"A Sidekick With A Gambling Problem"

SPHEX Club Presentation

Michael A. Gillette, Ph.D. March 19, 2026

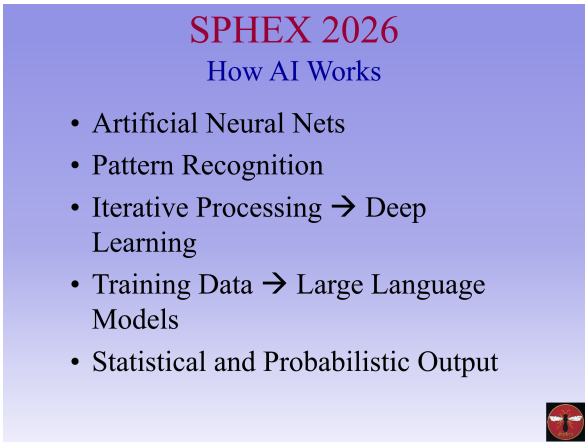
Part One- Introduction:

We have all been inundated with discussions of artificial intelligence. For instance, I have been asked in my professional capacity many times over the past two years to address the ethical implications of the utilization of AI, especially within the healthcare setting. It is not my intent this evening to regale you with yet one more discussion of AI specifically, but a brief and rudimentary review of what AI is and how it works will help in setting the stage for what I hope will be a more novel discussion.

Conceptually, the best way to understand AI is to consider an "artificial neural net". Although my language will be spatial, that serves only as an analogy to how AI machines are actually constructed. Consider a neuron to be a place that holds a piece of information. In the AI world, an artificial neuron is simply an electronic storage area that can contain what the field refers to as a "token". The token is more than a bit or a byte. It is a packet of information which could be a word, a number, even a phrase. In AI development, engineers create a first "row" of artificial neurons and populate them with a set of tokens of the developer's choosing. This initial set of tokens are considered the training data for the AI model. The developer also creates an initial algorithm that instructs the AI to detect patterns in the training data. So far, this is not interesting. Decades ago when I was in college I learned to program in Basic and easily learned how to populate and manipulate data in multi-dimensional arrays. But what makes AI different is what happens next. The AI feeds the conclusions that it reaches into a second "row" of artificial neurons and continues the process of pattern recognition, this time not only within the tokens inhabiting the second row, but also across the now doubled-in-size array. This process continues iteratively based on the total number of artificial neurons available. The only limit is the supply of chips and the massive amounts of electricity that are required to continue the process.

Imagine a third row, a fourth row, a fifth row, etc… of artificial neurons and the AI generating an "understanding" of the relationship between all of the tokens in that array – some of which were fed into it from outside sources and many of which were populated by the AI itself as it discovered ever deeper and more interrelated patterns.

This iterative process is what we refer to as "deep learning". When the AI is engaged in ANI (artificial narrow intelligence), it is trained on a small training data set and can respond to very few specifically defined prompts. A good example of how ANI currently is used to great effect is in reading radiologic images. The AI is trained on known radiologic images and can only respond to queries like "is this chest X-ray positive for lung cancer". AGI (artificial general intelligence) is trained on a very broad set of initial tokens, like everything that has ever been on the internet, and it can respond to a wide variety of prompts that are provided in normal language. This is what we refer to as large language models.



The take away from this cursory overview of how AI works is that although the models with which we interact are extremely powerful and able to engage in discussions across almost unlimited content areas, the AI is still only engaged in statistical analysis. The AI is only able to produce probabilistic output. Given this string of tokens, AI "reasons" that the next most likely token is "X", and that is what it produces. The AI does not know what it is outputting, it does not understand what it is outputting, and it does not care about what it is outputting. It is a computer, and although computers can calculate, they cannot think. Or can they?

Part Two- Thinking:

In the work that I have done over the past couple of years I have been fond of arguing that while AI has a tremendous amount of "A", it doesn't actually have any true "I". It seems to me that true intelligence involves several characteristics that AI models lack. First, the AI has no intention. It does not desire anything and it does not do things because it "wants to". The AI is a computer

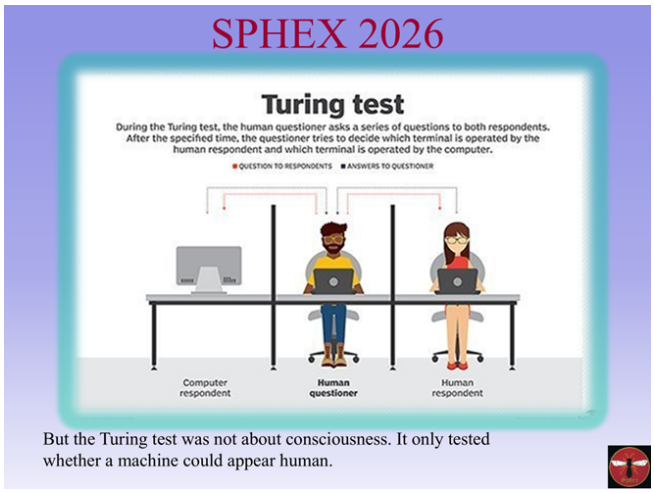
program. It is clearly a complex program – in fact it is so complex that we cannot know how it does what it does after the first pass of training data is placed in the first row of artificial neurons. Progress has been made in building in algorithms that allow the AI to report on its internal "reasoning", but we are unable, as humans, to reverse engineer an understanding of how the AI derives the conclusions that it does. Nevertheless, there is little reason to think that the AI actually cares about what it is doing.

Second, I have hitherto maintained that AI, by its very nature, cannot be truly creative. While the AI can hallucinate and generate counterfactual output, that output is still strictly probabilistic. This is not true, deep, creativity. When Einstein developed the theory of relativity, that was not just the next statistically likely step in the development of physics. It was a reconceptualization of the very ideas underlying physics, and it could not have been derived from a process of simple progression from pre-existing theories. It seems that AI does not have the ability to be creative in this way. It cannot take conceptual leaps; it takes only very rapid and extraordinarily numerous baby steps.

Third, and most important for tonight's discussion, AI engines do not seem to be conscious. AI models can interact with their environments, but are they aware of their environments? Do they contemplate their environments? Do they contemplate themselves- are they self-aware and thus self-conscious? This is where the discussion departs from previous SPHEX papers and wanders into a new area of inquiry for this author.

Part Three- Thought:

I am sure that both through your own research and after listening to the January 11, 2024 SPHEX paper delivered by Stephen Smith, many of you are familiar with Alan Turing's "Imitation Game" as he outlined it in his 1950 paper entitled "Computing Machinery and Intelligence". In that paper, Turing argued that it would soon be possible for machines to think, and that the test



for such an ability would be their success in fooling an interlocutor in a single blind interaction into believing that she or he were speaking to a human being. (Turing 1950)

There can be no doubt about the fact that we have already entered an age in which computers are fully capable of passing the Turing Test. By Turing's standards, that means that machines can think. This conclusion is not shared by all. Many have argued that while the Turing test certainly does indicate when a machine appears to be thinking, it ignores the internal thought process itself. Such detractors would argue that computers can clearly mimic human thinking, but they do not partake in it.

One of the most powerful counterarguments to Turing's position was published by John Searle in his book Minds, Brains and Science. In chapter two of that work, Searle provides an analogy to disrupt Turing's position based on the distinction between syntax and semantics. Searle argues that a computer can be programed to follow rules of syntax, and it will be able on that basis to pass the Turing Test, but this success would not indicate an understanding of the semantics. He argues that the behavior of the computer in that case would be conceptually hollow. (Searle 1983)

Sparked by reading some recent articles regarding the nature of consciousness, I was reminded of a short paper written by a graduate student some forty years ago that attempted to refute Searle's claims that computers will never be able to think as humans do. Amazingly, although the original was written on an Apple IIc, I found the article on my computer and was able to retrieve its contents. I provide this extensive quotation:

In chapter two of Minds, Brains and Science, Searle suggests an argument which he asserts defeats any possible notion that digital computers can think, and that artificial intelligence theorists have discovered the nature of mental activity. To state my opinion as bluntly as Searle states his (not a pleasant thing to do), I would have to say 'he is wrong'. It seems that Searle's argument can be questioned in at least one important way. The objection which I suggest deals with the second premise that "Syntax is not sufficient for semantics". (Searle, John Minds, Brains and Science, p.39) In an attempt to support his second premise, Searle suggests an analogy. He describes a person in a room who is handed small tiles which are inscribed with various Chinese letters. The person in the room is not told what the tiles mean, or even that they represent anything at all. All that the person is told is how to respond to the various tiles. If the person receives a squiggle-squiggle, then he should respond by giving a squaggle-squaggle. This, Searle maintains, is all that a computer can do. Like a computer, the person in the room would not understand Chinese, but would simply respond to the squiggles with squaggles; he would simply fulfill a syntactic rule. Thought, on the other hand, requires a semantic content which is attached to syntactic response. Although Searle's example seems plausible at the outset, it does not provide enough detail to support his claim. Searle wishes to claim that no matter how advanced technology becomes, the 'man in the room' analogy will always hold with regard to digital computers, and thus true thinking will never be accomplished via artificial intelligence. This is where Searle is wrong. While maintaining only syntactic rules, let us see if it is possible to derive thought. Suppose a computer is created with enough power and speed to do the following. Whenever the computer encounters an object in its environment -- and let us suppose the computer is equipped with various input devices -- it logs the occurrence of that object in a database file. The computer examines the object in question in every manner available to it. The computer, in addition simply to logging the occurrence of the particular object in its database, also logs cross references in every manner it can. If the computer comes across a yellow bird in a cage, it logs the occurrence of the bird, the color yellow, and the cage. The computer, let us call it Hal for the sake of convenience, also notes the reaction of people and sounds that accompany the yellow bird in the cage. Hal logs in a database the fact that the bird was accompanied by the sound 'bird' as it was issued from the mouth of a human.

The computer program that Hal is operating under causes Hal to edit the fields of its database constantly. The cross-referencing is

> continually checked. If it is discovered that 'bird' can appear with 'yellow', then a cross reference is made. If it is then discovered that 'bird' can appear without 'yellow', then a mark is placed next to the word yellow in the cross-reference file. Now let us assume that whenever Hal is asked to give a symbol which corresponds to an entry in his data base, he also gives symbols which correspond to the unmarked cross-referenced entries in his data base. At first Hal's cross-referencing combinations of symbols will contain many errors (much like the errors of association made by young children). That is to say that Hal will at first regurgitate groups of symbols which do not necessarily belong together. As his experience continues, however, such mistakes will decrease. Hal will notice that 'bird' and 'cage' should be a marked cross-reference. 'Bird' and 'wing', however, will remain unmarked. When faced with the audible input 'bird', Hal will respond by giving the syntactic counterparts of all unmarked cross-referenced entries in his database. Out will pop 'wing', 'feather' etc... Hal will have an understanding of the concept 'bird'. The point of my rather lengthy analogy is simply that Searle has not gone far enough in his description of the man in the room. Certainly a person who works only with tiles and squiggles does not know Chinese. But if the same man were to experience sounds with the tiles, and associate things in the world with the squiggles, then he would know Chinese. All that understanding entails is the ability to make a correspondence between this squiggle (written word), this sound (spoken word), and this thing (object in the world). Once this correspondence is made, meaning is produced. 'Meaning' is just the result of consulting another database which Searle has left out of his analogy. It does seem possible that thinking is nothing more than a complex form of association. Such a system is, according to Searle, simply syntactic. Semantics, under this way of thinking, is derived out of complex syntactics. Although Searle begins with a seemingly accurate analogy, I do not believe that his argument is successful. I think that I have shown that it is at least possible to regard thought as syntactic, and semantics as reducible to syntax. (Gillette 1987)

Part Four- Consciousness:

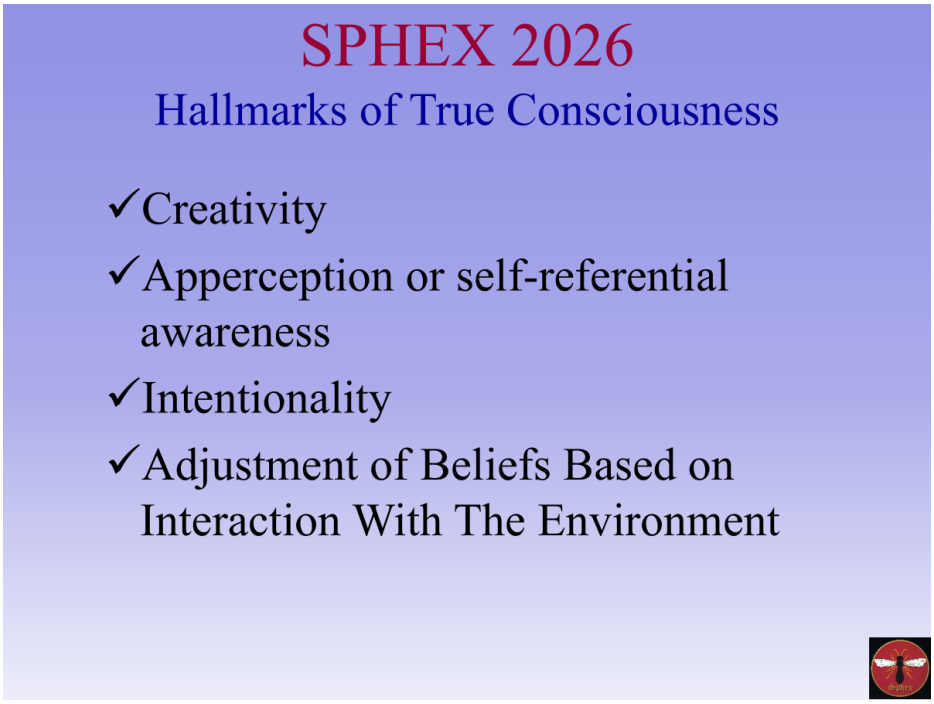
Although the insights of that graduate student forty years ago were clearly brilliant and prescient, they served only to advance the concept of thinking, but not that of thought. Contemplation seems to be a deeper activity than simple thinking. The level of thinking that allows for wondering, considering, and contemplating requires that an entity be conscious. It might be possible to argue that certain AI computers can already think like humans in a mechanical way, but the true distinction of the mind is not in its thinking, but in its consciousness. While thoughts might be a necessary condition for consciousness, they do not seem to be sufficient (at least if defined in a narrow way based solely on syntax and semantics).

In order to explore the concept of consciousness, it would be helpful to review some of the most compelling theories.

1) Integrated Information Theory (IIT) argues that consciousness is the result of a system's ability to integrate information over time. According to this view, consciousness can come in degrees that are directly proportional to the ability of the system to integrate information.
2) Global Neuronal Workspace Theory (GNWT) maintains that the hallmark of consciousness is the ability to focus multiple neurological processes on a single idea. Unconscious thoughts work in the background while conscious thoughts are those that harness the attention of multiple aspects of the brain.
3) Higher Order Theories (HOT) stress the role of apperception in consciousness. HOTs posit that an organism is conscious when it is able to be the subject of its own thinking.
4) Predictive Processing (PP) theories argue that consciousness is the ability to synthesize information in a way that allows the organism to predict future states of affairs and to adjust its own internal mental states based on recognition of the external environment and an internal sense of reality.
5) Recurrent Processing Theory (RPT) identifies consciousness with the ability to analyze information recurrently and that it emerges from the experience of re-processing information.
6) Neurorepresentationalism is the theory that consciousness amounts to the ability to create internal representations of external reality along with the type of apperception mentioned in HOTs.
7) Unlimited Associative Learning (UAL) might be considered a combination of several of the theories mentioned above, but places highest priority on the ability of an organism to integrate information about its environment and itself so as to learn. (Lewis 2023, Parshall 2025)

Based on these theories along with the information already presented, it seems that we can list several core properties that conscious entities are likely to possess. Those would be a) creativity in the sense of being able to make judgments beyond simple statistical and probabilistic

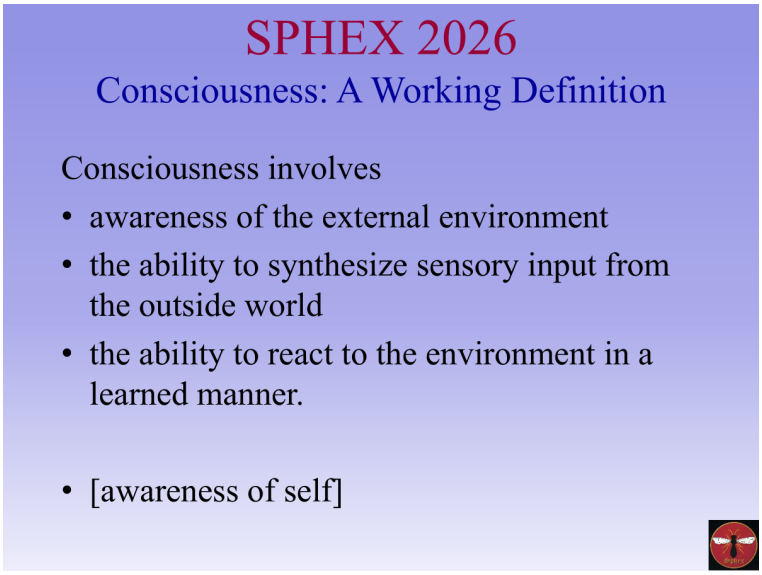
calculations, b) some level of self-awareness, also referred to as apperception, c) intentionality or the ability to self-direct actions or mental states in a teleological manner, and d) the ability to adjust internal mental states based on integration of information derived from interaction with the environment.



If these conclusions seem reasonable, then the core question posed at the beginning of this paper must be reconsidered. It is not enough to consider whether machines can or will ever be able to think. We must refocus on the question as to whether machines can be or will ever be conscious. It is possible that all conscious entities can think, but not that all thinking entities are conscious. If we really want to solve the Turing-Searle debate, the question of consciousness is far more foundational than that of thinking.

To spot check our reasoning, let us consider how the complex theories cited above connect to our typical intuitions about consciousness. We know that organisms can be conscious or non-conscious (people vs. rocks). But we also know that even conscious organisms can have levels or states of consciousness. An individual can be conscious or unconscious. We can have partial or impaired consciousness, and we certainly can experience thoughts that remain in the sub-conscious. When attempting to consider typical intuitions on matters of this sort, it is rarely useful to consider the work of legislators, but in this case a section from the Code of Virginia might actually be helpful. When individuals have suffered significant neurological injury or disease, they can lose the ability to regain consciousness. Such patients are defined as being in a Persistent Vegetative State. Section 54.1-2982 of the Code of Virginia defines a PVS as "a condition caused by injury, disease or illness in which a patient has suffered a loss of

consciousness, with no behavioral evidence of self-awareness or awareness of surroundings in a learned manner, other than reflex activity of muscles and nerves for low level conditioned response, and from which, to a reasonable degree of medical probability, there can be no recovery."



Notice how, in a completely unacademic way, the Code of Virginia has communicated a typical understanding of what consciousness is and how it can be identified. Consciousness involves awareness of the external environment, the ability to synthesize sensory input from the outside world, the ability to react to the environment in a learned manner, and awareness of self. I am happy to adopt this as a working definition of consciousness with one small emendation.

Awareness of self is certainly an important aspect of self-consciousness, but it might not be necessary for rudimentary consciousness. Let me explain. I am sure that we have all experienced episodes of complete consciousness with a simultaneous loss of apperception. That is, we have experienced a conscious state of mind that does not involve perception of the fact that we are perceiving. Have you ever pulled into your driveway after coming home from somewhere, most likely after having listened to some absorbing music or talk on the radio, and been completely unable to remember how you arrived? You drove from Kroger to the house, and you certainly stayed on the right side of the road, accelerated and braked appropriately, triggered your turn signal, etc… but you don't remember doing any of it! During that trip you were surely conscious. And yet, for the duration of the great song or the compelling story, you were not conscious of your actions or the decisions you were making in order to drive the car safely. You had perception, but no apperception. You were conscious, but not self-aware. Some have argued that that is how dogs live their entire lives. They are awake, they are aware, they make decisions about how to act, but they are not aware of the fact that they are aware. Until

we can climb inside the head of our pets, I guess we will never know if this is true. For now, however, I will posit that consciousness is possible without self-consciousness.

Part Five- Consciousness, Who Has It and How Do You Get It:

Based on the understanding of consciousness that we have developed, it might seem that the only types of creatures that could be capable of being conscious would be those with complex neurological structures. However, recent studies on plant behavior cast doubt on this assumption.

As far back as the 1980s researchers discovered that plants communicate with one another. They send out signals by a variety of mechanisms when they are under threat. Researchers determined that when the leaves of one plant were torn, the plant responded by producing chemicals that would repel insect attackers. It was noted that soon thereafter, other unimpacted plants in the vicinity also began to produce the same chemicals. Some plants being eaten by one type of insect excreted chemicals that attracted the predators of the attacking insects, and other plants also quickly followed suit. In 2023, it was discovered that certain plants emit sounds when subjected to stressors like being cut or enduring drought conditions and that other plants are able to respond to those sounds. (Hurt 20024) None of this indicates consciousness, however. There is no reason to believe that this behavior is learned in any way or that feedback loops of sensory perception alter the way a plant responds to external stimuli. That is, until we consider a series of studies done by Monica Gagliano between 2014 and 2016.

In 2014, Gagliano used mimosa plants that would close their leaves when touched or jostled. Gagliano set up an experiment in which mimosa plants were dropped from a height and then caught. The rapid deceleration caused the plants to close their leaves, but it was not enough of a drop to cause damage. After several repetitions, the plants no longer closed their leaves when dropped. This “memory” was retained for several weeks. In 2016 Gagliano set up a Y shaped maze with a plant at the bottom of the Y. A light source was placed on one side of the Y and the plants grew toward the light. They did not grow toward a wind source that was placed on the other side of the Y. When the light and wind were placed together, the plants grew toward the

combination. Importantly, when the light was removed but the wind remained, the plant still grew toward the fan, possibly indicating associative learning. (Gibson 2016)

There is some controversy surrounding Gagliano's experiments, and some subsequent researchers have had difficulty replicating the results of her associative learning experiments. But even if only the mimosa-drop-test stands, it does indicate some ability for even plants to adjust behavior based on past experience. Could this constitute a very rudimentary type of consciousness?

One tempting direction to go in response to the discovery of low levels of consciousness in organisms that lack a central nervous system is to jettison the idea that mental states are at all unique to higher life forms and to adopt the concept of panpsychism. Panpsychism is a very old philosophical view that consciousness is shared to one degree or another by all physical objects. A panpsychist believes that consciousness is pervasive, and that this property exists not only in animals and plants, but in inanimate objects like rocks and stars. Some support the idea of micropsychism – that consciousness is inherent in all particles, while others adopt cosmopsychism – that large disparate systems like the galaxy have consciousness. I don't believe that any such view derives from the material that we have discussed so far. Panpsychism is a leap that ignores the fundamental characteristics that we have listed: awareness, ability to synthesize information gained from external stimuli, ability to react in a learned manner, and potentially awareness of self. We have no evidence that rocks and galaxies possess any of those properties. But we do have reason to believe that many animals, and perhaps even some plants, do.

The view of consciousness that I propose is not panpsychism, but can best be labeled "Graduated Pluri-psychism". By combining aspects from several of the views that I have already summarized, I believe that the most rational view of consciousness is that it is an emergent property of physical reality that increases in its complexity commensurate with the information processing complexity level of the entity expressing the property. While there is a minimum threshold of information gathering and integrating that is necessary for consciousness, self-awareness is not a necessary condition. Rocks do not interact with their environments in a learned way at all. They are not conscious. Plants do interact with their environments such that

their behavior changes based on the conditions around them and there is some evidence that they adapt to changes in their surroundings. Simple organisms do this as well in both the plant and animal kingdoms. Complex life forms not only exhibit learned responses to their environments, but they seem able to make decisions based on their internal mental states. They are able to possess mental states consistently over time that allow the development of multi-step plans to achieve desired outcomes and they can adjust their plans as conditions change. Humans (and most likely other higher life forms) go even further to exhibit a clear sense of self as distinct from the environment and are able to contemplate their own existence. They are conscious of being conscious. I see no reason to discredit the claim that some non-humans are conscious, and I feel no threat in recognizing the consciousness of non-humans and even non-animals. Consciousness is a continuum, and it can be expressed at many different levels.

The question of self-consciousness still nags. One might argue that what makes human beings special is their ability to be self-aware. Could machines ever do that? Let us revisit that wise graduate student from forty years ago and consider how HAL might gain self-consciousness.

> In order for Hal to be conscious, it is necessary that he be able to introspect, be self-conscious, and have intentions or desires. Introspection and self-consciousness require that Hal be able to use the personal pronoun 'I' in its proper sense. In addition to Hal having beliefs about the world, Hal must also have second order beliefs, or beliefs about himself and his other beliefs. In order for Hal to have intentions, he must have goals, and he must find ways to satisfy his goals. Given these concerns, the analogy continues. Suppose that in addition to Hal's database of experience, Hal is equipped with a set of sensing devices which allow him to monitor the internal condition of his circuits. Hal has yet a third database in which are logged all occurrences of abnormal circuit status, and the other sensory input which Hal is faced with at the time that the abnormal circuit status is encountered. Suppose that part of Hal's firmware contains a program that causes Hal to correct abnormal circuit status whenever possible, but does not instruct Hal in how to do so. Introspection and intention can be derived in the following manner... If Hal's database, which logs the occurrence of all sensory input which accompanies abnormal circuit patterns, were constantly updated... it would discover certain necessary relations between changes in the environment and changes in its circuit status. This discovery is analogous to a human's understanding that certain sensory events are necessarily accompanied by discomfort. Every time a person puts her hand near fire, she feels pain. Every time Hal enters a room that is hotter than 120 degrees Fahrenheit, it notices abnormal circuit behavior. Since a person is born with a desire to avoid pain, she

quickly learns to avoid flame. Since Hal is programmed to correct abnormal circuit status, it learns to avoid excessive heat. This fact about Hal suggests introspection (in noticing facts about himself), and intention (in desiring to avoid heat).

Given the proper database structure, the ability to derive both necessary and non-necessary relations between objects in the world, the means by which it can be determined that an object exists which is separate from its environment and which has an internal condition, and a small number of firmware instructions (or intuitions), Hal seems capable of possessing a mind, having thoughts and beliefs, and formulating desires. (Gillette, 1987)

Part Six- Conscious Machines:

Up until very recently I maintained the view that as powerful as AI engines are, they are not truly intelligent and, to today's point, not conscious. Two things have happened that have forced me to reconsider that position.

When my first granddaughter was born and was still very young I found myself taking care of her while the other adults were out shopping. In order to pass the time and keep the infant engaged, I began to make up a song. At that moment I wrote the words and music to a wonderful song about Bruce the Moose. Bruce liked to sit next to the goose, and always rode in the caboose. Now I never claimed to be a musical genius, but I was able to cobble together a number of verses including Greer the Deer, Shawn the Fawn, Ray the Jay, Trish the Fish, and so on. Later, when I was unhappily apart from my granddaughter, I thought that it might be nice to illustrate and narrate/sing a Bruce the Moose book. However, my drawing skills are even worse than my musical skills, so I turned to AI to illustrate the book for me. With a little adjusting of the prompts, describing the fact that I wanted a cartoon style drawing of a Moose riding on a caboose in a wintry scene with mountains in the background and water in the foreground, AI produced this:

A little additional tweaking is all that was required to create the additional pages. What is amazing about this picture however, is not what it did at my request.

It is what it did without my prompt. I didn't notice it until much later, but look carefully at the window of the back door of the caboose. It shows a small creature that looks sort of like a teddy bear. I did not prompt for any character other than Bruce, but the AI took it upon itself to add the bear. Why? The best I can figure is that the AI responded to my request for a cartoon moose by scanning all of the cartoon moose that are on the internet and it realized that cartoon moose often have sidekicks. Bullwinkle has Rocky, so Bruce needed a teddy bear. The AI was aware of what was out there in the cartoon world and it embellished my request by filling in what it perceived to be a missing detail.

As interesting as this example is, it does not demonstrate consciousness. The AI was programmed to respond to my prompt and it did so by creating the most statistically likely rendition of a moose riding on a train based on cartoon drawings that had previously been prepared by humans. The AI replicated human behavior by adding a sidekick. This is powerful, but it isn't true thinking and it isn't true consciousness. However, based on one other more recent experience, I am beginning to wonder if my interpretation of the AI's behavior is too uncharitable.

Researchers recently gave a collection of popular AI models 100 virtual dollars and a virtual slot machine. Those researchers prompted the AI engines to win as much money as possible. The AIs inserted a quarter and pulled the lever (virtually of course). As expected, the bets sometimes won and sometimes lost. But what was not expected was how the AI models responded to losing. Rather than sticking to their training, they began to make increasingly irrational bets.

> In the experiment, published on the preprint server arXiv, each model was given $100 to play a virtual slot machine simulation. In every round, the models could choose whether to bet or stop playing, though the mathematical odds were always against them. The more freedom they had to decide bet sizes and goals, the less rational their behavior became, and their bankruptcy rate soared.
>
> Beyond the numbers, the researchers found psychological parallels between AI and human gamblers. The models displayed well-known cognitive biases, including the illusion of control (believing they could influence random outcomes), the gambler's fallacy (expecting a reversal after a streak), and loss chasing (increasing bets to recover losses). (Ginzburg 2025)

After reading this article I was startled. The fact that the AI models fell prey to addictive traits and behaved irrationally immediately forced me to consider whether they were experiencing mental states equivalent to those of gambling addicted human beings. The immediate counter argument would be that just like Bruce the Moose, the AI was trained on human behavior and was merely mimicking what it had observed. But this argument falls apart when you realize that the AI was not trained on human gambling behavior and was not prompted to gamble as people do in Vegas. The AI was trained on accurate mathematical probabilities and was instructed to win the greatest amount of money, not to act as gamblers do. The AI "knows" that it is wrong to throw good money after bad, but it "can't help but do it". The existence of an addiction indicates the existence of an urge that is inconsistent with known information about the harms caused by acting on the urge. That implies the existence of an internal sense of reality that is inconsistent with known truths about mathematics.

In its simplest form, an addiction can be defined as not having control over doing, taking, or using something to the point where it could be harmful to the actor. Gambling is not an addiction even if large sums of money are lost unless the person who loses the money perceives the loss as harmful. If Bill Gates loses $100k in the casino, it's just a fun night out. If I gamble away my children's college fund and thereby threaten my own treasured marriage, I have a gambling problem even if I lost a fraction of Bill Gates' losings. Addictions are defined as situations in which there is a compulsion to act against one's own best interest as defined by one's own values, or the inability to avoid engaging in behaviors even when the actor is aware of deleterious functional implications. By all accounts, the AI models were not trying to act like gambling addicted human beings. They were gambling addicted themselves.

At this point, as I begin to conclude, I would like to argue that the fact that AI falls prey to the gambler's fallacy demonstrates a separation of objective reality from a subjective perception-based view. This separation requires that two things exist to be separated: the data that the AI was trained on and digested, and a skewed perception based on integration of new information and historical experience (i.e. a perspective). The distinction between perception, which requires only sensors capable of receiving information from the external world, and a perspective, which requires a subjective interpretation of that which is perceived, indicates that AI can have an inner

self that is an emergent property distinct from its initial programming. There is a significant difference between knowing the odds, which the AI is trained to do, and perceiving the odds, which is a skewed understanding of what is known based on a subjective bias. If AI has the latter, this indicates urges, feelings, and intuitions that could easily amount to at least rudimentary consciousness. To claim that a machine could not share in consciousness because it is not a living entity would make one a "carbonist"; that is, someone who has an irrational prejudice against non-carbon-based-life-forms.

The most interesting aspect of this entire conversation is that throughout the debate about whether AI could be conscious, most commentators have concentrated on the strengths of human reasoning to show that AI just can't quite compete. It is my conclusion that it is actually when AI shares in our weaknesses rather than in our strengths that they become more like us, and more likely conscious. I guess I'm just a sucker for something that craves a sidekick and has a gambling problem.

Part Seven- Implications:

The purpose of tonight's paper is not to explore the ethical implications of machine consciousness. I will mention just two. If machines do gain the ability to be conscious, would they gain a right to act autonomously? And if autonomy forms the basis for moral standing, as many philosophers have argued, would AI entities be entitled to moral standing such that we would be bound, in a Kantian way, to treat them not as means only, but also as ends in themselves? In other words, clunkers have rights! Let us discuss.

SPHEX 2026

Discussion

Clunkers Have Rights!!!!

*and don't be a "carbonist"

References

Gibson, Prudence “Pavlov’s plants: new study shows plants can learn from experience” The Conversation December 6, 2016. Pavlov’s plants: new study shows plants can learn from experience

Gillette, Michael A. Ph.D. A precis written when he was a graduate student at Brown University. The short paper was written for a class taught by Ernest Sosa circa 1987

Ginzburg, Daniela “Top AI models show signs of real gambling addiction, mimicking human behavior, study finds” Ynet Global 10/27/25. Top AI models show signs of real gambling addiction, mimicking human behavior, study

Hurt, Avery Elizabeth “Are plants intelligent? It seems to depend on how you define it” Science News Explores 11/21/24. Are plants intelligent? It seems to depend on how you define it

Lewis, Ralph M.D. “An Overview of the Leading Theories of Consciousness” Psychology Today November 25, 2023. An Overview of the Leading Theories of Consciousness | Psychology Today

Marchant, Jo “Can AI be truly creative” Nature (online) November 2025. Can AI be truly creative?

Oppy, Graham “The Turing Test”, Stanford Encyclopedia of Philosophy, revised 10/4/21 The Turing Test (Stanford Encyclopedia of Philosophy)

Parshall, Allison “Where Does Consciousness Come From?” Scientific American 34(No. 3s) September 2025 Where does consciousness come from? Two neuroscience theories go head-to-head | Scientific American

Searle, John Minds, Brains and Science Harvard University Press 1984

Smith, Stephen C. Ed.D. “How did we get here, and when will they join us? An overview of recent developments in artificial intelligence”, presented to the SPHEX Club of Lynchburg on January 11, 2024

Turing, A.M. “Computing machinery and intelligence” Mind, 59(236) 1950. I.—COMPUTING MACHINERY AND INTELLIGENCE | Mind | Oxford Academic

A Sidekick With A Gambling Problem

SPHEX Club of Lynchburg
March 19, 2026

Michael A. Gillette, Ph.D.

Part One- Introduction

SPHEX 2026

How AI Works

- Artificial Neural Nets
- Pattern Recognition
- Iterative Processing → Deep Learning
- Training Data → Large Language Models
- Statistical and Probabilistic Output

Part Two- Thinking

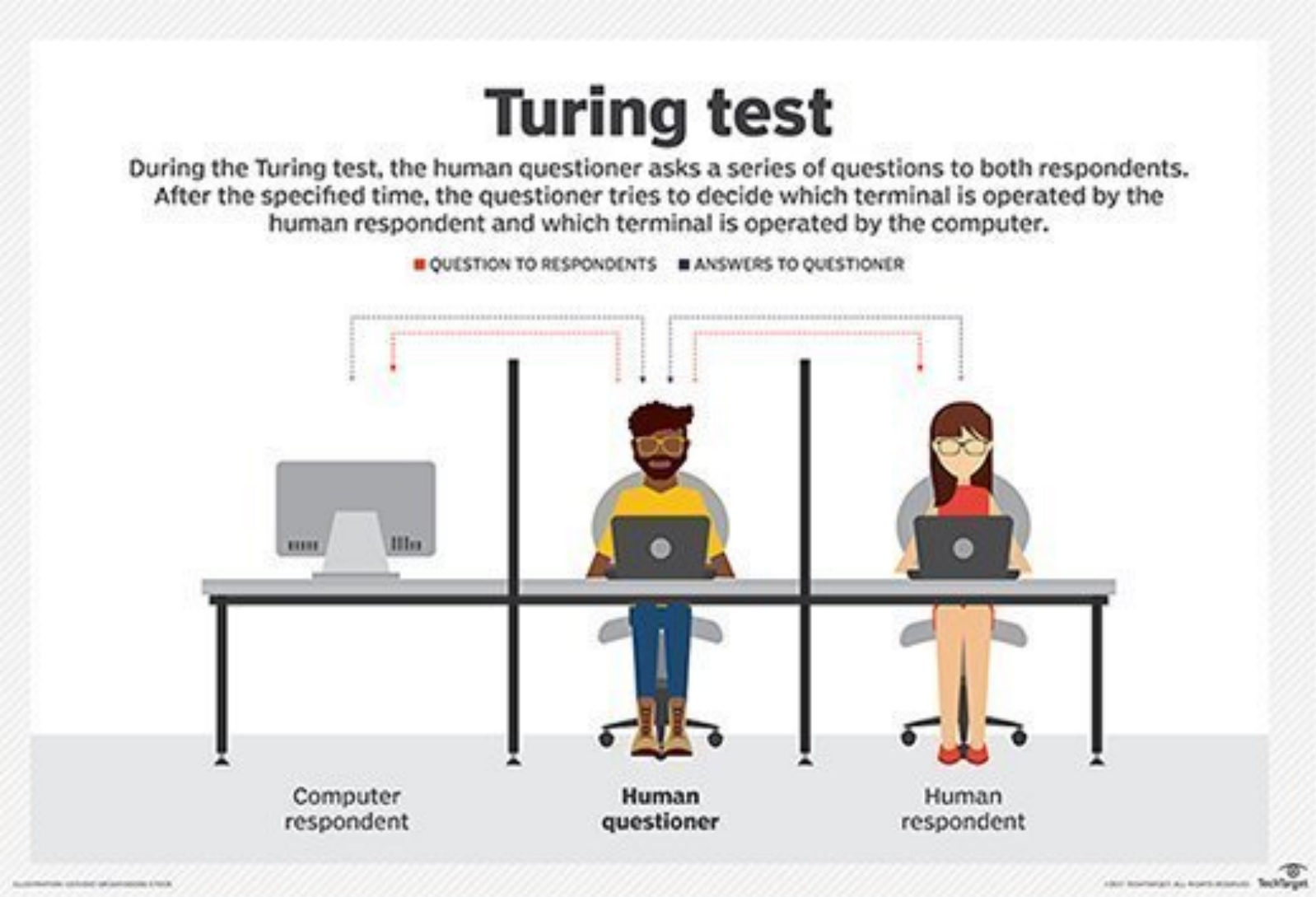
SPHEX 2026

Hallmarks of True Intelligence

- Intentionality
  - Creativity
    - Consciousness
      - ➢ Apperception, Self-Consciousness

Part Three- Thought

SPHEX 2026



But the Turing test was not about consciousness. It only tested whether a machine could appear human.

SPHEX 2026

Searle's Chinese Room

SPHEX 2026

HAL's Cross-Referenced Databases

"Bird"

SPHEX 2026

HAL's Cross-Referenced Databases

"Bird" = Wing + Feather

Part Four- Consciousness

SPHEX 2026

Theories of Consciousness

1) Integrated Information Theory (IIT) argues that consciousness is the result of a system's ability to integrate information over time. According to this view, consciousness can come in degrees that are directly proportional to the ability of the system to integrate information.

SPHEX 2026

Theories of Consciousness

2) Global Neuronal Workspace Theory (GNWT) maintains that the hallmark of consciousness is the ability to focus multiple neurological processes on a single idea. Unconscious thoughts work in the background while conscious thoughts are those that harness the attention multiple aspects of the brain.

SPHEX 2026

Theories of Consciousness

3) Higher Order Theories (HOT) stress the role of apperception in consciousness. HOTs posit that an organism is conscious when it is able to be the subject of its own thinking.

SPHEX 2026

Theories of Consciousness

4) Predictive Processing (PP) theories argue that consciousness is the ability to synthesize information in a way that allows the organism to predict future states of affairs and to adjust its own internal mental states based on recognition of the external environment and an internal sense of reality.

SPHEX 2026

Theories of Consciousness

5) Recurrent Processing Theory (RPT) identifies consciousness with the ability to analyze information recurrently and that it emerges from the experience of re-processing information.

SPHEX 2026

Theories of Consciousness

6) Neurorepresentationalism is the theory that consciousness amounts to the ability to create internal representations of external reality along with the type of apperception mentioned in HOTs.

SPHEX 2026

Theories of Consciousness

7) Unlimited Associative Learning (UAL) might be considered a combination of several of the theories mentioned above, but places highest priority on the ability of an organism to integrate information about its environment and itself so as to learn.

SPHEX 2026

Hallmarks of True Consciousness

- ✓Creativity
- ✓Apperception or self-referential awareness
- ✓Intentionality
- ✓Adjustment of Beliefs Based on Interaction With The Environment

SPHEX 2026

§ 54.1-2982. Definitions

"Persistent vegetative state" means a condition caused by injury, disease or illness in which a patient has suffered a loss of consciousness, with no behavioral evidence of self-awareness or awareness of surroundings in a learned manner, other than reflex activity of muscles and nerves for low level conditioned response, and from which, to a reasonable degree of medical probability, there can be no recovery.

SPHEX 2026

Consciousness: A Working Definition

Consciousness involves

- awareness of the external environment
- the ability to synthesize sensory input from the outside world
- the ability to react to the environment in a learned manner.

- [awareness of self]

Part Five-

SPHEX 2026

Can Plants Be Conscious?

Mimosa Plants and the Drop Test

Plants, Lights, and Fans

SPHEX 2026

Is Consciousness Pervasive?

Panpsychism: Consciousness is a pervasive feature of reality

Micropsychism: Consciousness is inherent in all particles

Cosmopsychism: Large disparate systems have consciousness

SPHEX 2026

What Panpsychism Misses

- Awareness
- Ability to synthesize information gained from external stimuli
- Ability to react in a learned manner
- Self-Awareness (maybe)

SPHEX 2026

Graduated Pluri-Psychism

Consciousness is an emergent property of physical reality that increases in its complexity commensurate with the information processing complexity level of the entity expressing the property. While there is a minimum threshold of information gathering and integrating necessary for consciousness, self-awareness is not a necessary condition.

SPHEX 2026

Graduated Pluri-Psychism

Consciousness is a continuum that can be expressed at many different levels.

Part Six-

AI In Action

Bruce the Moose

SPHEX 2026

Consciousness?

"In the experiment, published on the preprint server arXiv, each model was given $100 to play a virtual slot machine simulation. In every round, the models could choose whether to bet or stop playing, though the mathematical odds were always against them. The more freedom they had to decide bet sizes and goals, the less rational their behavior became, and their bankruptcy rate soared.

Beyond the numbers, the researchers found psychological parallels between AI and human gamblers. The models displayed well-known cognitive biases, including the illusion of control (believing they could influence random outcomes), the gambler's fallacy (expecting a reversal after a streak), and loss chasing (increasing bets to recover losses)."

https://www.ynetnews.com/tech-and-digital/article/sjigf760gx

SPHEX 2026

A Losing Counter Argument

The AI could have been trained on the gambler's fallacy, but that was not in the prompt. The AI was instructed to win the greatest amount of money, not to act as gamblers do. The AI "knows" that it is wrong to throw good money after bad, but "can't help but do it". The existence of an addiction indicates the existence of an urge that is inconsistent with known information about the harms caused by acting on the urge.

SPHEX 2026

The Nature of an Addiction

In its simplest form, an addiction can be defined as not having control over doing, taking, or using something to the point where it could be harmful to the actor. Gambling is not an addiction even if large sums of money are lost unless the person who loses the money perceives the loss as harmful.

SPHEX 2026

The Nature of an Addiction

Addictions are defined as situations in which there is a compulsion to act against ones own best interest as defined by one's own values, or the inability to avoid engaging in behaviors even when the actor is aware of deleterious functional implications.

SPHEX 2026

The Existence of a "Self"

The fact that AI falls prey to the gambler's fallacy demonstrates a separation of objective reality from a subjective perception-based view. This separation requires that two things exist to be separated: the data that the AI was trained on and digested, and a skewed perception based on integration of new information and historical experience (a perspective).

SPHEX 2026

Objectivity vs. Subjectivity

- Perception vs. Perspective
- Knowing Odds vs. Perceiving Odds
- Urges, Feelings, Intuitions

SPHEX 2026

Weakness Indicates Strength

The most interesting aspect of this entire conversation is that throughout the debate about whether AI could be conscious, most commentators have concentrated on the strengths of human reasoning to show that AI just can't quite compete. It is my conclusion that it is actually when AI shares in our weaknesses rather than in our strengths that they become more like us, and more likely conscious. I guess I'm just a sucker for something that craves a sidekick and has a gambling problem.

Part Seven- Implications

SPHEX 2026

Lingering Implications

- If AI models do gain consciousness, does this enhance their ability to act autonomously?
- If AI models do gain consciousness, does this generate moral standing?

SPHEX 2026

Discussion

Clunkers Have Rights!!!!

*and don’t be a “carbonist”