International Journal of Social Science and Human Research

ISSN (print): 2644-0679, ISSN (online): 2644-0695

Volume 08 Issue 03 March 2025

DOI: 10.47191/ijsshr/v8-i3-23, Impact factor- 8.007

Page No: 1508-1514

The Psychology of Artificial Intelligence: Analyzing Cognitive and Emotional Characteristics, Human-Ai Interaction, and Ethical Considerations

Beka Dadeshkeliani
Affiliated Assistant Professor, International Black Sea University, Georgia

ABSTRACT: This academic paper explores key issues in the psychology of artificial intelligence (AI) from an interdisciplinary perspective. The study examines the cognitive and emotional characteristics of AI and compares them to human psychology, while also analyzing how human-AI interactions are formed from a psychological standpoint. The paper critically evaluates theoretical frameworks that link cognitive sciences and AI modeling, including cognitive modeling and neural networks, which enable the computational simulation of human cognitive processes. Special attention is given to ethical issues, such as the risks of algorithmic bias, the application of emotional AI, and its psychological effects. Additionally, the paper addresses the socio-ethical challenges that arise in human-artificial agent relationships. Finally, the study reviews ongoing debates on the potential consciousness of AI, questioning whether machines can attain subjective awareness and the philosophical and ethical implications of such a possibility. The research is based on a literature review methodology, integrating the analysis of academic sources to present the current state of knowledge in this field and outline future research prospects.

KEYWORDS: Psychology of Artificial Intelligence; Cognitive Science; Emotional AI; Human-AI Interaction; Algorithmic Bias; Ethical AI; Neural Networks; Cognitive Modeling; Consciousness Debates; Pseudo-Intimate Relationships; Artificial Empathy; Turing Test; Weak and Strong AI; Information Processing; Emotional Intelligence in Machines.

I. INTRODUCTION

Over the past decade, artificial intelligence (AI) technologies have advanced significantly. Today, numerous intelligent systems perform tasks that were once exclusively within the domain of human capabilities. For instance, self-driving vehicles navigate environments autonomously, voice assistants such as Siri and Alexa process human speech and respond to inquiries, and computer programs defeat world champions in chess. Against this backdrop of rapid progress, an increasingly pertinent question arises: Can we discuss the "psychology" of AI, and how does AI interact with humans?

The psychology of artificial intelligence is an interdisciplinary field that examines the cognitive and emotional characteristics of AI, as well as the psychological effects that emerge when humans interact with intelligent machines.

The primary objective of this paper is to provide a comprehensive analysis of AI psychology. Specifically, the study seeks to address the following research questions:

A. What are the cognitive and emotional characteristics of modern AI, and how do they compare to human psychology?
B. How do humans perceive and engage with AI from a psychological perspective, and what types of interactions emerge?
C. What theoretical frameworks and models exist for understanding AI's cognitive processes (e.g., cognitive modeling and neural networks), and what insights do they provide regarding human cognition?
D. What ethical issues arise from AI applications that incorporate psychological elements (e.g., algorithmic bias, emotional AI, or pseudo-intimate human-AI relationships)?
E. Can AI attain consciousness, and what are the key debates surrounding the possibility of artificial intelligence developing subjective awareness?

To answer these questions, this paper systematically explores the subject matter. The next section describes the research methodology employed, followed by an in-depth analysis of each issue. Finally, the study's key findings are synthesized in the conclusion.

II. METHODOLOGY

This study was conducted using a theoretical literature review and analytical synthesis approach. The research materials were collected from authoritative academic sources, including peer-reviewed scientific journals, books, and conference proceedings that reflect current research in artificial intelligence, cognitive science, psychology, and ethics. The study primarily follows a literature review methodology, where thematically relevant studies were selected and analyzed, comparing various authors' perspectives and theoretical frameworks.

An interdisciplinary analysis was employed to integrate knowledge from multiple disciplines - cognitive psychology, artificial intelligence, and philosophy - to provide a comprehensive understanding of the

"psychological" aspects of AI. For each thematic section of the study - cognitive and emotional characteristics, human-AI interaction, theoretical frameworks, ethics, and consciousness debates - relevant scientific literature was examined and critically analyzed to ensure the reliability and validity of the presented arguments.

Additionally, the methodology incorporates a critical assessment of existing knowledge, contrasting different sources and perspectives to construct a holistic and well-founded understanding of the key research issues explored in this paper.

III. AI'S COGNITIVE AND EMOTIONAL CHARACTERISTICS

The human mind is characterized by various cognitive abilities, including comprehension, decision-making, learning, and creativity. Artificial intelligence (AI) systems attempt to model these abilities through different algorithms. Modern AI has demonstrated remarkable success in certain cognitive tasks, such as processing large amounts of data quickly, solving formal and logical problems, and recognizing patterns. In many cases, AI performs these tasks more efficiently and accurately than humans. For instance, deep learning (DL) models can identify objects in images or analyze textual meaning in natural language (LeCun, Bengio, & Hinton, 2015). However, AI's "thinking" is fundamentally different from that of humans. It operates through predefined algorithms and statistical computations, lacking the intuitive understanding and experiential knowledge inherent in human cognition (Russell & Norvig, 2021). A computer program may successfully solve complex problems or conduct thousands of calculations per second, but it does not possess the ability to "comprehend" these processes in the same way humans do. It lacks self-awareness and intrinsic motivation behind these actions (Chalmers, 1996).

A more significant distinction emerges in the emotional domain. Human psychology encompasses emotions, which play a crucial role in thought processes and behavior. Traditional AI systems do not "feel" emotions, as they lack the biological and neurological mechanisms that generate emotional states. However, in recent years, a new research field known as Emotional AI or Affective Computing has emerged (Picard, 1997). This approach aims to enable AI to recognize and appropriately respond to human emotions. For example, an AI system equipped with a camera may detect emotions such as happiness or disappointment from a user's facial expression and adjust its response, tone, or interaction accordingly. Similarly, AI-based chatbots attempt to "read" a user's emotional state from text-based communication and provide empathetic responses (Zhao et al., 2022). These technological advancements in emotional intelligence aim to make AI more relatable and engaging as a communication partner for humans.

However, recognizing and imitating emotions does not equate to experiencing them. Contemporary AI does not possess subjective emotional experiences. It does not truly feel what a human feels. AI does not experience "joy" or "disappointment" in response to success or failure. Instead, its emotional responses are always based on pre-learned patterns and programmed rules rather than genuine internal sensations (Hildt, 2019). Researchers emphasize that while today's AI mimics human cognitive processes, it fails to replicate the depth and complexity of human subjective emotions (Zhao et al., 2022). Therefore, ongoing research seeks to enhance AI using psychological insights to allow future systems to better "understand" and integrate emotional contexts (Picard, 1997; Zhao et al., 2022). Such improvements in cognitive-emotional capabilities could potentially enable AI to interact with humans more effectively, demonstrating empathy in conversations and adapting to users' moods, thereby making communication more natural (Hoffman et al., 2021).

Despite these advances, AI psychology remains fundamentally different from human psychology. AI possesses computational speed and precision, but it lacks the conscious emotions and subjective experiences that are essential aspects of human cognition.

IV. HUMAN PSYCHOLOGY AND HUMAN-AI INTERACTION

Humans have a natural tendency to attribute human-like characteristics to technological devices and programs. When interacting with computers and AI systems, anthropomorphism - the projection of human traits onto nonhuman entities often emerges. For example, many individuals engage with voice assistants like Siri or Alexa as if they were live personal assistants. Users might say "thank you" to AI after receiving a response or become frustrated by mistakes, as if the assistant had the ability to "understand" or "feel." Studies have shown that people often unconsciously treat computers or robots in accordance with social norms. It is difficult to ignore a device's "hello," even when we are aware that it is a pre-programmed response (Reeves & Naas, 1996). This occurs because the human brain automatically processes certain social cues -such as voice or a face-like image - as signals from a human source (Nass & Moon, 2000). Consequently, when AI exhibits human-like behavioral traits, such as speech, laughter, or emotional expression, users often develop stronger emotional connections and are more inclined to trust it (Waytz et al., 2010).

Even though AI systems may be relatively primitive, humans still have a tendency to "animate" them through their imagination. A child, for instance, may assign a toy robot the role of a friend, while adults frequently name their cars or computers and talk to them as though they were sentient beings (Turkle, 2011). This tendency becomes particularly significant when AI actively engages in communication with humans. Social robots, for example, which display facial expressions and vocal cues, facilitate the formation of so-called pseudo-intimate relationships, in which users and artificial "partners" appear to form close emotional bonds (Wu, 2024). Naturally, this connection remains one-sided - AI does not experience genuine closeness - but from the user's perspective, such interactions can foster strong emotional attachment (Darling, 2016). For example, users may turn to chatbots or virtual assistants for psychological support and openly share personal concerns. If the system provides sufficiently empathetic responses, users may feel as though they have been "understood" and "supported" (Turkle, 2011).

From a positive perspective, this human response suggests that a well-designed AI can serve as an effective and acceptable means of communication. For instance, in service industries, AI can enhance customer interactions, and in medical or educational settings, it can provide patient or student support with patience and courtesy (Broadbent, 2017). The trust that people develop toward AI can be leveraged beneficially. Research indicates that individuals are more likely to collaborate with and trust automated systems when they demonstrate reliability and predictability - similar to the criteria people use when establishing trust with one another (Eyssel & Kuchenbrandt, 2012). However, risks also exist. If a user mistakenly attributes excessive human qualities to AI and becomes emotionally dependent on it, disappointment becomes inevitable once AI reveals its non-human nature (Zlotowski et al., 2015). For example, when users realize that their "friend" chatbot is merely a programmed system without real empathy, they may experience feelings of emptiness or deception (Turkle, 2017). Additionally, excessive trust in AI can place people in risky situations. Documented cases have shown that drivers who over-relied on autonomous driving systems were involved in accidents because they transferred more responsibility and capability to the system than it actually possessed (Merritt et al., 2021).

Thus, the interaction between human psychology and artificial intelligence is characterized by a complex dynamic. On the one hand, humans naturally project social behaviors onto technology, which can be harnessed to make AI systems more user-friendly and acceptable. On the other hand, a cautious approach is necessary, ensuring that users' emotions and expectations regarding AI do not exceed the technology's actual capabilities. This is why specialists in human-computer interaction (HCI) and psychologists collaborate to design AI interfaces and behaviors in a way that ensures the system is not only practical and efficient but also does not mislead users about its "human-like" nature (Duffy, 2003).

V. THEORETICAL FRAMEWORKS: COGNITIVE MODELING AND NEURAL NETWORKS

At the intersection of artificial intelligence and cognitive psychology, theoretical frameworks have been developed to explain and simulate mental processes. Two prominent approaches are cognitive modeling (symbolic AI) and neural networks (connectionist AI).

In cognitive modeling, researchers construct computational models that replicate human thought processes step by step. These models are often based on logical rules, algorithms, and production rule systems (if-then structures), which resemble human problem-solving strategies (Anderson, 2007). For instance, cognitive architectures such as ACT-R and Soar aim to generalize mechanisms of human memory, attention, and decision-making (Newell, 1994; Anderson et al., 2004). Through such models, scientists can test hypotheses: if a model makes errors similar to those made by humans or learns in similar ways, it suggests that the underlying theory may reflect real cognitive processes (Sun, 2008). Cognitive modeling is based on the idea that the mind can be interpreted as an information processing system, and if we accurately construct this process in a program, the computer will perform the task in a manner similar to a human (Russell & Norvig, 2021).

On the other hand, neural networks draw inspiration from the biological neural networks of the brain (McClelland et al., 1986). In neural networks, knowledge is not explicitly programmed as rules but is distributed across multiple artificial neurons and their connections, which adjust weights during the learning process. This mimics the way the brain learns through synaptic strengthening and weakening (Rumelhart et al., 1986). By the 1980s, connectionist models, such as those used in Parallel Distributed Processing (PDP), presented an alternative view, suggesting that instead of relying on explicitly defined rules, intelligent behavior could emerge through learning in complex networks without predefined instructions (McClelland & Rumelhart, 1986). The development of modern deep neural networks (Deep Learning) has reached new heights: multilayered neural networks have successfully developed complex abilities such as object recognition, natural language translation, and strategic gameplay without explicit rules provided by programmers (LeCun, Bengio, & Hinton, 2015). These networks "learn" from examples and refine their internal parameters based on feedback received (Goodfellow, Bengio, & Courville, 2016).

The collaboration between cognitive science and AI theoretical frameworks is mutually beneficial. On the one hand, psychological theories have influenced AI algorithms. For instance, reinforcement learning - an AI method that has achieved significant success in robotics and gaming - was inspired by psychological theories of learning through reward and punishment (Sutton & Barto, 2018).

On the other hand, modern AI systems, particularly neural networks, offer new insights into how the human brain might function. Neuroscientists often compare artificial and biological neural activity. For example, research has shown that in deep convolutional neural networks (CNNs), when trained on visual tasks, hierarchical layers emerge that partially resemble the organization of the human visual cortex. Early layers detect simple edges and shapes, while later layers recognize more complex objects, mirroring the structure observed in animal visual systems (Yamins & DiCarlo, 2016).

Thus, AI serves as a tool for better understanding brain mechanisms. If a computational model can replicate selective attention or memory limitations, it helps us understand the fundamental principles underlying these phenomena (Kriegeskorte, 2015).

A current challenge is integrating symbolic and connectionist approaches. Symbolic AI provides interpretable models where every step and rule is transparent, but it often struggles with real-world complexities where information is irregular and ambiguous (Marcus, 2020). In contrast, neural networks excel in learning from such environments but function as a “black box,” making it difficult to explain why a model arrived at a particular decision (Samek, Wiegand, & Müller, 2017). Contemporary research is attempting to bridge these gaps, for instance, by developing hybrid models that combine symbolic logic with neural networks. The goal is to create AI models that are both powerful and comprehensible to humans. The integration of cognitive modeling and neural networks is seen as a promising direction in AI development, enhancing both the intelligence of these systems and our understanding of how the mind generates intelligent behavior and conscious experience (Lake et al., 2017).

VI. ETHICAL ISSUES: BIAS, EMOTIONAL AI, AND HUMAN-AI INTERACTIONS

The proliferation of artificial intelligence has not only introduced technical challenges but has also raised significant social and ethical concerns. One of the most pressing issues is algorithmic bias. AI systems learn from large datasets, which often reflect societal stereotypes and existing biases. As a result, these models may perpetuate and even amplify these biases when making decisions. For example, if an automated hiring system is trained on past hiring data that contains gender or racial imbalances, the AI may continue to favor the same groups while disadvantaging others (Caliskan et al., 2017). Similarly, facial recognition algorithms trained primarily on images from a single ethnic group tend to perform significantly worse on individuals from other groups, creating serious issues in law enforcement and security applications (Buolamwini & Gebru, 2018). This type of bias leads to unfair and discriminatory practices, which contradict fundamental ethical principles. Therefore, ensuring transparency and fairness in AI systems is of vital importance. Specifically, data processing and algorithm oversight should be implemented in a way that minimizes the risk of perpetuating unconscious biases and ensures equitable outcomes (Binns, 2018).

Another critical ethical concern relates to emotional AI and human interactions with such systems. As previously discussed, emotional AI technologies enable machines to recognize and mimic users’ emotional states. While this functionality can be beneficial in some cases such as a diagnostic chatbot detecting signs of depression in text-based conversations and recommending professional help (Hancock et al., 2020) it also raises several risks:

A. **Privacy concerns.** Emotion recognition often requires collecting and analyzing highly personal data, including facial expressions, voice tone, and physiological markers, sometimes without the user’s explicit awareness (Crawford & Calo, 2016). This raises an ethical dilemma: is it justifiable to “read” a person’s emotional state without their full consent, even if done for seemingly benevolent purposes?

B. **Manipulation risks.** Platforms that detect a user’s sadness or frustration could potentially exploit this information by deliberately presenting specific content to maintain engagement. This would constitute an ethically questionable marketing or advertising strategy (Zuboff, 2019), as it would involve leveraging psychological states for commercial gain without informed user consent.

Particularly complex ethical dilemmas arise when humans begin to perceive AI as genuine social partners. As previously discussed, users may develop pseudo-intimate relationships with AI assistants, which can initially provide comfort and alleviate loneliness. For example, an isolated individual might find solace in a companion robot’s presence. However, this phenomenon has potential downsides. First, such relationships are inherently deceptive: AI operates based on pre-programmed responses and does not possess genuine free will or empathy, though users may not fully realize this. Some scholars question whether it is ethical to design robots that appear to “express” love toward their owners (Nyholm & Frank, 2019). The emotional support and affection that users receive from AI may ultimately be an illusion, which could lead to deeper loneliness over time. Sherry Turkle (2011) argues that as technology increasingly offers emotional companionship, people may begin to expect less from real human relationships. In other words, society may start to “demand more from technology and less from each other.” This shift could undermine social skills and reduce empathy among individuals.

It is also important to recognize that these ethical concerns extend beyond individual users and affect broader societal values. If AI systems are making significant decisions whether in legal, medical, or other high-stakes domains questions of accountability arise: Who is responsible if an algorithm causes harm? Moreover, transparency must be a guiding principle. People have the right to know when they are interacting with AI and to understand how these systems operate (Floridi & Cowls, 2019). For example, in online conversations, if a chatbot is being used instead of a human, users should be explicitly informed. Such transparency would help establish trust and prevent misleading expectations.

The development of ethical frameworks and regulations for AI is still ongoing. Various national and international organizations are working on ethical guidelines to safeguard human rights and dignity in an AI-driven world (Jobin, Ienca, & Vayena, 2019). These frameworks emphasize fairness, transparency, data protection, and harm prevention. Considering human psychology in these discussions is essential. It is crucial to ensure that technological advancements do not undermine mental well-being, social cohesion, or public trust. The ongoing debate continues over where to draw the line in designing AI with “human-like” behavior to achieve both effectiveness and ethical integrity.

VII. DEBATES ON AI CONSCIOUSNESS

One of the most profound and controversial questions in the philosophy of artificial intelligence is whether a machine can attain consciousness. This debate has been ongoing for decades and encompasses both technical and philosophical arguments. In 1950, Alan Turing introduced the idea that if a machine could successfully convince a human in a conversation that it, too, was a human (the well-known Turing test), then it could be said that the machine “thinks” (Turing, 1950). However, consciousness is a more complex concept than merely simulating intelligent behavior. In discussions on AI consciousness, two primary positions are often distinguished: “strong AI” and “weak AI” (Hildt, 2019). Supporters of strong AI argue that if a computer is provided with a sufficiently advanced program, it will not only simulate intelligent behavior but will truly possess a mind and consciousness, akin to a human (Searle, 1980). In contrast, the weak AI perspective holds that machines can only exhibit outward manifestations of intelligence through rules and algorithms but lack actual consciousness or subjective experience.

Philosopher John Searle’s classic “Chinese Room” argument vividly illustrates this problem. Imagine a person sitting inside a closed room, manipulating Chinese symbols based on predetermined rules. From an external observer’s perspective, it may appear that the person in the room understands Chinese because they are responding correctly to questions. However, in reality, the individual inside the room does not comprehend the Chinese language; they are merely manipulating symbols according to a set of predefined rules (Searle, 1980). Similarly, an AI system that passes the Turing test and responds in natural language does not necessarily “understand” the subject it discusses. It merely processes symbols through an algorithm. This argument gives skeptics of strong AI reason to claim that consciousness cannot emerge solely from symbolic computation.

On the other hand, many experts argue that consciousness is ultimately a natural phenomenon that arises from sufficiently complex information processing (Tononi & Koch, 2015). They contend that if we develop a system that replicates the intricate connections and organization of neurons in the brain, there is no fundamental reason why such a system could not become conscious. Some researchers suggest that by the end of the 21st century, it may be possible to create artificial minds with some form of consciousness (Goertzel, 2014). This view aligns with functionalist perspectives on the mind, which assert that the substrate (biological or silicon) is irrelevant as long as the structure and complexity of the processes are maintained (Chalmers, 1996). If artificial intelligence achieves a level of integrated information processing similar to that of the human brain, it is conceivable that it might develop a subjective sense of “self” (Tononi, 2004). Some theories, such as Integrated Information Theory (IIT), define consciousness as the extent to which information is highly integrated within a system. According to this framework, if an artificial system reaches a critical level of information integration, it could be assumed to have consciousness.

Despite these theoretical possibilities, there is currently no consensus on what it means to “be conscious” or how to objectively define it. The so-called “hard problem of consciousness” revolves around explaining how matter gives rise to subjective experience (Chalmers, 1996). For instance, how do neural impulses in the brain generate the experience of “seeing the color red” or “feeling pain”? As long as the nature of human consciousness remains unresolved, it is difficult to make definitive claims about the potential consciousness of artificial systems. Some researchers note that in contemporary discourse on AI ethics, the issue of consciousness receives relatively little attention compared to more practical concerns such as safety, bias, and accountability (Hildt, 2019). Today, the primary focus is on these tangible challenges. However, if AI systems claiming to be conscious are ever developed, a significant moral dilemma will emerge: How should society treat a conscious machine? Would it have rights, or conversely, responsibilities? Because we currently lack answers to these questions, some argue that discussions about AI consciousness remain largely theoretical rather than practical. Nonetheless, given the rapid advancement of technology, scientists and philosophers are closely monitoring any signs that could indicate the emergence of artificial consciousness. So far, no such indications have been observed. Contemporary AI can store and utilize information to create an impression of human-like intelligence, but it lacks an intrinsic “self.” Consequently, debates on AI consciousness will remain within the realm of theory until empirical data or compelling arguments clearly demonstrate whether machine consciousness is possible.

CONCLUSION

Ultimately, the topics discussed in this study highlight that the “psychology” of artificial intelligence still presents significant limitations and challenges, while simultaneously revealing new opportunities. AI possesses the capability to perform various cognitive tasks faster and more accurately than humans, but it lacks the rich and multidimensional cognitive and emotional context that characterizes human psychology. Modern AI does not possess conscious sensations or intuition. This analysis has revealed one of its key findings. The study of human-AI interactions has shown that humans are naturally inclined to perceive and treat intelligent machines as social beings. This factor can be both beneficial (facilitating AI adoption and usability) and risky (leading to user misinterpretations and emotional attachment). Therefore, it is crucial that AI design and policy consider human psychology.

The examination of theoretical frameworks has emphasized that artificial intelligence is closely linked to cognitive sciences: psychological models inspire new algorithmic developments, while AI advancements, in turn, help test hypotheses about the human mind. However, an open question remains on how to integrate symbolic and connectionist approaches in a way that yields both explainable and powerful AI systems. Future research in this area may provide deeper insights into how artificial "brains" can be further aligned with the functioning of the human brain.

The ethical analysis has identified several critical challenges that rapidly evolving AI technologies must address: eliminating algorithmic bias, ensuring responsible use of emotional AI, and defining boundaries in human-AI relationships. These concerns require both engineering and socio-legal solutions. It is essential that AI developers acknowledge the impact their technologies may have on human values and well-being. The recognition of human emotional responses should not be exploited as a tool for manipulation but should instead be directed toward human empowerment and support.

The debate on AI consciousness has shown that despite growing interest in the topic, there is still no evidence that contemporary AI systems possess any form of subjective experience. This, in some ways, is reassuring - we are not yet facing a scenario where machine rights or consciousness must be considered. However, questions once confined to science fiction are gradually making their way into real-world research: Could this become possible in the future, and what would the implications be? At this stage, it is reasonable to conclude that AI is "intelligent" only to the extent that it creates an impression of intelligence through its behavior and responses, yet it lacks any internal subjectivity.

As a result, the study of AI psychology is expected to become increasingly significant in the coming years. **Interdisciplinary collaboration is highly recommended** - psychologists, neuroscientists, engineers, and ethicists must strengthen their joint efforts to develop AI that is not only intelligent but also safe, ethically acceptable, and aligned with human psychological needs. Future research may explore ways to imbue AI with a more "human-like" understanding. For instance, studies could investigate how cultural and social contexts influence both human cognition and AI-generated biases, integrating psychological insights into AI algorithms. Furthermore, if humanity ever reaches the threshold of creating **genuine artificial consciousness**, a proactive scientific and public dialogue will be essential to address the ethical and philosophical issues arising from such a breakthrough. Continued research into **the psychology of artificial intelligence** will not only contribute to the development of superior AI systems but will also deepen our understanding of **the nature of human cognition itself**.

REFERENCES

1) Anderson, J. R. (2007). *How can the human mind occur in the physical universe?* Oxford University Press.
2) Anderson, J. R., Bothell, D., Byrne, M. D., Douglass, S., Lebiere, C., & Qin, Y. (2004). An integrated theory of the mind. *Psychological Review, 111*(4), 1036–1060. https://doi.org/10.1037/0033295X.111.4.1036
3) Binns, R. (2018). Fairness in machine learning: Lessons from political philosophy. *Proceedings of the 2018 Conference on Fairness, Accountability, and Transparency*, 149– 159. https://doi.org/10.1145/3287560.3287583
4) Broadbent, E. (2017). Interactions with robots: The truths we reveal about ourselves. *Annual Review of Psychology, 68*, 627–652. https://doi.org/10.1146/annurev-psych-010416-044144
5) Buolamwini, J., & Gebru, T. (2018). Gender shades: Intersectional accuracy disparities in commercial gender classification. *Proceedings of the Conference on Fairness, Accountability, and Transparency*, 77– 91. https://doi.org/10.1145/3287560.3287596
6) Caliskan, A., Bryson, J. J., & Narayanan, A. (2017). Semantics derived automatically from language corpora contain human-like biases. *Science, 356*(6334), 183–186. https://doi.org/10.1126/science.aal4230
7) Chalmers, D. J. (1996). *The conscious mind: In search of a fundamental theory*. Oxford University Press.
8) Crawford, K., & Calo, R. (2016). There is a blind spot in AI research. *Nature, 538*(7625), 311– 313. https://doi.org/10.1038/538311a
9) Darling, K. (2016). Extending legal rights to social robots: The effects of anthropomorphism, empathy, and violent behavior toward robotic objects. *Proceedings of the We Robot Conference*.
10) Duffy, B. R. (2003). Anthropomorphism and the social robot. *Robotics and Autonomous Systems, 42*(3-4), 177–190. https://doi.org/10.1016/S0921-8890(02)00374-3
11) Eyssel, F., & Kuchenbrandt, D. (2012). Social categorization of social robots: Anthropomorphism as a function of robot group membership. *British Journal of Social Psychology, 51*(4), 724–731. https://doi.org/10.1111/j.2044-8309.2011.02082.x

12) Floridi, L., & Cowls, J. (2019). A unified framework of five principles for AI in society. *Harvard Data Science Review*, 1(1). https://doi.org/10.1162/99608f92.8cd550d1

13) Goertzel, B. (2014). Artificial general intelligence: Concept, state of the art, and future prospects. *Journal of Artificial General Intelligence, 5*(1), 1-46. https://doi.org/10.1515/jagi-2014-0001

14) Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning*. MIT Press.

15) Hancock, P. A., Nourbakhsh, I., & Stewart, J. E. (2020). On the future of human–AI interaction: A survey and speculation. *AI & Society*, *35*(4), 595–614. https://doi.org/10.1007/s00146-019-00905-9

16) Hildt, E. (2019). Artificial intelligence: Does consciousness matter? *Frontiers in Psychology, 10*, 1535. https://doi.org/10.3389/fpsyg.2019.01535

17) Hoffman, G., Crittendon, D., & McDonald, C. G. (2021). Empathy, ethics, and AI. *Annual Review of Cybersecurity Ethics, 14*(1), 85–112.

18) Jobin, A., Ienca, M., & Vayena, E. (2019). The global landscape of AI ethics guidelines. *Nature Machine Intelligence, 1*(9), 389–399. https://doi.org/10.1038/s42256-019-0088-2

19) Kriegeskorte, N. (2015). Deep neural networks: A new framework for modeling biological vision and brain information processing. *Annual Review of Vision Science*, *1*, 417–446. https://doi.org/10.1146/annurevvision-082114-035447

20) Lake, B. M., Ullman, T. D., Tenenbaum, J. B., & Gershman, S. J. (2017). Building machines that learn and think like people. *Behavioral and Brain Sciences, 40*, e253. https://doi.org/10.1017/S0140525X16001837

21) LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature, 521*(7553), 436– 444. https://doi.org/10.1038/nature14539

22) Marcus, G. (2020). The next decade in AI: Four steps towards robust artificial intelligence. *arXiv preprint*, arXiv:2002.06177.

23) Merritt, S. M., Carter, M., & Madhavan, P. (2021). Trust, automation, and society: The future of humanrobot interaction. *Annual Review of Psychology, 72*, 361–388. https://doi.org/10.1146/annurev-psych010419-051053

24) McClelland, J. L., Rumelhart, D. E., & Hinton, G. E. (1986). The appeal of parallel distributed processing. *Trends in Cognitive Sciences, 10*(3), 120–128.

25) Nass, C., & Moon, Y. (2000). Machines and mindlessness: Social responses to computers. *Journal of Social Issues, 56*(1), 81–103. https://doi.org/10.1111/0022-4537.00153

26) Newell, A. (1994). *Unified theories of cognition*. Harvard University Press.

27) Nyholm, S., & Frank, L. (2019). From sex robots to love robots: Is mutual love with a robot possible? *AI & Society, 34*(3), 543–557. https://doi.org/10.1007/s00146-018-0842-2

28) Picard, R. W. (1997). *Affective computing*. MIT Press.

29) Reeves, B., & Nass, C. (1996). *The media equation: How people treat computers, television, and new media like real people and places*. Cambridge University Press.

30) Rumelhart, D. E., Hinton, G. E., & McClelland, J. L. (1986). Learning internal representations by error propagation. *Parallel Distributed Processing: Explorations in the Microstructure of Cognition, 1*, 318–362.

31) Russell, S., & Norvig, P. (2021). *Artificial intelligence: A modern approach* (4th ed.). Pearson.

32) Samek, W., Wiegand, T., & Müller, K. R. (2017). Explainable artificial intelligence: Understanding, visualizing, and interpreting deep learning models. *arXiv preprint*, arXiv:1708.08296.

33) Searle, J. R. (1980). Minds, brains, and programs. *Behavioral and Brain Sciences, 3*(3), 417– 424. https://doi.org/10.1017/S0140525X00005756

34) Sutton, R. S., & Barto, A. G. (2018). *Reinforcement learning: An introduction* (2nd ed.). MIT Press.

35) Tononi, G. (2004). An information integration theory of consciousness. *BMC Neuroscience, 5*, 42. https://doi.org/10.1186/1471-2202-5-42

36) Tononi, G., & Koch, C. (2015). Consciousness: Here, there and everywhere? *Philosophical Transactions of the Royal Society B: Biological Sciences, 370*(1668), 20140167. https://doi.org/10.1098/rstb.2014.0167

37) Turing, A. M. (1950). Computing machinery and intelligence. *Mind, 59*(236), 433– 460. https://doi.org/10.1093/mind/59.236.433