Can Turing Machines Possess Intrinsic Intentionality?

Zhen Wang

This paper explores the question of whether Turing machines, particularly artificial intelligence (AI) systems, can exhibit intrinsic intentionality — defined as the capacity to interpret internal processes and generate meaningful outputs. This paper then discusses Searle's Chinese Room Argument (1980), which challenges the possibility of machines' intrinsic intentionality, as well as the syntactic theory that suggests otherwise. This theory suggests that internalized syntactic processes suffice for creating intrinsic intentionality. Rapaport used Helen Keller's experience to illustrate how the internalization of symbols may create intrinsic intentionality (2007) . Finally, this paper raises objections to syntactic semantics as a solution to Turing Machines to acquire intrinsic intentionality. It argues that AI symbols can only be about intrinsically meaningless tokens without phenomenon experience. Drawing on Jackson's Knowledge Argument (1982), the paper contends that intrinsic intentionality requires a mental process to be about a phenomenal experience.

1. Introduction

For humans, our mental activities have meaning. To say that all raccoons are mammals is not merely a logical proposition that all *A*s are *B*. For us, a raccoon means a bandit-looking furry creature with four limbs and various other characteristics. We can visualize one with our mind's eye and imagine how it moves or sounds. Computers are Turing machines that manipulate inputs based on sets of instructions. An artificial intelligence system may contain a class called mammal which has a subclass called raccoon in its storage. But does a raccoon mean anything

to this system when it processes a raccoon? I will first discuss Searle's Chinese Room Argument as a negative answer to this question. Then, I will present and evaluate the theory of syntactic semantics which argues that internalized syntactic processes are meaningful on their own. Finally, I will argue against the syntactic semantics theory by arguing that the grasp of meaning requires intrinsic intentionality, which requires phenomenon consciousness.

2. Two Types of Intentionalities

In keeping with influential works in Philosophy of Mind, I use the term intentionality to mean "the power of a process to be directed at or about certain things like objects, properties, and states of affairs."[67] There are two types of intentionality: original intentionality and derivative intentionality.[68] A book, for example, can refer to many objects or concepts through its texts. However, it only does so when a reader interprets it. So, the book only has derivative intentionality that affords its interpretability. Such intentionality was given by the author of the book and reconstructed by its readers. Original intentionality is the capability of delegating representations to objects and interpreting objects from representations. Therefore, original intentionality exists only in the interpreters of the book. For the purpose of this essay, I will refer to original intentionality as intrinsic intentionality. This is because the word "original" may carry a connotation of authorship in the legal sense. An interpreter of words in a book possesses original intentionality not because they are the first to delegate certain meaning to the words, but

[67] Searle, 1980; Haugeland, 1990; etc. Dietrich et al. 2021, p. 93
[68] Haugeland, 1990

because they are capable of delegating *any* meaning to them.

3. The Chinese Room Argument and Intentionality

The problem of machines and meaning is not about derivative intentionality. The outputs of machines like a calculator or a large language model (LLM) can usually sustain human interpretation. This is because their symbols can be translated into a human language, and their syntax can be defined to only allow interpretable outputs. If you take care of the syntax, the *derivative intentionality* will take care of itself. However, it is far from clear whether a machine can possess intrinsic intentionality — the power to interpret its internal processes and produce sensible output that is also meaningful to itself. This is an apparent feature of human cognitive systems. We can *interpret* what we think (our internal processes), what we say, and even much of what others say. A famous argument against the possibility of artificial intelligence (AI) having intrinsic intentionality is the Chinese Room Argument proposed by Searle (1980). He wondered whether the human mind works like a Turing machine, a purely formal system. He concludes that if we work like that, we would not be able to even interpret our own languages.

Suppose you are locked inside a room with an input slot and an output slot. The input you receive is written in a language completely strange to you. There is a handbook that outlines how you should respond upon encountering any kind of input. So, being a good rule follower, you produce correct responses and insert them into the output slot. To an outsider who understands the strange language, it is as if the room has a native speaker of that language.

Searle notes that no matter how good you are at manipulating the inputs to produce the outputs, the language means nothing to you. Searle notes that in the Strange Language Room, you behave just like a computer processor. The handbook is like a program written by intelligent programmers. While you do not understand that strange language, the book enables you to pretend to understand. Therefore, if an otherwise-intentional being like yourself cannot derive intentionality from formal syntactic operations, there is no reason to believe a computer processor can. What gives us intrinsic intentionality must not be formal syntactic manipulation.

For the machine to possess intrinsic intentionality, it needs to be able to interpret its own processes and figure out what they are *about*. Searle argues that human brains have "proper causal powers" to possess intrinsic intentionality. Searle does not argue that our intentionality must represent the outside world. Those proper causal powers refer to the physical-chemical processes and the biological structure of an organism's brain.[69] This implies that a brain-in-a-vat would possess intrinsic intentionality, whereas a silicon-based robot that can act entirely indistinguishable from humans never could. However, Searle makes no argument defending how biological processes, *but not* electronic processes, can give rise to intrinsic intentionality. If this claim is not taken for granted, then another compelling theory of intrinsic intentionality should be considered.

4. Syntactic Semantics

In response to Searle (1980), Dietrich et al. (2021) discuss the syntactic semantics theory of

[69] Searle 1980, p. 442.

intentionality. Proponents of syntactic semantics believe that a formal system is sufficient to generate intrinsic intentionality. Rapaport (2007) uses the life story of Helen Keller to argue that intentionality arises when all semantics are properly internalized. Helen Keller was both blind and deaf since childhood, yet she could learn to communicate using finger gestures, signs, and eventually English. Rapaport argues that Keller had been living in a version of Searle's Strange Language Room for almost her entire life. The following quote from Keller's autobiography suggests that she manipulated the English symbols based on syntactic rules: "I did not know that I was spelling a word or even that words existed; I was simply making my fingers go in monkey-like imitation."[70] As she mastered the syntax, it was obvious that she does understand English, and English *means* something to her. Dietrich et al. (2021) summarize that the key to syntactic semantics is the internalization of external symbols. Once they are appropriately internalized by the agent, the symbols are intrinsically intentional.

Her example seems to suggest that formal syntactic manipulation can be sufficient for intrinsic intentionality. Computers are good at syntactic manipulation, so perhaps they can possess intrinsic intentionality as well. Under Rapaport's syntactic semantics theory, symbols can be said to be *about* each other via a process called variable binding. This process lets a variable name refer to an entity. Variables and objects can be defined in terms of each other and constitute each other. The AI interprets a variable by following its references. For example,

[70] Keller, 1905, p. 35, cited in Rapaport, 2007, p. 395.

suppose an AI system with cameras detected a raccoon sleeping on the grass.[71] The object recognition algorithms determined that the object was a raccoon. So, the AI instantiated a Raccoon in its environmental model with the following *fields* (encapsulated information in object-oriented programming languages):

Raccoon #3942	
Class	Animal
Sub-class	Raccoon
Colour	R: 23 G: 21 B: 27
Distance	5
Ground Velocity	2

So, Raccoon #3942 referred to the combination of all its properties/fields. Variables such as “distance” and “ground velocity" referred to numbers five and two. As the robot approached, it startled the raccoon who increased its velocity away from the robot. So, the robot retrieved those variables and incremented them as such:

[71] cf. Dietrich et al., 2021, pp. 98-9.

Raccoon #3942	
Class	Animal
Sub-class	Raccoon
Colour	R: 23 G: 21 B: 27
Distance	Distance + 4
Ground Velocity	Ground Velocity + 5

The AI being able to follow references is a sign of interpretation according to the syntactic semantics theory. The raccoon can also exist in relation to other objects in the robot's environmental model. For instance, a new field in Raccoon #3942 called Surface can indicate the surface on which it stands. Surface can be bound to a grass chunk. The grass chunk can also have a field that is bound to Raccoon #3942. If the environment model is set up properly, the AI system can simulate interactions between objects and run counterfactual scenarios. This does seem to approach the power of human intentionality about other objects. Note that AI may behaviorally simulate intentional beings like humans. The physics simulation above can afford it to perform some goal directed actions. However, the question of whether variable binding captures all that is required for having intrinsic intentionality still remains to be open.

5. Meaningless Symbols Do Not Produce Meaning

On the table representations of Raccoon #3942, I intentionally (no pun intended) included a field called Color to raise suspicion about the syntactic semantics theory. The AI system represented colors using the intensity of primary colors: red, green and blue. The raccoon's color may fall under gray to a sighted human. But what is the Color field *about* to the AI system? It seems that it is really about a collection of three integers: R, G, and B. Then what does each of them mean? A knowledgeable computer scientist may program all we know about color science into the AI system. However, does that give it any idea about what red, green, or blue means? This scenario is analogous to Mary's (the color scientist) situation in Jackson's Knowledge Argument for qualia (1982). I believe if we *somehow* programmed the phenomenal experience of seeing colors into the AI system, it would learn something new. Without being able to experience any color, the AI's color field cannot be color.

An objection is that the syntactic AI system can experience colors via the camera connected. The experience involved the sensor registering lights of different frequencies, the processor writing data into the memory, and so on. So, there is no need to *somehow* program the phenomenal experience because the system could already experience colors. My response is that attributing phenomenal experience to camera sensors and processors risks anthropomorphizing mechanical processes. There are three premises for my response.

(1): Phenomenal experience requires levels of dynamical emergence.

(2): The light-sensitive material in a camera's sensor does not sustain the levels of emergence.

(3): AI systems designed for accomplishing computation do not sustain the levels of emergence.

The first premise is based on works of Terrence Deacon on biological anthropology and neuroscience. Deacon (2013) argues that phenomenal consciousness requires three levels of emergence from thermodynamic (homeodynamic) processes to morphodynamic, teleodynamic, and higher-order teleodynamic processes.

> ... this second-order teleodynamics is analogous to the way that the teleodynamics of interacting organisms within an ecosystem can contribute to higher-order population dynamics, including equilibrating (homeodynamic) and self-organizing (morphodynamic) population effects… the tendency for population-level morphodynamic processes to emerge in the recursive flow of signals within a vast extended network of interconnected neurons is critical to the generation of mental experience … This tangled hierarchy of causality is responsible for the special higher-order sentient properties (e.g., subjective experience) that brains are capable of producing, which their components (neurons) are not.[72]

He argues that sentience is the result of organisms (perhaps not exclusive to biological ones) engaging in self-creative and self-bounding tendencies. I argue that AIs that work like a Turing Machine function only at the thermodynamic level, and are neither self-creative nor self-bounding. The same can be said about the camera sensor. Therefore, I take (2) and (3) as true. If all my premises are true, it follows that:

[72] Deacon, 2013, p. 510.

(4): the phenomenal experience of color cannot come into existence by connecting a light sensor and processors without additional emergent processes.

If we "interrogate" the AI system for the meaning of a color, it can only respond with other ungrounded symbols. Note that Searle (1980) would not even consider those as symbols because they are not interpretable for machines (p. 422). For a symbolic AI (in contrast to artificial neural networks), I grant Rapaport (2007) that a symbol can refer to the symbol(s) that it was bound to. For artificial neural networks (ANNs), there are no longer distinct high-level symbols interpretable to humans. Their operations consist of layers of threshold logic units (TLUs) and store information in their connection weights.[73] They are trained with input data and using algorithms like error backpropagation to modify thresholds in TLUs to produce desirable outputs.[74] This means that they are Turing Machines that perform syntactic operations on their inputs. However, what do the syntactic operations mean? The compiler of a program translates executable high-level computer instructions into low-level instructions. Eventually, the codes are translated into machine-readable binary instructions. Binary instructions are actualized in the silicon as different voltages in wires and logical gates. Nowhere in these processes could a phenomenal experience seem to emerge.

If a symbol is not fundamentally about a phenomenal experience, what could it be about? My answer is meaningless tokens. For a person who has never experienced the color red, there

[73] Kruse et al., 2013, p. 15
[74] Ibid, 34, 66.

is no amount of mental gymnastics they can do to make ⟨*R*: 255, *G*: 0, *B*: 0⟩ about this color 🟥 (a red colored square). For a person who can see red, they can try to imagine a color that is outside of the human's visible spectrum. We can think about the light's (or electromagnetic radiation) physical or thermal properties because they can translate into our experience, but we can never think of that color.

I propose that a mental process is intrinsically intentional *if and only if* it is about a phenomenal experience. A problem with Rapaport's (2007) analogy that Helen Keller lived in a Strange Language Room is that she lived *the human experience.* She experienced emotions and sensations. Her concepts of water, cake, coldness, and textures of objects were all grounded in the sensations that they cause. This is vastly different from a purely symbol manipulator such as our AI friend above. All its symbols only refer to other symbols, whereas Keller's finger plays could refer to phenomenal experiences.

A corollary of this proposal is that not all human mental processes are intentional. Processes about purely syntactic constructs are only derivatively intentional. For example, when I *only* think of the number two, it does not refer to any phenomenon. It could refer to the successor of one or the predecessor of three in the domain of integers, but those references are only syntactical because both one and three are also mere syntactic constructs. In contrast, to think of two apples is about the phenomenon of them; a combination of their colors, smell, taste, texture, etc. Of course, I can think of "two apples" as an abstract symbol. This would make the

thought *only* derivatively intentional. The act of interpreting the symbol can ground it to something phenomenal and thus make it intrinsically intentional.

6. Conclusion

To build an AI that thinks like humans, intrinsic intentionality is an important feature that needs to be included. The human mind is intrinsically intentional because we can interpret what our own mental activities are about. Searle's Chinese Room Argument (1980) demonstrates that no amount of syntactic manipulation can give rise to intrinsic intentionality. He further argues that only biological brains are capable of generating intrinsic intentionality, but he does not give sufficient evidence for this claim. Therefore, it seems promising that the syntactic semantics theory could tackle the challenge posed by Searle (1980). Rapaport (2007) proposes that appropriately internalizing symbols into a system is sufficient to create intrinsic intentionality, regardless of human brains or Turing machines. He suggests that Helen Keller learned a human language via a similar process. I argue that Rapaport understated the importance of Keller's phenomenal experience as a human being. It was the human experience that provided something to ground her symbols onto. I propose that a process is intrinsically intentional if and only if it is about a phenomenal experience. I am not convinced that any Turing machine-based AI has phenomenal experience. Thus, they are not intrinsically intentional. However, I do not exclude the possibility of AI acquiring phenomenal experience someday. How AI might gain phenomenal experience is an important question for future research.

Works Cited

Deacon, T. W. (2013). “Incomplete Nature: How Mind Emerged from Matter.” W. W. Norton.

Dietrich, E., Fields, C., Sullins, J.P., van Heuveln, B., & Zebrowski, R. (2021). “The Strange Case of the Missing Meaning: Can Computers Think About Things?”. In Great Philosophical Objections to Artificial Intelligence: The History and Legacy of the AI Wars (pp. 87–134). London: Bloomsbury Academic. Retrieved November 30, 2023, from http://dx.doi.org/10.5040/9781474257084.ch-005

Haugeland, J. (199 0). “The Intentionality All-Stars.” *Philosophical Perspectives*, *4*, 383–427. https://doi.org/10.2307/2214199

Keller, H. (1905). *The story of my life*. Garden City, NY: Doubleday (1954).

Jackson, F. (1982). “Epiphenomenal Qualia.” The Philosophical Quarterly (1950-), 32(127), 127–136. https://doi.org/10.2307/2960077

Kruse, R., Borgelt, C., Klawonn, F., Moewes, C., Steinbrecher, M., & Held, P. (2013). “Threshold logic units.” In Texts in Computer Science (pp. 15–35). Springer London. http://dx.doi.org/10.1007/978-1-4471-5013-8_3

Rapaport, W. J. (2007). "How Helen Keller used syntactic semantics to escape from a Chinese Room." Minds and Machines, 16(4), 381–436. https://doi.org/10.1007/s11023-007-9054-6

Searle, J. (1980). "Minds, brains, and programs." Behavioral and Brain Sciences, 3(3), 417-424. doi:10.1017/S0140525X00005756