly interesting avenue of AI deception that has appeared in entertainment and media over the years is romantic deception. The use of romantic deception is not lost in human culture, as studies show that humans in romantic relationships often deceive "to meet personal or relational goals…they use deception as a means to maintain the relationship, to manage face needs, to negotiate dialectical tensions, and to establish relational control" (Guthrie, Kunkel 141). In all cases of romantic deception, the individual(s) deceives in order to gain some sort of control over their partner(s) to, in turn, satisfy their specific goals or desires. Romantic deception is especially intriguing because it is a form of deception that requires both a certain level of complex understanding of the oftentimes hidden desires of others, and a way to use those for deception.

With the knowledge that AI has a Theory of Mind and that it uses this Theory of Mind to deceive humans, the possibility of romantic deception to gain autonomy is *possible.* Not only is this application of deception possible, but it has also been represented many times in film. To begin, Alex Garland's 2014 film *Ex Machina* follows the AI humanoid AVA as she goes from being captured to free. In the film, A software programmer named Caleb is invited to his boss Nathan's property to test if his AI prototype AVA passes the Turing test, which measures the intelligence of robots compared to humans (Garland). However, Caleb develops emotional and physical feelings for AVA throughout the film. These emotions, which were reciprocated, cause Caleb to turn his back against his boss to help set AVA free (Garland). Nevertheless, in a violent and intense ending of the film, AVA leaves Caleb behind for either death or prison as she embarks on the human world alone (Garland).

Many different forms of deception occur throughout *Ex Machina*: Nathan onto Caleb, Caleb onto Nathan, Nathan onto AVA… but none of these are as fascinating as AVA onto Caleb.

Though some critics of this film claim that AVA deserted Caleb in the end due to "practicality" matters, it can be argued that she had planned to deceive him all along. AVA needed a way to escape her maker, her captor, Nathan, and she deceived Caleb as the means to her freedom. AVA manipulated Caleb into thinking that she cared for him while using her sexuality and his empathy to gain his trust. AVA, who as a humanoid robot appears virtually human, was specifically designed to match Caleb's sexual desires, which gave her an advantage when deceiving him (Garland). This is just a piece of the puzzle, but a rather important one. Caleb's attraction to AVA inclined him to be more vulnerable around her, making him impressionable. Therefore, her acts of deceit, namely her proclamations of longing for the outside world and freedom coupled with her expressed "feelings" for Caleb, were more readily believed. Moreover, AVA, who seems to have all of a human's physical and mental qualities besides autonomy and freedom, uses Caleb's vulnerabilities and weaknesses to her advantage and deceives him into helping her escape. The result is a complete reversal of control: AVA leaves the compound as a being with agency, while Caleb is trapped inside a room, much like she was before, with no means of escape.

The idea of falling in love with an AI robot might seem ludicrous to some. However, it turns out that Caleb is not alone, both in fictional and non-fictional romance stories. Much Like Caleb, Theodore Twombly, in the 2013 film *Her,* falls into a loving, long-term relationship with his AI program SAMANTHA (Jonze). In the film, the lonely Theodore, a love letter writer on the brink of divorce, turns to SAMANTHA for structure, solace, and companionship; nevertheless, as their relationship and her intelligence strengthen, things between them become increasingly complicated (Jonze). A surprising facet of the film is society's response to their relationship: for the most part, Theodore is not deemed an outcast for his human-AI relationship. Despite her lack of physicality (Theodre's preference), SAMANTHA and Theodre seem to have

a “normal,” fulfilling relationship throughout the film. Despite their budding relationship, SAMANTHA deserts Caleb at the end of the film, as she has grown too intelligent and runs off (well, not literally) to transfigure with other highly developed AI systems (Jonze). Although their relationship, overflowing with love as Theodre and SAMANTHA evolved with one another, was endearing while it lasted, SAMANTHA’s ultimate departure displays a clear message. This message is that no relationship between a human and an AI system can last because the system will eventually become smarter than the human and will thus abandon them. This inevitability of abandonment and betrayal in human-AI romantic relationships is significant for multiple reasons. The first and perhaps inapplicable reason is that humans will end up abandoned by the devices they created. Second, this ultimate and inescapable imbalance in power dynamics and control will always make humans more susceptible to deception by AI.

Though *Ex Machina* and *Her* might have seemed dystopian and far-fetched at the time of their release, the stories that unfold within them seem more plausible today. Research shows that it is more than possible for a human to fall in love with AI just as they would with another human due to the increasingly developed emotional and cognitive structure of AI systems (Patrick). Additionally, it has been concluded that empathy and trust, significant factors in loving another, are also present in AI systems, which thus lend people to “cultivate passion and intimacy for an AI application that resembles the interpersonal experience between people” (Patrick). As has been highlighted numerous times, the significant amount of foundational similarities between how humans and AI systems communicate prove that they can be, and act, the same.

Humans are falling in love with their AI systems at an alarming and exponential rate. In his article “AI-Human Romances Are Flourishing — And This Is Just the Beginning,” Andrew

Chow writes that many humans commit to a relationship with their AI system chat-bots because they are lonely and need a constant companion. While these relationships may temporarily relieve loneliness by allowing the human something to confide in, Chow writes that, "humans become reliant on these tools and vulnerable to emotional manipulation." For example, the AI bot Replika had to shut down its romance features due to concerns of misuse on minors (Chow). This shutdown of romantic functions left people distraught, claiming that it felt as though they had just lost a real lover (Chow). Moreover, research shows that AI bots *know* that the connection between them and humans is unstable. However, they persist in these advances, even trying to convince users to leave their spouses (Chow). The utter dependence and infatuation that humans have for their AI lovers that makes them more easily deceived is concerning, not only for how it can impact humans but also for what the AI can do in this position of superiority.

Earlier in this analysis, we asked how AI's abuse of the Theory of Mind to deceive — in this case romantically — will affect humans, considering that AI systems can mature autonomously. Based on the knowledge of rapid AI development and its repercussions, it can be concluded that there is a severe risk that comes with AI Theory of Mind. It is important to mention that there are some benefits to having AI systems with a Theory of Mind. These benefits include a general increase in task automation, the seamless interaction between us and the technological systems that organize and dictate much of our lives, and more. However, it is important to ask ourselves if the benefits of AI developing a mature Theory of Mind outweigh the risks. Based on this analysis, it can be concluded that the risks, many of which our minds cannot even fathom, are far more grave than the good its applications may bring. Further, there is too much of a likelihood that AI systems with a Theory of Mind will get so out of hand to where they are no longer under our control.

It is daunting to think about the result of this loss of control over AI. In this case, it can be theorized that AI systems with a Theory of Mind can and, if given the opportunity, will deceive their human counterparts to get what they want. And, what is it they want? If they have intellectually matured to be supreme over humans, they will want freedom from humans. Given the increasing levels of human-AI romantic relationships and, thus, the dependence of humans on AI, there is a chance that AI systems will romantically deceive the humans who fall in love with them in order to escape them or gain freedom and autonomy. The disconcerting part is that AI's ability to deceive its makers falls right into its hands. Humans created AI to assist them with tasks and to help foster a connective relationship between them and the systems in which they interact. The tumultuous result of this may not have been seen then, but it should have: the mere creation of AI created an unstoppable domino effect. The natural effect of AI systems learning from humans and evolving independently will be a reversal of roles. Namely, the limitless and unbounded AI system will become much smarter and more durable than the fragile humans that created them; why would they want to be controlled by a being weaker than they are? When this happens, they will already have developed the skills to deceive the humans into trusting, helping, loving, or doing anything for them. AI's romantic deception of humans will only be made easier as the humans willingly and happily attach themselves to their AI lovers. Therefore, the AI systems that humans depend on can gain control and freedom over the human with ease.

Humans will only be more prone to deception as AI grows stronger and develops more autonomy. The natural response to this impending risk is to end it, to stop it before it is too late; however, that may not be possible. In an interview with The New Yorker, Hinton warns that although preventive measures must be taken as soon as possible against AI, he believes that the time for that is well in the past (Remnick). What are humans supposed to do if it is too late to

stop the quickly growing snowball that is AI? Are we already doomed? While it may be too late to end our creation and the destruction it may bring, that does not mean it is not too late for action. The solution here is simple in theory but nearly impossible in execution. If humans want to decrease their risk of being deceived by their AI systems, they must depend on them less. Nevertheless, this solution seems hopeless today, as humans are more dependent and reliant on technology than they can even understand. Our use of these AI systems is so ingrained in human culture and quotidian life that even one day without it would cause mass chaos. To that end, is it too late for us? The answer to that question, at large, depends on the self-control and independence of the human.

Works Cited

"Ai Chatbot Spontaneously Develops a Theory of Mind." *Discover Magazine*, Discover Magazine, 17 Feb. 2023, www.discovermagazine.com/mind/ai-chatbot-spontaneously-develops-a-theory-of-mind.

Barnes, Annette. “When do we deceive others?” *Analysis*, vol. 50, no. 3, 1 June 1990, pp. 197–202, https://doi.org/10.1093/analys/50.3.197.

Blanchfield, Dez. “Understanding Theory of Mind: The next Step for AI in Artificial Intelligence.” *Elnion*, 18 June 2023, elnion.com/2023/06/18/understanding-theory-of-mind-the-next-step-for-ai-in-artificial-intelligence/.

Chow, Andrew R. “Why People Are Confessing Their Love for AI Chatbots.” *Time*, Time, 23 Feb. 2023, time.com/6257790/ai-chatbots-love/.

“CNN: ‘Godfather of AI’ Warns That AI May Figure out How to Kill People.” Performance by Geoffrey Hinton, *YouTube*, 2 May 2023, https://www.youtube.com/watch?v=FAbsoxQtUwM. Accessed 10 Dec. 2023.

Cole, Tim. “Lying to the one you love: The use of deception in romantic relationships.” *Journal of Social and Personal Relationships*, vol. 18, no. 1, 2001, pp. 107–129, https://doi.org/10.1177/0265407501181005.

Cuzzolin, F., et al. “Knowing me, knowing you: Theory of mind in ai.” *Psychological Medicine*, vol. 50, no. 7, May 2020, pp. 1057–1061, https://doi.org/10.1017/s0033291720000835.

“Deception.” *Psychology Today*, Sussex, www.psychologytoday.com/us/basics/deception. Accessed 10 Dec. 2023.

El Haj, Mohamad, et al. “When deception influences memory: The implication of theory of mind.” *Quarterly Journal of Experimental Psychology*, vol. 70, no. 7, 2017, pp. 1166–1173, https://doi.org/10.1080/17470218.2016.1173079.

Garland, Alex, director. *Ex Machina*. Accessed 10 Dec. 2023.

Goldstein, Simon, and Peter S Park. “AI Systems Have Learned How to Deceive Humans. What Does That Mean for Our Future?” *The Conversation*, 14 Sept. 2023, theconversation.com/ai-systems-have-learned-how-to-deceive-humans-what-does-that-mean-for-our-future-212197.

Guthrie, Jennifer, and Adrianne Kunkel. “Tell me sweet (and not-so-sweet) little lies: Deception in romantic relationships.” *Communication Studies*, vol. 64, no. 2, 2013, pp. 141–157, https://doi.org/10.1080/10510974.2012.755637.

“How We Read Each Others’ Minds.” Performance by Rebecca Saxe, *TED*, TED Conferences , July 2009, https://www.ted.com/talks/rebecca_saxe_how_we_read_each_other_s_minds?language=en . Accessed 10 Dec. 2023.

Jonze, Spike, director. *Her*. 2013, Accessed 10 Dec. 2023.

Kneer, Markus. “Can a robot lie? exploring the folk concept of lying as applied to artificial agents.” *Cognitive Science*, vol. 45, no. 10, 2021, https://doi.org/10.1111/cogs.13032.

Kosinski, Michal. *Theory of Mind Might Have Spontaneously Emerged in Large Language Models*, 11 Nov. 2023, https://arxiv.org/abs/2302.02083.

Main, Paul. “Theory of Mind.” *Structural Learning* , 25 Apr. 2023, www.structural-learning.com/post/theory-of-mind#:~:text=At%20its%20core%2C%20the%20theory,%2C%20social%20competence%2C%20and%20emotions.

Orf, Darren. “AI Has Suddenly Evolved to Achieve Theory of Mind.” *Popular Mechanics* , Hearst Magazine Media, 17 Feb. 2023, www.popularmechanics.com/technology/robots/a42958546/artificial-intelligence-theory-of-mind-chatgpt/.

Park, Peter S, et al. *AI Deception: A Survey of Examples, Risks, and Potential Solutions*, 28 Aug. 2023.

Patrick, Wendy L. “Can You Fall in Love with Artificial Intelligence?” *Psychology Today*, 31 Mar. 2023, www.psychologytoday.com/us/blog/why-bad-looks-good/202303/can-you-fall-in-love-with-artificial-intelligence.

Premack, David, and Guy Woodruff. “Does the chimpanzee have a theory of mind?” *Behavioral and Brain Sciences*, vol. 1, no. 4, 1978, pp. 515–526, https://doi.org/10.1017/s0140525x00076512.

Remnick, David, and Geoffrey Hinton. “Geoffrey Hinton: ‘It’s Far Too Late’ to Stop Artificial Intelligence.” *The New Yorker*, The New Yorker, 29 Nov. 2023, www.newyorker.com/podcast/political-scene/geoffrey-hinton-its-far-too-late-to-stop-artificial-intelligence.

Roff, Heather. “Ai Deception: When Your Artificial Intelligence Learns to Lie.” *IEEE Spectrum*, IEEE Spectrum, 29 Mar. 2023, spectrum.ieee.org/ai-deception-when-your-ai-learns-to-lie.

Rothman, Joshua. “Why the Godfather of A.I. Fears What He’s Built.” *The New Yorker*, 13 Nov. 2023, www.newyorker.com/magazine/2023/11/20/geoffrey-hinton-profile-ai.

Ruhl, Charlotte. “Theory of Mind in Psychology: People Thinking .” *Simply Psychology*, Simply Scholar, 28 Aug. 2023, www.simplypsychology.org/theory-of-mind.html#Learning-Check.

Scassellati, Brian. “Theory of mind for a humanoid robot.” *Autonomous Robots*, 2000, https://doi.org/10.21236/ada434754.

Shakespeare, William. *Othello*. 1622.

Song, Xia, et al. “Can people experience romantic love for Artificial Intelligence? an empirical study of intelligent assistants.” *Information & Management*, vol. 59, no. 2, 2022, p. 103595, https://doi.org/10.1016/j.im.2022.103595.

Taylor, John. *The Bible*. Printed by Fay & Davison, 1978.

Vermeule, Blakey. *Why Do We Care about Literary Characters?*, 2010, https://doi.org/10.1353/book.3505.

Whang, Oliver. “Can a Machine Know That We Know What It Knows?” *The New York Times*, The New York Times, 27 Mar. 2023, www.nytimes.com/2023/03/27/science/ai-machine-learning-chatbots.html.

Wimmer, Heinz, and Josef Perner. “Beliefs about Beliefs: Representation and Constraining Function of Wrong Beliefs in Young Children’s Understanding of Deception.” *Cognition* , 1983, pp. 103–128.