Proceedings of the 2nd International Conference on Machine Learning and Automation
DOI: 10.54254/2755-2721/97/20241408

The Mindset of AGI Unexplainability in Human-computer Interaction Scenarios from Turing's "Computing Machinery and Intelligence"

Heyang Chen

School of AI and Advanced Computing, Xi'an Jiaotong-Liverpool University, Suzhou, Jiangsu, 215006, China

Heyang.Chen21@student.xjtlu.edu.cn

Abstract. The emergence of AI-generated content (AIGC) can be traced back to as early as 1950, when Alan Turing introduced the famous "imitation game" in his paper *Computing Machinery and Intelligence*, which proposed a method to determine whether a machine possesses "intelligence." Since the introduction of the GAN model by Goodfellow et al. in 2014, the issue of autonomy in AIGC has not seen any breakthroughs. However, due to the challenges posed by robustness and the lack of explainability, society is already beginning to anticipate the social issues and anxieties that might arise with the advent of autonomous artificial general intelligence (AGI). The increasing influence of AI technology on society has further driven concerns about the ethical implications of both AIGC and AGI. Specifically, the relationship between human-computer interaction (HCI) and AI ethics—particularly the role of explainable AI—has become increasingly crucial. Merely understanding the issue of non-explainability from a technical standpoint is no longer sufficient to form a principled basis for AI ethics. In fact, Turing to some extent foresaw the possibility that AI's future development would face such issues. This paper seeks to offer a direction that moves beyond the traditional AI ethics research framework by reinterpreting Turing's original question and analyzing some of the objections he raised. The goal is to provide a new mindset for exploring the necessary modes of thinking for human-computer interaction in the era of AGI.

Keywords: Turing Test, Unexplainability, Artificial General Intelligence, Human-Computer Interaction, Explainable AI.

1. Introduction

At present, AIGC (AI-generated content) typically focuses on specific creative tasks, corresponding to narrow AI, also known as Artificial Narrow Intelligence (ANI). Unlike AIGC, AGI (also known as strong AI) is capable of addressing a wide range of tasks without explicit programming for each individual task [1]. AGI represents one of the ultimate goals of AI research; however, its development is accompanied by the challenge of explainability issues. The ethical concerns associated with non-explainable AI, such as transparency, bias, privacy security, explainability, and trust [2], arise due to the difficulty in interpreting the "black-box" nature of deep neural networks. These ethical challenges and requirements posed by non-explainability must be addressed in the design of human-computer interaction systems. Otherwise, AGI will remain a theoretical concept and cannot be fully deployed in

DOI: 10.54254/2755-2721/97/20241408

human society for production and daily life. As autonomous and anthropomorphic AI becomes increasingly prevalent, and as the potential realization of AGI approaches, these challenges and ethical demands will only grow in importance.

Efforts to address these challenges and requirements have primarily focused on research into explainability, which is fundamental to the ethical development of AI. Explainability enables stakeholders—including end users, regulators, and developers—to understand the underlying principles behind AI-generated outcomes. From this, the concept of Explainable Artificial Intelligence (XAI) has emerged, which is considered critical for establishing trust, accountability, and transparency in AI systems [3]. To some extent, contemporary society finds itself in a similar position to that of Turing: constrained by technology, Turing could only speculate about future developments through a series of assumptions about machines that could pass the imitation game. Likewise, the black-box nature of deep learning remains unresolved, necessitating that research into explainability move beyond mere technical understanding. This paper will explore the intrinsic relationship between Turing's perspectives and the issue of AI explainability, proposing a reinterpretation of Turing's original question "Can machines think?" while seeking to answer its reformulation from a broader, non-technical viewpoint.

2. Replacing Turing's original question "Can machines think?"

Professor Yang Qingfeng from the Institute of Technology Ethics for Human Future at Fudan University has pointed out that there are two typical approaches in the diverse paths of AI ethics [4]: one centered on human authority and another based on technicism, with XAI representing a typical example of the technicism approach. Another study suggests that explainability is not an inherent attribute of a model, but rather depends on the perception and capacity of the individual receiving the explanation [5]. Even when a model is made fully transparent, it does not guarantee that the recipient will grasp all the information or avoid feeling overwhelmed. This understanding is influenced by the recipient's current knowledge, the purpose for which the explanation is sought, and other human factors. Therefore, research in XAI needs to be human-centered, which aligns with the other approach identified by Yang. In his paper, Turing posed the question "Can machines think?" and addressed it by presenting nine "Contrary Views" (Objections) as shown in table 1:

Table 1. Contrary Views on the Main Question

1	The Theological Objection
2	The 'Heads in the Sand' Objection
3	The Mathematical Objection
4	The Argument from Consciousness
5	Argument from Various Disabilities
6	Lady Lovelace's Objection
7	Argument from Continuity in the Nervous System
8	The Argument from Informality of Behaviour
9	The Argument from Extra-Sensory Perception

Among the objections raised by Turing, five of them (1, 2, 4, 5, and 6) largely revolve around the subjective thinking of humans at that time, which bears a strong resemblance to the human-centered perspective in XAI research. This further supports the notion that the development of XAI is fundamentally a design challenge, where HCI researchers and design practitioners can offer insights, solutions, and methodologies by understanding users' needs and expectations when interacting with AI systems, thereby making AI more explainable [5]. Moreover, in a future where technology becomes more advanced and AI ethics more refined, updates and further interpretations of these objections will likely remain aligned with their core principles. Based on this consideration, this paper suggests that Turing's thoughts can guide human approaches to addressing explainability issues in HCI applications.

DOI: 10.54254/2755-2721/97/20241408

From the perspective of the generational development of AI, Turing's definition of a "conceivable machine capable of passing the Turing Test" and the concept of discrete state machines will ultimately be replaced by AGI. In this paper's framework, such digital computers are endowed with sufficient storage capacity, fast enough processing speeds, and are appropriately programmed to possess autonomy, with their storage capacity clearly exceeding 10^9 units. For example, GPT-3, which boasts 175 billion parameters, exemplifies this evolution. While the number of parameters cannot directly translate into storage capacity in the traditional sense, it reflects the model's ability to "store" knowledge from the vast datasets on which it was trained. Turing believed that similar technological breakthroughs render the original question "Can machines think?" meaningless, though he did not explicitly clarify why. Instead, he introduced the concept of a "learning machine" to explain this viewpoint. He posited that once the necessary engineering advances were made, programming a simulation of a child's brain and subjecting it to proper education could eventually result in an adult brain. This process consists of three parts [6]:

(a) The initial state of the mind, say at birth,

(b) The education to which it has been subjected,

(c) Other experience, not to be described as education, to which it has been subjected.

Similarly, the evolution of AIGC can also be divided into three stages: the auxiliary stage, the assistance stage, and the autonomous stage [7], which aligns with Turing's hypothesis. The auxiliary stage corresponds to the early development of AIGC technology, where AI serves primarily as a tool. During this phase, AI models are small in scale, with simple architectures and limited learning capabilities, resembling a child's brain. Correspondingly, their explainability is higher—in fact, they do not face issues of non-explainability seen in more advanced models. These early models rely on predefined templates or rule-based frameworks for simple content generation, falling far short of the flexible and realistic content creation seen today. Next, the assistance stage corresponds to the current landscape of AIGC technology, where large-scale pretraining aligns with what Turing referred to as "education." These pretrained models possess a vast number of parameters, and as a result, their explainability decreases. Lastly, the autonomous stage corresponds to the era of AGI, where models exhibit autonomy and generality. AIGC can perform real-time perception, precise cognition, and autonomous content creation. The relationship between AI and content generation evolves from a tool-based auxiliary or assistive role to one where virtual entities engage in content creation and interaction, akin to an adult brain.

It can be argued that the more "intelligent" a model becomes, the less explainable it is. Turing noted that an important feature of such a learning machine is that the teacher often does not fully understand what is happening inside, yet can still predict its behavior to some degree. This statement has two interpretations: the former indicates that the brain is a black box, while the latter reflects the relationship between explainability and the recipient. This ties back to the earlier discussion of XAI, where the issue of non-explainability needs to move beyond the technical framework and adopt a human-centered approach. An example would be when users engage in conversations with a large language model: even if they do not fully understand how the model works, it is not difficult for them to anticipate receiving a response after sending a message. Human-centered XAI shifts the focus of explainability from the AI system to the user, emphasizing that the design of AI technologies should focus on people's needs and use user experience (UX) as a criterion to evaluate each instance of HCI [5].

The arguments so far seem to dismiss Turing's original question, "Can machines think?" outright. More precisely, the question has not been "replaced" but rather negated. This may stem from an insufficiently clear definition of the "subject." The next section of this paper will delve further into the issue of the "subject" and will attempt to explain why pondering the question "Can machines think?" is meaningless while proposing a relatively suitable replacement.

3. The objections from theology and fear: the philosophical conundrum posed by explainability

The shift in the subject of explainability is essentially a move toward understanding artificial intelligence's "intelligence" from a functionalist perspective. Stuart J. Russell, in *Artificial Intelligence: A Modern Approach*, introduced the concept of the rational agent [8]. A rational agent is defined as an

DOI: 10.54254/2755-2721/97/20241408

entity that acts to achieve the best possible outcome or, in uncertain conditions, the best-expected outcome based on the available information. Rational agent frames AI as systems that perceive their environment and take actions to maximize their success according to a performance measure. Russell effectively defines intelligence through the relationship between the agent and its application scenarios, thereby avoiding metaphysical issues [4]. Unlike general tools invented by humans so far, AI not only performs cognitive functions like perception, cognition, decision-making, and action but also has unique affective computing capabilities in HCI, such as emotions and empathy. Based on these two functions, this paper proposes two possible categories for HCI scenarios in the AGI era: the first is **cognitive scenarios**, where AGI exercises its cognitive functions, and the second is **affective scenarios**, where AI performs emotional or affective computations. The human-centered XAI approach, using UX as the evaluation standard for HCI, as mentioned at the end of Section 2, becomes relatively reasonable under this definition. These scenarios encompass a substantial portion of human activities, such as high-risk decision-making tasks and the handling of intimate relationships. In such cases, the evaluation of task completion is objective and even quantifiable, without involving human subjective consciousness. For instance, in the case of AI-assisted driving, user experience can be assessed through factors such as route selection, travel time, and incident-free driving. It is easy to assess the performance of an autonomous driving system and design a fixed framework and paradigm that can be applied to all autonomous driving scenarios. This kind of evaluation system for cognitive scenarios effectively addresses the key issues that mainstream XAI seeks to overcome, fostering user trust in human-machine interactions without requiring technical understanding of the AI's decision-making process.

The limitations of the technological path thinking can lead to anxiety and fear in human-computer interaction, particularly as AI functionalities continue to evolve. This section will discuss Turing's "The Theological Objection" and "The 'Heads in the Sand' Objection" together, as both arguments are closely linked to these emotional of fear. The theological objection asserts, "God has given an immortal soul to humans, but not to any machine; hence, no machine can think." This view suggests that thinking is a trait that marks human superiority. While this explanation may not be rigorous, human superiority is indeed challenged by AIGC. In human-computer interaction, a proper understanding of how a system works is referred to as the user's "mental model" [9]. From a functionalist perspective, users' understanding cannot be improved simply by explaining the system's inner workings. With the added dimension of losing the sense of human superiority, users' mental models may fail to resolve cognitive dissonance, ultimately leading to anxiety and fear in human-computer interaction. In other words, this relates to the issue of AI replacement, where the development of self-awareness and automation may result in feelings of being superseded. In the ethics of artificial intelligence, the transcendence problem may arise when AGI develops capabilities that surpass current regulatory, moral, or social norms. This development complicates efforts to govern or predict AGI behavior using existing legal or ethical frameworks, underscoring the necessity for a new conceptual framework that accommodates the unprecedented complexity and autonomy of AGI systems.

In the context of AGI, discussing these two objections—"The Theological Objection" and "The 'Heads in the Sand' Objection"—is highly appropriate. The challenge to human superiority essentially revolves around the transcendence problem, which touches upon the ultimate direction of AI, namely AGI. The fear stemming from the transcendence problem is situated in the realm of mental or psychological activity, specifically concerning whether machine intelligence can surpass human intelligence. When examined solely from the perspective of intelligence, it becomes clear that the AI replacement issue is an inevitable derivative of the transcendence problem—a deeper or more tangible manifestation of it, and ultimately a philosophical issue with social implications. The replacement problem raises the question of whether human practical activities can be replaced by intelligent machines. Turing expressed hope that machines would eventually be able to compete with humans in purely intellectual domains. He mentioned that activities such as chess might be a good starting point. In 1997, IBM's "Deep Blue" defeated world chess champion Garry Kasparov, demonstrating that AI can replace human activities in certain areas, functioning as an independent agent to some degree. A report suggests that in the coming decades, up to 30% of jobs could potentially be automated by AI [10]. This indicates

DOI: 10.54254/2755-2721/97/20241408

that the replacement problem has broader societal impacts, such as economic inequality and unemployment. If certain groups benefit more from these technological advancements, it could exacerbate existing societal divisions. Thus, it is necessary to acknowledge both the inevitability of AI's autonomous self-awareness and the possibility of its role in replacing human knowledge production. This ontological view ultimately leads back to the problem of AI's explainability. By shifting the subject, the agency and subjectivity of intelligent systems, rather than merely their technical evolution, become the focus by moving away from theological objections. This reorients the conversation towards a new question: "What are we actually communicating with?" This question represents the replacement of Turing's original inquiry, "Can machines think?" While Turing's question seeks an objective answer centered around intelligent machines, the new question introduces a subjective perspective, with answers that vary depending on the user experience in different HCI scenarios, aligning with the human-centered exploration of XAI. In the remainder of this paper, the "new question" will refer to this: "What are we actually communicating with?"

4. The argument from consciousness and various disabilities

The subjective adaptability of the new question can be further explained through the argument from consciousness. This argument posits that machines cannot truly think or understand because they lack consciousness and subjective experience. Turing acknowledged the mystery of consciousness but suggested that a complete understanding of it is unnecessary for recognizing machine intelligence. It is essential to distinguish between consciousness and perception: robots can have perception, and even share perceptual abilities similar to those of humans. The explanation of the first part of this statement is that AI perception refers to the ability of artificial intelligence systems to interpret and understand information from their environment, akin to how humans perceive the world around them. This involves the use of sensors and data processing techniques, allowing machines to recognize objects, understand speech, and interpret visual and auditory signals [11]. Simply put, robots can be constructed with the ability to see and hear. To some extent, they can interpret what they perceive. For instance, machines designed to analyze facial images can determine whether the faces express happiness, sadness, anger, or other emotions. They can rationally and accurately identify the emotional state of a user. The second part of the statement suggests that AI can exhibit behavior patterns similar to humans while performing these perceptual tasks, though the internal representational structures may differ. According to Russell's view, rational agents select optimal perceptual outcomes based on specific contexts. Human-centered XAI requires users to focus on the relationship between AI behavior and the application scenario. Under the framework of the new question, intelligent machines are viewed as alienated, independent entities distinct from humans, yet possessing spatial memory representational structures similar to humans (the possibility of autonomous consciousness) [4, 12].

This homomorphism can be understood from the user's subjective perspective. The objection from consciousness argues that emotional recognition models do not truly grasp the meaning behind their tasks—programs that assess emotions from an image do not actually experience those emotions. Turing quoted Professor Jefferson's statement: "Not until a machine can write a sonnet or compose a concerto because of thoughts and emotions felt…could we agree that machine equals brain—that is, not only write it but know that it had written it" [6]. In Section 3 of this paper, two types of HCI scenarios in the AGI era were defined. While the earlier discussion focused on the first type, this section uses the second scenario—affective scenarios—as an example to distinguish perception from consciousness. Specifically, it explores the intrinsic relationship between the argument from consciousness and the human-centered approach through the lens of human-AI intimate relationships. The reason affective computing is classified as a major category in AGI applications is that it is considered a key driver in developing human-centered AI and human intelligence: people expect to build cognitive intelligent systems in HCI scenarios that can distinguish and understand human emotions while providing sensitive and friendly real-time responses [13]. In typical intimate relationships, Timmerman emphasizes the importance of mutual trust and emotional closeness, as well as the necessary conditions for intimacy, such as self-disclosure and open communication [14]. In intimate relationships, both parties should at

DOI: 10.54254/2755-2721/97/20241408

least be “capable of interaction”—for instance, between two people or between a person and a pet. This paper posits that such “interaction” involves a collision of consciousness between the two parties, but given the shift in the subject of explainability, only the consciousness of the HCI user needs to be considered. The explanation of AI, therefore, pertains to the explanation of human-like behavior [4]. This leads to a simplistic definition of AI as “thinking and acting like humans,” which reflects how users subjectively evaluate their experience in human-AI intimate relationships within affective scenarios.

Intimacy plays a critical role in individual identity formation, psychosocial development, and satisfaction with social support [14]. Human-AI intimate relationships, as envisioned in the AGI era, could significantly impact the social cognition that individuals form over extended periods of social experience. Applying the UX-based evaluation system for HCI, it is evident that compared to cognitive scenarios, the outcomes in affective scenarios are nearly impossible to quantify in a unified way. This is because personal values in intimate relationships must be validated through consensus, allowing individuals to feel understood and accepted within the relationship [14]. Such feelings are subjective and vary from person to person. At this point, answering the question “What are we actually communicating with?” becomes highly appropriate and effective. Given the homomorphism of machine intelligence, users in HCI scenarios can treat the communication counterpart as an autonomous individual rather than a tool. For example, in the case of autonomous driving, a significant portion of users’ distrust of AI stems from the fact that when an accident occurs, users have no way of knowing the cause of the error. Even if the cause is identified, they cannot find someone to hold accountable—being angry at a "tool" has no practical significance. This situation contrasts sharply with interactions involving a human driver, as the notion of "a driver driving" aligns with common sense, and the driver’s qualifications are recognized by social evaluation systems (e.g., a driver’s license or assessments by a taxi company). Even in the event of an accident, the user has an interactive counterpart, and their need for understanding is met. From the perspective of the new question, users no longer regard autonomous driving systems as mere tools but as independent individuals replacing human drivers. This individual might belong to a new group or species and exhibit homomorphism with humans.

Evidently, human-centered XAI can become increasingly complex in specific scenarios. From the Argument from Various Disabilities, it can be inferred that the trust issues in human-machine interaction, as exemplified earlier, can be described as judgments made by individuals based on principles of scientific induction. These trust issues are often unproven, much like how scientists cannot definitively prove whether the error rate is higher for autonomous driving systems or human drivers. Nevertheless, a user might still say, "I trust human drivers more because I know I’m actually communicating with a driver." This aligns with the social cognition formed by individuals over extended periods of social experience. In affective scenarios, such trust issues can be described as the user’s ability to “judge whether the listener has the capacity to ‘understand’,” as this is evidently a prerequisite for individuals to feel “understood” in a relationship. To some extent, this capacity reflects consciousness. Similarly, when interacting with independent entities exhibiting homomorphism, it remains unprovable whether such entities possess this capacity, even if their internal structure is entirely different from the human brain. Nonetheless, children often talk to toys or name their dolls: "Mark is my friend." These homomorphic entities exist solely within the child’s subjective cognition, but they still represent a form of human-centered XAI. This phenomenon may be related to shifts in societal ideology, as not everyone would experience the same sense of connection. This paper does not attempt to provide a conclusion on this matter but simply offers it as a direction for further reflection.

5. Conclusion

This paper investigates the issue of explainability in artificial intelligence (AI) within the context of human-computer interaction (HCI) scenarios, analyzing the limitations of Turing’s original question, “Can machines think?” in light of modern AI advancements. Based on assumptions about the era of general artificial intelligence, this paper underscores the significance of human-centered explainable AI and highlights the central role of user experience in evaluating the success of HCI. The original question posed by Turing is replaced with a new question: “What are we actually communicating with?” To

DOI: 10.54254/2755-2721/97/20241408

achieve effective human-machine interaction, research on explainability must shift from a purely technical focus to a UX-oriented approach. This means concentrating on how users comprehend and accept the behavior of AI systems, rather than solely focusing on the internal mechanics of the system. Through this human-centered approach, people can better design and develop AI systems that meet users' needs and expectations. Moreover, in the face of the potential emergence of AI's autonomous consciousness and the issue of human replacement, new ethical frameworks must be developed to address users' anxieties and fears, ensuring that the development and deployment of AI systems align with societal norms and ethical standards. The advancement of AI is not merely a technical challenge; it encompasses significant social and philosophical considerations. As AI systems become increasingly complex and autonomous, it requires the public to reconsider the relationship between humans and machines and explore how to fully harness AI's potential while preserving human values and ethical principles.

In future research, it is recommended to further investigate how human-centered XAI principles can be applied to different HCI scenarios and how to design and evaluate AI systems that foster user trust and acceptance. Additionally, there is a need to explore how to balance technological advancement with social ethics in AI development, ensuring that AI technologies bring positive impacts to human society. To conclude with Turing's words, "We can only see a short distance ahead, but we can see plenty there that needs to be done."

References

[1] Xu, B. (2024). What is Meant by AGI? On the Definition of Artificial General Intelligence. arXiv preprint arXiv:2404.10731.

[2] Bostrom, N., & Yudkowsky, E. (2018). The ethics of artificial intelligence. In Artificial intelligence safety and security (pp. 57-69). Chapman and Hall/CRC.

[3] The Ethics of Understanding: Exploring Moral Implications of Explainable AI https://www.ijsr.net/archive/v13i6/SR24529122811.pdf

[4] Qingfeng.Y (2020) Rethinking AI ethical principles from the problem of artificial intelligence https://philosophy.fudan.edu.cn/08/2b/c24650a264235/page.htm

[5] Liao, Q. V., & Varshney, K. R. (2021). Human-centered explainable ai (xai): From algorithms to user experiences. arXiv preprint arXiv:2110.10790.

[6] Turing, A. M. (2009). Computing machinery and intelligence (pp. 23-65). Springer Netherlands.

[7] Guo Quanzhong, Zhang Jinyi. AI+Humanities: the development and trend of AIGC[J]. Journalism Enthusiast, 2023(3):8-14.

[8] Russell, S. J., & Norvig, P. (2016). Artificial intelligence: a modern approach. Pearson.

[9] Norman, D. (2013). The design of everyday things: Revised and expanded edition. Basic books.

[10] Bessen, J. (2018). AI and jobs: The role of demand (No. w24235). National Bureau of Economic Research.

[11] What Is Perception in Machine Learning? https://labelyourdata.com/articles/machine-perception-in-artificial-intelligence

[12] Mini, U., Grietzer, P., Sharma, M., Meek, A., MacDiarmid, M., & Turner, A. M. (2023). Understanding and Controlling a Maze-Solving Policy Network. arXiv preprint arXiv:2310.08043.

[13] Wang, Y., Song, W., Tao, W., Liotta, A., Yang, D., Li, X., ... & Zhang, W. (2022). A systematic review on affective computing: Emotion models, databases, and recent advances. Information Fusion, 83, 19-52.

[14] Timmerman, G. M. (1991). A concept analysis of intimacy. Issues in mental health nursing, 12(1), 19-30.