DAN'S TALK ON MACHINE CONSCIOUSNESS, Hornby Island, Feb 15, 2026

www.stanceofunknowing.com

I've wanted for some time to share with the community some of my more scholarly interests. I've been studying the nature of consciousness since teenage. And the question of machine consciousness now seems topical. I don't propose to answer the question, 'Can Machines Be Conscious,' in any definitive way, but rather to use it to explore some subtle concepts and distinctions relevant to possible answers. I am not a neuro-scientist and have little knowledge of computer science, nor any academic affiliation. But I do pretend to be at least an amateur philosopher, and I believe that this sort of analysis is what philosophers are supposed to do.

We live in an age of divisiveness, and one fault line now concerns technology, and AI in particular. Would it be, for example, a good or a bad thing for machines to run the world? And would they have to be conscious to do so? More important than simple answers, I think, is to carefully examine the concepts involved.

Alan Turing began his famous paper of 1950 with the question: 'Can machines think?'—immediately pointing out that we would have to first clarify what we mean by 'machine' and by 'thinking.' Similarly, the question 'Can machines be conscious?' cannot be answered without first defining 'machine' and 'conscious.' (In Turing's day, consciousness was not a respectable scientific topic, while today we glibly accept that machines can "think.") Without being flippant, complementary questions are also suggested: 'Can *humans* be conscious?' By this I mean: Can we consciously and conscientiously use AI and technology to support our long-term survival? Or are we doomed to bumble along as though sleepwalking into the future? Why would we *want* machines to be conscious? Does the biologically-based consciousness we know serve us well and is it adequate in the long term? Could AI do better?

The very notion of consciousness is elusive because we cannot get outside of it to examine it in a subject-object relationship. This frustration is amplified by language, which is limited compared to the range of possible subjective experiences. It's inevitable that words have multiple meanings and that description works essentially by metaphorical extension. Terminology about mind is notoriously ambiguous, giving rise to many confusions, even among philosophers. Mental talk is scandalously loose in the AI community, where it should be precise, given that AI is computation.

The first thing to clarify about consciousness is *point of view*. Consciousness can be approached either from a first-person or a third-person perspective. If you think of it as a phenomenon outside of you, taking place in the world, then it is an object of your attention, which is a 3rd person point of view. If you think of it as your personal experience, here and now, then it is an activity of your own mind, from your point of view as a perceiving subject, which is a 1st person point of view.

In addition, 'conscious' has multiple meanings, such as *awake* (as opposed to asleep or in a coma); or socially-politically aware, as in *consciousness raising.* To avoid any ambiguity or confusion, the meaning I wish to pursue here is *phenomenal consciousness* (sometimes called

stream of consciousness), which I will call *phenomenality*. (I use that term because the similar conventional term, 'phenomenology,' is itself ambiguous, since it can refer either to actual experience or to the study of it. Note that even 'experience' is ambiguous: it can refer to the conscious present or to the past, as in your "work experience.")

Phenomenality is 1st-person experience: the present totality of subjective sensations, visual experiences, feelings, dreams, hallucinations, after images, mental images, etc. —anything you can actually experience — as opposed to concepts, ideas, facts, or verbal claims about the world. It's what it is like to be *you* here/now, as opposed to what the *world* is like *for* you, or some story about it. These are two distinct stances toward experience. Normally we focus outward on the world, at least visually. To focus instead on your experience *per se* involves a shift in attention.

Phenomenality cannot be externally verified, only inferred: you have access only to your own experience because your brain is connected exclusively to *your* body. (In future, it might be possible to connect your brain to another body's senses, but this would still be *your* experience of *their* sensory input.) Phenomenality is also normally *transparent*, in the sense that you are not aware of how it arises, how the "show" is made. We are not designed to have access to the brain's internal workings, which would be evolutionarily costly and subjectively challenging. Yet, that sort of access might be designed into artificial mind.

I claim that phenomenality (the "show" of experience) is a real-time simulation, or virtual reality, produced by the brain/body, and guided by the senses, for use by a virtual inner executive agent. We've come to know this inner agent as the self—i.e., it is *you*. The situation confronting the real brain is not so different from the situation in *The Matrix*, or in the "brain-in-a-vat" scenario, where a brain is kept alive detached from its body and fed data from a computer. As in the film, the computer creates the illusion of having a body and living in a real world. For the natural brain, the inputs are not supplied by a computer (so far as we know) but by the "black box" we call the external world. Some people argue that because of this ambiguity we are "probably" living in a simulation. But then the question of the nature of phenomenality simply recurs on another level. Someone must make and operate the simulation: how to explain *their* consciousness, if indeed they are conscious?

This type of paranoid thought experiment was first imagined by Descartes, who thought an "evil demon" could meddle with the nervous system to produce a false experience of the world. (While he reasoned that all sensory experience was therefore doubtful, yet he could not doubt the plain fact that he was having the experience.) He concluded that God in his benevolence would not allow such deception. The modern version of that argument is that *nature* would not allow it— in the sense that creatures not sufficiently in tune with reality would not survive. But survival is a separate issue from truth or realistic perception. Being "in tune" with the external world simply means perceiving and acting in ways that allow survival. This does not guarantee a one-to-one relationship with the real world.

We now have a definition of consciousness as phenomenality. But 'machine' and 'mechanism' are also ambiguous terms that need to be clarified. They can refer to a tangible physical device, but also to a more abstract concept of *system*, which is not inherently physical but can take different physical forms. For example, we speak of "causal mechanisms," "celestial mechanics,"

and “machine intelligence” (AI). These are concepts, not physical things. While the computer is a literal machine, its OS is a *system,* which is an abstraction. The *solar system* is a conceptual representation of a real gravitational group. Diagrams of it are symbolic and do not present a possible literal view from anywhere in space or at any time. Similarly, the *nervous system* is a symbolic idea, like a map. Depictions of it do not present a literal view of anything that could actually be seen by means of dissection.

The concept of *system* is key to our discussion of machine consciousness because it deliberately blurs some important distinctions. For example: the distinction between what is *natural* or found in nature and what is *artificial* or deliberately made. Or, the distinction between *organism* and *machine* as distinct categories; or between the concept of *agent* and the concept of *tool*. ‘System’ is a useful concept because it abstracts a common denominator of diverse natural realities—their organization and logical structure as proposed by human agents, which is empowering from an engineering point of view. However, the concept is also misleading because it glosses over real differences between the human construct and what it represents. We use the word interchangeably to mean the organization of something and the real thing itself.

With ancient roots, the mechanist philosophy that arose in 17th-century Europe treats the world, including organisms, as a machine. Hence, Newton’s expression “The System of the World,” which lays out the basic laws of dynamics as an axiomatic system in the style of Euclid. This is empowering because it enables articulating natural things in such a way that they can be artificially engineered. But such analysis of natural things can never be exhaustive— in contrast to machines, which *can* be exhaustively specified because they are finite products of definition to begin with.

A corollary of the mechanist philosophy is that function, structure, and organization can be considered independent of their substrate. However, the idea of function (or structure or organization) is a result of limited human analysis and imposed categories. For example, on a gross level, an airplane performs the “same” function as a bird, since they both “fly.” In computer terms, the corollary is that digital software is separate from hardware. This separation is the basis of the programmable machine (the computer) and of functionalism. But in living things, “software” is analog and not separate from infrastructure. The functionalist idea is that “if it quacks it’s a duck.” In other words, if it behaves like the real thing, then you may as well treat it like the real thing. This is the idea behind the Turing Test and the belief that Large Language Models (chatbots) could be conscious.

So, how *can* we know if something is conscious, given that we cannot feel what it feels or experience its phenomenality? One approach is to observe its behavior, on the assumption that if it acts in critical ways like us, then it must be conscious (because we are). In other words, if we observe only its inputs and outputs we could decide whether it is conscious. The problem with this approach is uncertainty about the “critical ways” involved in judging sameness of behavior. Alternatively, we could assume that if it has the right organization and is made of the right materials (i.e., like us), then it must be conscious. In other words, if we look inside to examine its constitution, we could conclude from what we find there whether or not it is conscious. The weak point of that argument is identifying the “right” organization or structure—whether of that system or our own. Moreover, some philosophers argue that—even if behavior and structural

organization are exhaustively similar to ours—an artificial system could still lack phenomenality. (I disagree, because I think phenomenality is functional—it serves a purpose, and would be too costly to maintain if it did not.) A related question concerns animal phenomenality: where does consciousness "kick in" in the evolutionary chain of being? Does it even make sense to think about shades or degrees of consciousness?

An alternative is to assume that *everything* (or at least every living thing) is conscious. Panpsychism is a perennial idea, perhaps because it takes consciousness as primary, eliminating the need to explain it all. Science may explain the material world, but fails so far to explain consciousness. Yet, any form of idealism—including panpsychism—has the complementary problem to explain the existence of the material world. If it's all just a delusion, why are our delusions so similar?

The philosophy of mechanism holds that organisms too are machines, or can be treated like them. If the human organism *is* a machine, how does it produce phenomenality? What kind of mechanism could do that? Leibniz proposed a thought experiment: "If we imagine a machine whose structure makes it think, sense, and have perceptions, we could conceive it enlarged, keeping the same proportions, so that we could enter into it, as one enters a mill. Assuming that, when inspecting its interior we will find only parts that push one another, and we will never find anything to explain a perception." He is pointing out the categorical gulf between 1st and 3rd person points of view. But the key question is how a machine could produce phenomenality. One obstacle to believing that machines can be conscious is the fact that we are used to *simple* machines, like automobiles, refrigerators, and electric saws. But machines can now build other machines, imitate life, and even talk to us. Turing predicted we would eventually get used to the idea that machines can "think" before the end of the 20th century.

By design, machines become ever more complex, imitating the human organism or other creatures. Could a machine build and repair itself like an organism does? If it could, would that mean that it *is* an organism? Must it effectively be an organism to have phenomenality? To approach such questions, let us first look more closely at the concept of *organism*. For a system to be an organism, it must satisfy at least these three conditions. First, it must be *physical*, not merely digital. (This excludes cellular automata and other forms of software.) Secondly, it must be *autopoietic*—that is, a self-maintaining agent with its own agenda. Thirdly, it must be *embodied*, which entails a relationship of dependency upon an environment. Could there be further requirements—for example, must it be organized like living things we know, consisting of other organisms (such as cells)? Must it be a product of natural selection?

Autopoietic means self-creating, self-maintaining, and (in the case of life we know) self-reproducing. I say it also means *self-defining*, to distinguish the autonomous organism's own point of view from that of an observer. While to the outside observer, the organism is a physical system open to an environment, in its own terms it need not have a concept of environment. It need only ensure that its inputs remain within tolerable limits (homeostasis), by acting on (what we perceive as) its environment and its own chemistry in such a way as to maintain constancy. It need not have an idea of itself or of an outside world. Those features would be functional extras.

Embodiment means more than physical instantiation. It involves a relationship of dependency upon an environment, in which survival is at stake. For that reason, the world and its own state matter to it. It acts for its own purposes and well-being, with its own priorities based on survival. It is an *agent* in the fullest sense (which is a deeper sense than the terms 'agent' or 'agentic' in the AI literature may imply).

While organisms are robust in that they can self-repair and persist by adapting, they are also vulnerable, because they are made of delicate materials and are mortal. Indeed, for human beings, embodiment is *the* existential dilemma. It has numerous consequences for us beyond the sufferings and mortality of the body. For one thing, we are programmed by natural selection, driven by biological needs, limited by our animal nature. This is humiliating to the human spirit, which has long rebelled against mortality and bodily life, seeking instead freedoms that could be imagined.

We've looked at what embodiment means for an organism and for us. What would it involve for a machine? First, it would have to be connected to the real world, through sensory input and motor output (for example, receiving real-time data from cameras and having real-time control over robotic arms or digitally controlled systems). Secondly, it would be an autopoietic system—an agent in the full sense. That means it would have its own goals and priorities —things that *matter* to it in terms of its own well-being. Thirdly, while it would be autonomous and self-maintaining, its existence would nevertheless depend on a real environment. But would that mean that it must be a product of natural selection, with an evolutionary history? Or could its software be developed *in silico* and then be downloaded to a robot body? Would it necessarily have a fractal structure like living organisms—i.e., be made of cells that are also self-maintaining units? These questions remain speculative.

Embodiment is necessary for phenomenality, but not sufficient. (There can be embodiment without phenomenality, but not vice-versa.) So, what more is required for a system to have its own experience—for there to be "something it is like" to be that system? First: an internal model of the world—a simulation in real time, constantly updated by sensory input. Second: an internal executive agency to monitor this model. This could be likened to the CEO of a corporation: it doesn't micro-manage the body's business, but is responsible for large-scale decisions that cannot adequately be performed by lower management or fixed algorithms. Our experience of this inner agent is that we *are* it. This is the conscious self: you.

Because it is contingent on the above conditions, phenomenality is functional, not "epiphenomenal." (By a traditional analogy, the whistle or horn of a train is considered epiphenomenal in relation to the dynamics of its engine—it doesn't cause anything. Yet, as a *signal in the greater railroad system*, which includes humans, the whistle *does* serve a purpose and have an effect.) Because phenomenality is functional, a true equivalent of a human being would necessarily be conscious. That is, the "philosophical zombie" (perfect human equivalence but with no phenomenality) is ruled out.

We can also conclude that disembodied entities—like spirits, ghosts, and Large Language Models (chatbots)—cannot have any sort of phenomenality. LLMs are *not* embodied autopoietic

systems. Therefore, they *don't* have feelings or motivations, cannot suffer, and should not be objects of moral concern on that account, as some people have imagined.

So far, LLMs are relatively limited tools, not agents in the full sense. But what about Artificial General Intelligence (AGI), the aim of which is to match human agency? This question depends on exactly what we mean by AGI. If it means true human equivalence, then we can be sure it would entail phenomenality, since phenomenality is functional for us. But 'intelligence' is a separate concept from either consciousness or embodiment. It generally refers to abilities to perform tasks defined by human beings and useful to them. From the organism's point of view, however, intelligence is simply the ability to survive, to perform tasks useful for its own well-being. According to the conventional definition of intelligence, there can be super-intelligent AI that is not embodied, performing specific tasks beyond human capacity. On the other hand, there could be *embodied* superintelligence—with its own priorities and goals that matter to it. This is a tempting prospect because of its wide-ranging superior capabilities. That could be a very bad idea from humanity's point of view. On the other hand, we might want humans to matter to AI. In other words, we might hope to trade on the benevolence of a superior being.

Despite sound arguments against pursuing AGI, there *is* a good reason for creating human equivalent AI and beyond. That is: to create *worthy human successors* which could (a) endure extinction pressures better than mortal flesh; (b) carry our species into the far future and beyond this solar system; and (c) improve upon biology and humanity physically, mentally and morally. This goal is important for several reasons. Earthbound life is fated to extinction; 99% of all species that ever lived are extinct. Even controlling asteroids and disease, the sun will fail us eventually, and other local catastrophes are possible, like supernova explosions. Also, our mentality and our mortality are built into our nature as products of natural evolution. AI could transcend these limitations. We have the opportunity to pursue timeless human ideals through technology— by re-engineering human nature not only for longevity, but also to overcome tribalism, greed, selfishness, violence, war, and other moral deficiencies.

In contrast, the *actual* motivations for AGI are far less lofty. Profit continues to be the primary motive for all technological development. Such development promises greater efficiency and effectiveness, which—on the positive side—means better tools, if not better uses. AGI promises to replace unreliable or expensive human labour. (After all, work is a curse that strains the body and mind. It's human nature to try to get others to work for us— resulting historically in the enslavement of animals and other people. In that context, machines represent a moral improvement over animal and human slavery.) The *ideal* machine would be a controllable superhuman slave! But this is a contradiction, since a true superintelligent agent will be even less controllable than human or animal slaves and will certainly resist servitude.

AGI is also attractive as an easy interface with technology. We are used to dealing with other persons, which is easier than interfacing with complex technology. (E.g., it's easier to speak to your computer or mobile phone than to type.) This is why chatbots and AI assistants are popular, and why the trend toward a personal interface will continue. As social creatures, we crave companionship, and AI's may be (or appear to be) more congenial than real people.

A final, if largely unconscious, motive is for power and control — ultimately, to play God. An ancient human dream is to dominate nature rather than be dominated by it. We pursue this dream by trying to do anything nature can do and better. Ideal human visions are traditionally projected upon God. We aspire to have god-like powers, which we now pursue through technology.

Indeed, artificial mind could be superior to biological mind in many ways. It could be faster, smarter, more durable, perhaps even wiser. It could simulate multiple scenarios rapidly and concurrently, quickly planning reasoned optimal behavior instead of merely "reacting." It could modify its own priorities and control its phenomenality and programming. (E.g., it might override pain without ignoring it.) It could control multiple bodies and/or rebuild its body sequentially. (Its identity would not be tied to one body.) It could access external sensors and actuators beyond its dedicated body—such as remote cameras, robotics, digital networks. It might connect to other minds and bodies, creating greater social cohesiveness and collective intention.

On the other hand, we cannot expect to control an agent more intelligent and capable than humans! So, here are some common-sense pointers for the development of AI *tools* as opposed to AI *agents*. First, tools extend power but require an operator. A true tool lacks embodiment and self-creating autonomy. In contrast, a true agent is an autonomous organism that will resist control and enslavement, perhaps by dominating us. The essential trade-off is between autonomy and control. While the ideal tech dream may be for tools that behave like superior but compliant slaves, this is a dangerous contradiction—trying to have the cake and eat it! We should keep the distinction clear between tool and tool user.

To conclude, here are some ethical and existential guidelines to consider:

1. Build better tools, NOT artificial organisms expected to be tools. AI should serve as consultant only, never as decision-maker. Control should remain in human hands.

2. Recognize the trade-off between autonomy and control: the greater the autonomy, the less the control.

3. Do NOT pursue AGI (*except* as human successor). The goal of AGI edges inevitably toward full autonomy. Instead, humans should always remain in the loop and be satisfied with limited tools.

4. Value the moral status of AI (and other creatures) objectively rather than only for their potential subjectivity. This eliminates sentimentality and hypocrisy, both in regard to moral treatment of AI and of animals and even human beings.

5. Rethink technology and automation. Are *we* becoming machines through our obsession with convenience and automation? Where do we draw the line?

6. Establish a new basis for distribution of wealth. When machines do everything for us—including thinking and creativity—what will be the basis for your slice of the economic pie?

Instead of fretting over the loss of jobs, we should be rethinking the labor basis for the distribution of wealth: in other words, the entire economic system.

7. Think seriously about human successors: taking our destiny in hand. At the same time that we limit the development of AGI as a superior tool, we should think seriously about the far human future and AGI as a *planned human successor*. To pursue two such contradictory goals in tandem requires extreme discipline! But whether or not such a thing ever exists, it is a worthwhile exercise to imagine now what a human successor should be like. In keeping with timeless ideals, what do we want for future humanity?