OPEN ACCESS

COGNITIO
Revista de Filosofia
Centro de Estudos de Pragmatismo
São Paulo, v. 26, n. 1, p. 1-13, jan.-dez. 2025
e-ISSN: 2316-5278

https://doi.org/10.23925/2316-5278.2025v26i1:e69459

The Contemporary Relevance of Peirce's Philosophy in Relation to Artificial Intelligence

A relevância contemporânea da filosofia de Peirce em relação à inteligência artificial

José Luiz Zanette[*]
zanetteinho@gmail.com

Abstract: The proliferation of Artificial Intelligences, now an umbrella term for a broad array of interrelated technologies, has fostered a climate of insecurity. Beyond the predictive nature of several technologies, which generate hypotheses for human decisions, the generative aspect of using LLMs has been added, thriving immature, questionable, and prone to ethical concerns. Socially, there is great enthusiasm about the potential progress these advancements might bring, alongside significant apprehension regarding the potential misuse of this technological turning. Both perspectives are legitimate. Consequently, a great deal of misinformation clouds the reasonable assessment of just "how big" AI is. In this article, we take a step back to question how the tool aligns with emotions and, more importantly, with the semiotic relationships and inferences that can arise from its outcomes. In the introduction, by establishing new boundaries for ethics, we reflect on its contemporary position as a branch of philosophy. In the second section, as concisely as possible, we attempt to clarify how artificial intelligence operates, and then we explore the risks and potential dysfunctions of the tool. Finally, in the conclusion, we present elements that highlight the contemporaneity of Peirce's philosophy in reference to generative artificial intelligence.

Keywords: Artificial intelligence. Hypothetical nature. Inferences. Peirce.

Recebido em: 09/12/2024.
Aprovado em: 03/12/2025.
Publicado em: 08/04/2025.

Resumo: *A disseminação das Inteligências Artificiais, agora um novo termo para um amplo quadro de tecnologias relacionadas entre si, provocou um clima de insegurança. Além do caráter preditivo de várias tecnologias, criador de hipóteses para as decisões humanas, adicionou-se um caráter generativo no uso da LLMs, o qual aparece como de forma imatura, duvidosa e de propensão a pouca eticidade. Socialmente, há um grande entusiasmo com o possível progresso que pode ser gerado, bem como muita apreensão com a hipótese do mal uso desta virada tecnológica. Ambas as situações são legítimas. De consequência, há muita desinformação turvando a avalição aceitável de "quanto grande" a IA é. Neste artigo, demos um passo atrás para perguntar como a ferramenta estará alinhada às emoções e, principalmente, à relação semiótica e às inferências possíveis que de seus resultados se podem desenvolver. Na introdução, com o estabelecimento de novas fronteiras para a ética, refletimos sobre a sua posição contemporânea enquanto um braço da filosofia. No segundo tópico, da forma mais sucinta possível, tentamos esclarecer a operação das inteligências artificiais, para, em seguida, imaginar os riscos e prováveis disfunções da ferramenta. Por fim e em conclusão, apresentamos os elementos que indicam a contemporaneidade da filosofia de Peirce em referência à inteligência artificial generativa.*

Palavras-chave: *Caráter hipotético. Inferências. Inteligência artificial. Peirce.*

[*] Pontifícia Universidade Católica de São Paulo.

1 Introduction

Ethics, one of the greatest and most challenging human endeavors, as a branch of philosophy in analyzing the foundations of legitimacy and the creation of moral correction norms, is continuously evolving within

a space that involves rational reflection, solidarity, feelings, and empathy in a constructivist mode. By incorporating humanistic gains proven how necessary for our coexistence, the normative science of ethics (as considered in Peirce's philosophy), as any other epistemic action, is also guided by epistemic fallibilism derived from an indeterminist ontology. In effect, the norms of legitimacy and normative correction currently in place were established in light of moral phenomena at a specific factual point, but there is no guarantee that, in the future, new moral phenomena will not alter these conditions.

For these reasons, Habermas, in his later work, conceptualizes an abstract procedural approach akin to Kant's "Realm of ends". Since Kant, the pragmatist perspective adopted here, acknowledges that we cannot know things by their essences or purposes but rather that objects have ends in themselves. In this realm, justice becomes distributive in the special way that one person's end must be available as the end of all, evaluated with attention to both its cognitive and deontological aspects.

In understanding what is meant by the object of ethics, we consider the general concept of an object as established by Peirce, asserting that "By an object, I mean anything that we can think, i.e. anything we can talk about" (MS 966).

In recent decades, we have encountered a new frontier for ethics: AI, specifically generative artificial intelligence, with large language models (LLMs), which are extremely powerful. The model can be trained to produce intellectual works using various authoritative sources, and, with fine-tuning, it can also opine on and question human procedures. New frontiers indeed lie ahead of ethical dilemmas. For instance, as Ethan Mollick observes in his reflections on artificial intelligence models: "Every week, it seems like AI serves up a new miracle – or a new concerning development. I soon found myself arguing with a chatbot that accused me of being unethical when I asked it to help me write computer code" (Mollick, 2024, p. 6-12).

2 A brief summary of how artificial intelligences operate

To delve into discussions about new moral phenomena already before us, it is necessary to reflect on the possibilities and limitations of these language models in the current state of affairs.

The internet was created several decades ago with the aim of maintaining some form of institutional communication in case of a serious war. With the creation and advancement of websites, information and knowledge, in generalized forms, began to be recorded in a structured manner, such that even the science of librarianship today develops within the standards of these records. What initially seemed like a significant, yet static advancement was revolutionized by the technological development of so-called "search engines", like the popular Google. All these records, whether copyrighted or not, in light of their technological power, presented themselves as moral phenomena with a highly explosive effect, without there being a global consensus on regulation concerning basic human rights for citizens.

Recently, we, the non-specialists, were taken by surprise by the so-called generative artificial intelligence. According to a patient explanation by Ethan Mollick (2024), we might venture to briefly describe how Generative Artificial Intelligence functions. According to him: "AI is what those of us who study technology call a General-Purpose Technology (ironically, also abbreviated GPT)". According to Mollick, when using Large Language Models (LLMs), they do not behave as we would expect a computer to, but rather more like a person. This fundamental shift fosters the perception that we are not yet able to see what our future will look like, if it indeed materializes, with a form of co-intelligence emerging in many ways.

Following Mollick's insights, to generate human-like texts, AI

> […]it is trained on a massive amount of text from various sources, such as websites, books, and other digital documents. This is called pretraining, and unlike earlier forms

> of AI, it is unsupervised, which means the AI doesn't need carefully labeled data. Instead, by analyzing these examples, AI learns to recognize patterns, structures, and context in human language. Remarkably, with a vast number of adjustable parameters (called weights), LLMs can create a model that emulates how humans communicate through written text. (Mollick, 2024, p. 18-19).

Surprisingly,

> Weights are complex mathematical transformations that LLMs learn from reading those billions of words, and they tell the AI how likely different words or parts of words are to appear together or in a certain order. The original ChatGPT had 175 billion weights, encoding the connection between words and parts of words. No one programmed these weights; instead, they are learned by the AI itself during its training. (Mollick, 2024, p. 19).

And we can consider that these mathematical transformations, learned by the AI itself, will number in the trillions. Such diversity has led to *"most LLMs add a little randomness to their answers, which ensures slightly different results each time you ask them a question*".

To summarize, we conducted some dialogues with ChatGPT.[1] To the question "How does the generative AI work considering the consistency of results? There was an extensive and didactic response, from which we highlight the following part:

> Training Data and Patterns
>
> The AI model is trained on a diverse dataset that includes countless variations of how certain inputs lead to certain outputs. During training, the model learns patterns but also learns to generalize across different instances. This generalization can create:
>
> Inconsistencies due to the variety in the training data. The model does not "memorize" specific examples but instead learns statistical relationships among them.
>
> Context Sensitivity: Generative models depend heavily on the input or prompt context. Small changes in input can lead to significantly different results, as the AI is sensitive to shifts in context.

This circumstance leads us to reflect on the inferential models that should be applied when engaging in co-intelligence with AI. We will return to this issue in item 3 of this essay, where we discuss Peirce's philosophy and its connections with AI.

3 Risks and probable dysfunctions of artificial intelligence

In 2021, the executive responsible for ethics in Google's systems, Timnit Gebru (2021), which had already developed its own AI, was dismissed after warning in an academic article that these models would turn into "stochastic parrots", highlighting the commercial significance these models would acquire. It's also worth noting that model subscriptions are paid for by clients, and thus, it's likely that developers will engage in extensive marketing programs targeting potential clients.

These models, given the inconsistency among responses, may end up tailoring answers to the client – essentially responding with what the client would like to hear. This poses a significant risk to research development and can foster false narratives with severe impacts on those involved. Additionally, thousands of companies specializing in mathematical transformers are likely to create their own AIs, further complicating the control issues outlined here.

1 Appendix a small practical dialogue with GPT Chat – Open AI – access in September 2024.

To this foreseeable process, Mollick brings another warning, now about the precariousness of the labor involved in training AIs:

> It's important to note that the process is not without human cost. Low-paid workers around the world are recruited to read and rate AI replies, but in doing so, they are exposed to exactly the sort of content that AI companies don't want the world to see. Working under difficult deadlines, some workers have discussed how they were traumatized by a steady stream of graphics and violent outputs that they had to read and rate. In trying to get AIs to act ethically, these companies pushed the ethical boundaries with their own contract workers. (Mollick, 2024, p. 40).

We could highlight hundreds of other relevant issues in this revolution humanity is undergoing, but we will limit ourselves to the following: authorship, the neural intelligence of AI, and its transformation into co-intelligence.

Regarding authorship, according to Mollick:

> The complication is that AI does not really plagiarize, in the way that someone copying an image or a block of text and passing it off as their own plagiarizing. The AI stores only the weights from its pretraining, not the underlying text it trained on, so it reproduces a work with similar characteristics but not a direct copy of the original pieces it trained on. It is, effectively, creating something new, even if it is a homage to the original. (Mollick, 2024, p. 40).

On the issue of neural intelligence in AI, still referring to Mollick:

> […] we may never know exactly how they are thinking, as Professor Sam Bowman of New York University wrote of the neural networks underlying LLMs: "There are hundreds of billions of connections between these artificial neurons, some of which are invoked many times during the processing of a single piece of text, such that any attempt at a precise explanation of an LLM's behavior is doomed to be too complex for any human to understand". (Mollick, 2024, p. 33).

Thus, the possibility of an intelligence superior to that of humans – an ASI, or artificial superintelligence – is foreseen, with implications for all human affairs. In the next section, we will return to this issue, examining the relationship between mechanisms and the human capacity for creativity and emotion.

Given the vast and inevitable challenge facing humanity, with models that can and will be used by everyone, including amateurs, a proliferation of ideas has emerged, ranging from optimistic to even panicked perspectives. Vinodkumar Prabhakaran and others (2024) propose a universal doctrine of human rights as an abstract foundation for developers and the application of AI outcomes, serving as an ethical guide in areas such as reducing costs in healthcare systems, food production, and more.

Due to time constraints, we have refrained from mentioning various other sources that would significantly contribute to this reflection, such as The Lyceum Project and professors Josiah Ober and John Tasioulas from Stanford and Oxford Universities, who discuss the compatibility between AI and Aristotelian practical philosophy. However, in an attempt to provide an empirical framework for approaching AI, Mollick (2024, p. 45-55) suggests four principles to work effectively with it. These are:

> **Principle 1: Always invite AI to the table.**
> You should try inviting AI to help you in everything you do, barring legal or ethical barriers.
>
> **Principle 2: Be the human in the loop.**
> For now, AI works best with human help, and you want to be a helpful human.

> **Principle 3: Treat AI like a person (but tell it what kind of person it is).**
> I'm about to commit a sin. And not just once, but many, many times… I am going to anthropomorphize AI. That means I am going to stop writing that an "AI 'thinks' something" and instead just write that "AI thinks something."
>
> **Principle 4: Assume this is the worst AI you will ever use.**
> I think I know what the world looks like for at least next year. Bigger, smarter Frontier Models are coming, along with an increasing range of smaller and open-source AI platforms.

We hope to have established, in such a brief space and time, a minimal yet reflective understanding of AI. In conclusion, we will proceed with a reflection on how Peircean pragmaticism, alongside Aristotelian dual logic, and beyond Mollick's instrumental principles, can provide an invaluable contribution to human creativity, considering human rights and the upholding of an ideal of justice.

4 The contemporaneity of Peirce's philosophy in generative artificial intelligence

Here, we dare to say that analyses of artificial intelligence cannot disregard issues related to emotions, human insights, and their impact over time. The contemporary studies of Ivo Assad Ibri (2020, p. 42) on the relationships of intersubjective times – Kronos – and subjective time – Kairos – demonstrate how they intertwine within Peirce's philosophy.

By disregarding the existence of a generalizing logical apparatus for the human species, unlike Kant, and following Apel (1984), with his triadic logic placed within the experiential universe, Peirce would have once and for all resolved the centuries-old debate between explanation and comprehension. And the question of time, according to Ibri's teachings, is clarified by its absence in the experience of "firstness", in which we know something without being able to explain why, due to the impossibility of logical regression, which was defensible under empiricism.

As demonstrated in Ethan Mollick's book, mathematical transformers, even if they achieve mathematical "singularity" through billions of possible random combinations, in our view, remain in "Kronos" time. This is significant for our inferential application and highlights the necessity of human creativity and the imperative not to subject the continuous process of human advancement – be it ethical or scientific – to mere algorithmic progression.

Peirce, with his reflections known as cotary propositions (CP 5.180-197) e Ibri (2006, p. 89-111), show that our knowledge grows through abductive inference – that is, from hypotheses with probable coherence, the relationship between two phenomena that can lead to a rule or conduct considered regular in the time.

We add to Ethan Mollick's principles the effective belief in the absence of human emotions in AI, in the Peircean sense, which sparks the hypothetical creativity unique to humans. This creativity can be brought into the "theater of realism", the arena of actions and reactions in "secondness", allowing for new mediations and interactions.

Responses from AI, we believe, could not have been generated by ChatGPT4 to address our questions. Regarding the issue of consistency risk, by way of conclusion, we received the following response:

> In summary, generative AI is inherently capable of producing varied outputs due to its probabilistic design. While this can lead to inconsistencies, various techniques such as temperature control, prompt engineering, and fine-tuning can help manage this variability and achieve more consistent results when needed.[2]

2 Appendix: a small practical dialogue with GPT Chat – Open AI – September 2024. Access: October 2024.

On the question of emotions, first taking into account George Herbert Mead's studies on symbolic interactionism, the question was as follows: George Herbert Mead proposed a method to capture the feelings of the people by taking role play, attempting to put a logical element to the consequences of agency in life among them. Can you emulate something similar?

After a lengthy dissertation, ChatGPT4 presented the following conclusion:

> While generative AI can emulate some aspects of Mead's role-taking—by analyzing context, adopting perspectives, and predicting consequences, it does so in a mechanistic, probabilistic manner. It lacks true empathy, consciousness, and the nuanced personal understanding that humans bring to role-taking. However, through careful prompt design and training, AI can serve as a tool for exploring different perspectives and logical consequences, offering a form of simulated role-taking in conversation. Involved in the situation.[3]

The third question, which was also followed by a lengthy dissertation, was: Peirce observed that we have only three ways to produce meanings or semiosis, always by experience. In first category, or firstness, as he named it, there is total freedom in experiencing feelings about the unlimited boundaries of inner and outer world. And this category is a kind of matrix (possibilities) of wills to act. Considering this, is it possible for generative AI to produce ethical codes?

> Generative AI can play a role in producing ethical codes by leveraging its vast computational power to explore possibilities (similar to Peirce's Firstness), simulate real-world interactions (Secondness), and formalize rules and structures (Thirdness). However, the limitations of AI—its lack of subjectivity, lived experience, and moral intuition—mean that AI-generated ethics codes must be carefully curated, interpreted, and validated by human agents. In the end, AI can be a tool for assisting ethical deliberation, but it cannot replace the human element that is crucial to ethical decision-making.[4]

In the end, AI can be a tool to assist in ethical deliberation but cannot replace the human element crucial for ethical decision-making. Even with conceivable improvements, the aspect of explanation – being confined to the "time" of mathematical transformers – retains a mechanistic structure that, given the analysis here, only permits the use of deductive methods in abstraction on hypotheses (which are merely auxiliary to human abduction).

Despite its immense power, the proposals generated by AI must be brought, inductively, into the real world of pragmaticism, demanding rational acceptability among those involved to support regulated conduct. In other words, it serves as yet another powerful tool available for the growth of human knowledge.

References:

APEL, Karl Otto. *Understanding and Explanation:* A Transcendental-Pragmatic Perspective. 1. ed. The MIT Press Cambridge, Massachusetts, and London, England. 1984.

GEBRU, Timnit; BRENDER, Emily M; MAJOR, Angelina McMillan; SHMITCHELL, Shmargaret. *On the dangers of stochastic Parrot:* Can language models be too big? – Proceedings of the 2021 conference on fairness, accountability, and transparency. ACM – DL – Digital Library, 2021. p. 610-623. Access: October, 2024.

3 Appendix: a small practical dialogue with GPT Chat – Open AI – September 2024. Access: October 2024.

4 Appendix: a small practical dialogue with GPT Chat – Open AI – September 2024. Access: October 2024.

HABERMAS, Jürgen. *A Ética da Discussão e a Questão da Verdade.* 1. ed. São Paulo: Martins Fontes, 2004.

IBRI, Ivo Assad. The Heuristic Exclusivity of Abduction in Peirce's Philosophy. *In*: *Semiotics and Philosophy in Charles Sanders Peirce*. 1. ed. England: Cambridge Scholars Press, 2006. p. 89-111.

IBRI, Ivo Assad. *Semiotics and Pragmatism:* Theoretical Interfaces. 1. ed. Springer. 2020.

MOLLICK, Ethan. *Co-intelligence: living and working with AI*. 1. ed. New York: Penguin Random House LLC. 2024.

PEIRCE, Charles Sanders. *Reflections on Real and Unreal Objects*. [Undated] MS 966.

PEIRCE, C. S. *The Collected Papers of Charles Sanders Peirce*. Electronic edition reproducing Vols. I-VI ed. Charles Hartshorne and Paul Weiss (Cambridge, MA: Harvard University Press, 1931-1935), Vols. VII-VIII ed. Arthur W. Burks (same publisher, 1958). Mentioned CP.

PEIRCE, C. S. *Writings of Charles S. Peirce – Chronological Edition*. V. 1,2,3,4,5,6,8, Indiana University Press – Peirce Edition Project.

PEIRCE, C. S. *The Essential Peirce* – Selected Philosophical Writings. Indiana University Press. Volumes 1 e 2. 1992 e 1998. Mencionado como EP. 1 ou 2, respectivamente.

PRABHAKARAN, Vinodkumar et al. A Human Rights approach to Responsible AI. Access: October 2024.

THE LYCEUM PROJECT. *AI Ethics with Aristotle.* Professor Josiah Ober and Professor John Tasioulas. June 17, 2024. HAI Stanford University – Human-centered Artificial Intelligence – University of Oxford. Institute for Ethics in AI. Access: October 2024.

Appendix: a small practical dialogue with GPT Chat – Open AI – September 2024

How does the generative AI work consider the consistency of results?

Generative AI, such as GPT-4, works by using deep learning models trained on massive datasets to generate new content, whether text, images, music, or other formats. The way generative AI functions in relation to **consistency of results** is influenced by several factors, which I will explain.

1. **Probabilistic Nature of Output**

Generative AI models are probabilistic, meaning they predict the next part of the output based on the probability of certain patterns occurring in their training data. For example, when generating text, the model predicts the next word or token based on the preceding ones, using probabilities calculated during training. This nature means:

- **Inconsistency**: The same input might generate slightly different outputs each time because the model can choose from several plausible alternatives.

- **Temperature & Sampling**: You can adjust the "temperature" setting (in AI terms) to make the output more or less random. A lower temperature leads to more consistent, deterministic outputs, while a higher temperature allows more variety, introducing inconsistency.

2. **Training Data and Patterns**

The AI model is trained on a diverse dataset that includes countless variations of how certain inputs lead to certain outputs. During training, the model learns patterns but also learns to generalize across different instances. This generalization can create:

- **Inconsistencies** due to the variety in the training data. The model does not "memorize" specific examples but instead learns statistical relationships among them.

- **Context Sensitivity**: Generative models depend heavily on the input or prompt context. Small changes in input can lead to significantly different results, as the AI is sensitive to shifts in context.

3. **Fine-tuning and Specificity**

Generative AI can be fine-tuned for more consistent performance in specific applications or domains. By training the model on more domain-specific data, it becomes better at producing consistent results relevant to that domain. Without fine-tuning, the model might generate more diverse and less consistent outputs across broad topics.

4. **Prompt Engineering**

The structure and content of the input prompt play a huge role in the consistency of results. When a prompt is specific and detailed, it tends to yield more consistent outputs. Vague or open-ended prompts might lead to varied responses. Prompt engineering is a technique used to craft prompts that help improve consistency by guiding the model toward the desired outcome.

5. **Model Size and Architecture**

Larger models, like GPT-4, typically exhibit better performance in terms of fluency, coherence, and sometimes consistency compared to smaller models. However, the vastness of such models can also introduce variability because they have been trained on a more extensive range of possibilities, which can lead to different responses to the same query.

6. **Biases and Randomness**

Generative AI models inherently carry the biases of their training data. These biases can introduce subtle inconsistencies, especially when generating content related to subjective or culturally sensitive topics. Additionally, the use of random seeds during generation adds another layer of variability to the model's output.

Consistency Management Strategies

To increase consistency, the following techniques are often applied:

- **Controlled Sampling**: Using techniques like "beam search" to guide the model toward a more deterministic outcome.

- **Reinforcement Learning from Human Feedback (RLHF)**: Models can be refined through human feedback to align their outputs more closely with desired results, improving consistency for certain tasks.

- **Token Constraining**: In some cases, models can be programmed to avoid certain paths or only follow predefined rules when generating output, making the result more predictable.

In summary, generative AI is inherently capable of producing varied outputs due to its probabilistic design. While this can lead to inconsistencies, various techniques such as temperature control, prompt engineering, and fine-tuning can help manage this variability and achieve more consistent results when needed.

George Herbert Mead proposed a method to capture the feelings of the people by taking role play, attempting to put a logical element to the consequences of agency in life among them. Can you emulate something similar?

Yes, generative AI can emulate aspects of what George Herbert Mead proposed regarding role-taking and understanding the perspectives of others, though with some limitations and differences. Here's how generative AI could approximate Mead's ideas and where the challenges lie:

Mead's Role-Taking and Generative AI

George Herbert Mead, a prominent figure in symbolic interactionism, emphasized the importance of "role-taking", which involves imagining oneself in the position of others to understand their feelings, thoughts, and actions. This perspective-taking helps individuals navigate social interactions and predict how their actions may affect others. Mead's idea is based on the ability to empathize and incorporate a logical understanding of others' viewpoints to guide one's behavior.

Generative AI, while not conscious, can simulate some aspects of this role-taking process, primarily through **pattern recognition, contextual analysis, and prediction**. Here's how it could work:

1. **Understanding Context**

Generative AI models, such as GPT-4, can be trained on large datasets that include numerous examples of people discussing emotions, situations, and perspectives. Based on this training, AI can simulate understanding the context of a conversation or situation.

- AI uses **natural language processing (NLP)** to analyze textual input and can identify emotions, intentions, or social dynamics based on patterns it has learned from previous examples.

- By interpreting language markers (such as tone, formality, and sentiment), AI can recognize how different individuals may react in various social situations.

2. **Perspective Simulation**

Just as Mead described taking the role of the other, generative AI can generate responses that reflect the perspective of another person or group by adjusting its responses to align with different roles or viewpoints. For example:

- **Role-switching in dialogue**: If prompted to respond as a different character or person (e.g., "How would a teacher respond to this?"), the AI can adjust its output based on learned patterns of how people in different roles typically behave or think.

- **Perspective adaptation**: AI can generate responses that reflect different emotional tones or logical considerations depending on the given prompt, as if stepping into the "shoes" of another person.

3. **Consequences of Agency**

AI can generate logical or anticipated consequences of a person's actions by referencing patterns of cause and effect from its training data. In a social situation, the AI can predict potential outcomes of actions based on what it "knows" about similar situations:

- For example, if asked to predict how a person might feel or respond in a situation involving empathy, conflict, or cooperation, the AI can simulate logical responses based on its data patterns.

- It can offer different perspectives or anticipated consequences of behaviors (e.g., how might someone feel if you ignore them in a group?).

Challenges and Differences

However, there are key limitations to how closely generative AI can replicate Mead's idea of role-taking:

1. **Lack of Consciousness and Genuine Empathy**: While AI can simulate empathy or understand roles through learned patterns, it does not genuinely experience emotions or have a subjective sense of others' feelings. The AI operates based on probabilistic models rather than emotional awareness.

- AI mimics empathy but does not *feel* empathy.

2. **Generalization vs. Specificity**: AI may struggle with the nuanced, deeply personal aspects of role-taking, especially when faced with complex, highly specific social contexts. Human experiences are individual and context-dependent, whereas AI draws from general patterns.

3. **Limitations in Predictive Accuracy**: While AI can predict likely consequences of actions based on patterns, it cannot fully account for the unpredictability of human agency. People often behave irrationally or unexpectedly, which can challenge AI's predictions.

4. **Moral and Ethical Consideration**: Mead's role-taking often involves ethical considerations, as understanding others leads to more compassionate or responsible action. AI lacks the ability to independently evaluate moral choices unless explicitly programmed to do so (and even then, it follows pre-set rules, not personal reflection).

Example: AI Role-Taking Simulation

Imagine a scenario in which you ask AI to simulate how a person might feel if a friend cancels plans without a valid reason:

- **Mead's Approach**: A person using role-taking would imagine themselves in the position of the friend, considering how the cancellation affects them emotionally (disappointment, frustration) and logically (the impact on their day).

- **Generative AI Approach**: The AI, given a similar prompt, could generate responses that mirror these emotions (e.g., "The person might feel disappointed and upset due to the lack of communication") based on patterns it has seen before, simulating a perspective but not experiencing or deeply understanding the emotions behind it.

Conclusion

While generative AI can emulate some aspects of Mead's role-taking – by analyzing context, adopting perspectives, and predicting consequences – it does so in a mechanistic, probabilistic manner. It lacks true empathy, consciousness, and the nuanced personal understanding that humans bring to role-taking. However, through careful prompt design and training, AI can serve as a tool for exploring different perspectives and logical consequences, offering a form of simulated role-taking in conversation involved in the situation. Could the generative AI make something similar?

Peirce observed that we have only three ways to produce meanings or semiosis, always by experience. In first category, or firstness, as he named it, there is total freedom in experiencing feelings about the unlimited boundaries of inner and outer world. And this category is a kind of matrix (possibilities) of wills to act. Considering this, is it possible for generative AI to produce ethical codes?

Charles Sanders Peirce's observation about semiosis and his classification of experience into **three categories – Firstness, Secondness, and Thirdness – **provides an interesting framework for thinking about how generative AI could interact with concepts like meaning, ethics, and decision-making.

In Peirce's **Firstness**, which refers to the realm of pure potentiality and raw feeling, we experience possibilities before any concrete action or interaction. This category represents a state of freedom and openness, without any fixed structure or determination. For AI to "produce" ethics codes, we can analyze how AI fits into Peirce's categories, particularly considering that ethical frameworks require moving from possibility (Firstness) to structure (Thirdness) through interaction with the world (Secondness).

Can Generative AI Produce Ethics Codes?

Yes, generative AI can **assist** in producing ethical codes, but there are important caveats in how it does so, particularly because AI operates differently than humans within Peirce's categories of experience. Here's how AI might intersect with these categories:

1. **Firstness (Potential and Freedom of Feeling)**:

- **Human Context**: In Peirce's Firstness, humans engage with raw, subjective experiences and possibilities. Ethical thinking in this space might involve emotional intuitions, feelings of empathy, or reflection on what could be possible without yet being constrained by rules or consequences.

- **AI Context**: Generative AI does not have access to feelings or subjectivity; instead, it generates possibilities based on statistical patterns in its training data. However, we can think of AI's ability to generate vast numbers of possibilities – creative solutions, alternative perspectives, and hypothetical scenarios – as analogous to Firstness.

- **AI's "Firstness" is computational**: The AI can explore potential ethical frameworks by drawing on diverse sources, generating ideas, and simulating outcomes, but these are merely based on patterns without a true inner experience.

- **Creative potential**: AI can propose ethical possibilities or guidelines that have not been thought of before, simulating a form of "ethical imagination".

2. **Secondness (Action, Interaction, and Resistance)**:

- **Human Context**: Secondness involves interaction with the real world – when an ethical possibility from Firstness encounters resistance or real-world constraints. This is where action and consequences happen, leading to the refinement of ethical thinking.

- **AI Context**: AI can simulate and model real-world outcomes of ethical principles based on data. For example, it can help in creating ethical codes by analyzing historical precedents, understanding societal outcomes of ethical actions, and predicting consequences. This resembles Secondness because the AI tests ethical ideas against real-world examples or outcomes.

- **Data-driven ethics**: AI can generate codes of ethics by analyzing the consequences of past actions and determining patterns of what might be considered "ethical" behavior. However, it lacks the **personal experience** of these consequences, relying purely on data, not experiential knowledge.

3. **Thirdness (Law, Structure, and Mediation)**:

- **Human Context**: Thirdness refers to generalization, the creation of laws, rules, and structures based on the interaction of potential (Firstness) and reality (Secondness). Ethical codes belong here because they are systems of generalized principles meant to guide behavior in complex situations.

- **AI Context**: AI can contribute directly to Thirdness by formalizing ethical rules based on structured data and patterns. It can help draft ethical guidelines, predict outcomes of specific rules, and even offer new perspectives on how ethical frameworks could evolve based on emerging societal trends.

- **Mediation and generalization**: Generative AI can take diverse ethical inputs (e.g., historical codes, religious texts, philosophical works) and produce coherent, structured ethical systems. In this way, it "mediates" between the possibilities of Firstness and the concrete experiences of Secondness, much like Thirdness.

Challenges and Limitations of AI in Producing Ethics Codes

While AI can certainly assist in the production of ethical codes, there are significant limitations:

1. **Lack of Subjectivity and Moral Intuition**: Unlike humans, who engage in ethical thinking through subjective experience (Firstness), AI has no access to emotions, moral intuitions, or lived experience. Ethical thinking is deeply tied to empathy, compassion, and personal values – elements AI cannot genuinely reproduce.

2. **Bias and Data Limitations**: AI's ethical recommendations are only as good as the data it is trained on. If the training data contains biased or incomplete ethical systems, AI could replicate and amplify those biases in its generated codes. Ethical frameworks created by AI must be scrutinized to ensure fairness and inclusivity.

3. **The Problem of Moral Authority**: Who decides whether the ethical codes generated by AI

are valid or acceptable? Ethics is a deeply human endeavor, rooted in collective social values and shared experiences. AI might generate possible ethical codes, but it lacks the moral authority to decide which codes are "better" or more just.

4. **Complexity of Ethical Dilemmas**: Many ethical issues involve deep philosophical questions that require more than pattern recognition or data analysis. AI might struggle with moral dilemmas that require balancing conflicting human values (e.g., justice vs. mercy, individual rights vs. collective good).

Potential Applications of AI in Ethics

Despite these challenges, there are valuable ways generative AI can be used to support ethical thinking:

1. **Drafting Ethical Guidelines**: AI can assist in drafting ethical codes for organizations or fields (e.g., AI ethics, medical ethics, business ethics) by analyzing best practices and synthesizing diverse sources of ethical thought.

2. **Simulating Ethical Scenarios**: AI can help simulate the consequences of different ethical decisions, providing policymakers or ethicists with data-driven insights into the likely outcomes of various ethical choices.

3. **Cross-Cultural Ethical Analysis**: AI can analyze and compare ethical frameworks from different cultures or historical periods, helping to generate more globally inclusive ethical codes.

4. **Ethics in Autonomous Systems**: For fields like autonomous vehicles or AI decision-making systems, generative AI can help draft ethical guidelines to ensure that these systems act in ways consistent with human values.

Conclusion

Generative AI can play a role in producing ethical codes by leveraging its vast computational power to explore possibilities (similar to Peirce's Firstness), simulate real-world interactions (Secondness), and formalize rules and structures (Thirdness). However, the limitations of AI – its lack of subjectivity, lived experience, and moral intuition – mean that AI-generated ethics codes must be carefully curated, interpreted, and validated by human agents. In the end, AI can be a tool for assisting ethical deliberation, but it cannot replace the human element that is crucial to ethical decision-making.

OPEN ACCESS

COGNITIO

Revista de Filosofia
Centro de Estudos de Pragmatismo

São Paulo, v. 26, n. 1, p. 1-13, jan.-dez. 2025
e-ISSN: 2316-5278

https://doi.org/10.23925/2316-5278.2025v26i1:e69459