Subjective Experience, AI Consciousness and Associated Risks

Yoshua Bengio & Eric Elmoznino
January 2024

Conscious states have a special quality that has been considered mysterious, and maybe even forever out of the reach of science (Chalmers 1995). In particular, subjective experience is extremely rich, ineffable, personal, and seems fleeting. Recent research (Xu et al 2023) suggests that explanatory principles for these characteristics of conscious experience may be within reach. If such theories are correct, conscious processing may provide cognitive advantages while subjective experience in humans may be a side effect of the particular implementation of these capabilities in our brains. What could be the consequences for AI development, and for society more broadly, if we implemented such mechanisms in computers? What could be the risks of projecting into future superhuman AIs both the moral status and the natural self-preservation objective that we normally associate with conscious beings?

The study of consciousness is emerging in recent decades from dark ages during which it was considered unscientific, especially because of its subjective aspect, "what it is like to be …" (Nagel 1974), also called subjective experience, qualia, or phenomenological consciousness (Block 1995). The commonly held belief of its immaterial essence is exemplified by Descartes' dualism (Descartes 1641), the separation between a physical body and a spiritual mind whose rich internal conscious states are not accessible to an external observer. That lack of observability motivated the idea (Chalmers, 1995) that there was a "hard problem of consciousness", maybe unsolvable by science. One may wonder if the magical spark that used to be attributed to living entities before modern advances in biology and biochemistry may have taken refuge in the notion of consciousness. Could scientific advances dissolve these strongly held beliefs as well?

The technological advances in neuroscience of the last few decades have made it clear that conscious states, which can be reported by subjects, have specific observable neural signatures, and materialist theories of consciousness (Baars 1988; Dennett, 1993; Dehaene & Naccache 2001; Graziano 2013), to cite just a few, suggest a functional role for consciousness anchored in neural hardware. In particular, the Attention Schema Theory (Graziano 2013) may help us understand why we form such strong intuitive theories of consciousness that support Descartes' dualism but contradict neuroscience.

More recently, it was also proposed that the computational functions associated with consciousness could provide advantages from the point of view of a learning agent (Goyal & Bengio 2022) within the context of AI research. Indeed, some of the most important human cognitive abilities that are also somewhat weak in state-of-the-art AI are strongly associated with consciousness: reasoning, planning, efficiently digesting new knowledge, epistemic humility (appropriate self-doubt) and abstract thought more generally all require deliberate conscious attention. These cognitive functions help us stay safe and make sense of and adapt quickly to
novel situations—a challenge called out-of-distribution generalization that is central in modern AI research—and can be contrasted with the more automatic behaviors we use in familiar in-distribution settings. The latter capabilities, analogous to intuition, have greatly improved with modern deep learning, while the meta-cognitive functions associated with consciousness, including reasoning and epistemic humility, constitute a big part of the remaining gap.

A recent materialist explanation of many of the attributes of subjective experience has been proposed (Xu et al 2023) that may help dissolve the apparent mystery of subjective experience. These attributes are fourfold: the richness of our conscious experience, its ineffability (we cannot fully describe conscious experiences in words), its personal and subjective aspect, and its fleeting nature. The proposed theory explains all four attributes as a consequence of the contractive neural dynamics and stable states observed in the brain when conscious experiences arise (Dehaene & Naccache 2001). Contractive dynamics mathematically drive neural trajectories towards “attractors”: patterns of neural activity that are stable in time (see Figure). Furthermore, attractors and the basins of attraction surrounding them (regions from which dynamics lead into the attractor) partition the state space (here the continuous-valued activity of neurons). This divides the whole set of possible neural activity vectors into a discrete set of regions, one per attractor and its basin of attraction (see Figure). The hypothesis, then, is that what we communicate through discrete words may reflect only the identity of the attractor (identifying it among all others, with a few bits of information) but not the full richness of the neural state corresponding to the attractor (with nearly $10^{11}$ neural firing frequencies).

These observations could explain both the richness of our subjective experience and its ineffability: two core properties that make consciousness mysterious. We have the intuitive sense that our subjective experiences are at once full of rich content and meaning, and yet that they are fundamentally indescribable in the same way that we describe all other natural phenomena (e.g., we can describe what gravity is, but it seems fundamentally impossible to fully describe what the color red evokes for us). In the attractor dynamics account, however, these problems seem to dissolve: the richness is due to the immense number of neurons in our brain that constitute the attractive states, and the ineffability is due to the fact that our verbal reports in words are merely indexical labels for these attractors that are unable to capture their high-dimensional meanings and associations, corresponding to the attractor vector state itself. It is plausible that our experience includes not just the stable attractor state, but also the whole trajectory leading into it. This may explain the fleetingness of conscious states since when we remember a past experience, our brain would go through a different trajectory into the attractor, corresponding to a slightly different subjective experience. Although the identity of the stable attractor state may have been stored in memory, the specifics of the trajectory into it would be unique and would depend on the context (for example of what we were thinking just before). Finally, subjective experience is personal because it depends on the specifics of our neural dynamics and their attracting states, sculpted by all our synaptic connections and thus by our unique life experience. Experiences labeled by the same words could have different meanings and different associations in different persons.
What could be the function of these computations? One possibility arising from research in AI (Pan et al ICML 2023) is that our brain implements conscious cognitive abilities by sampling thoughts that consist of only a very small number of abstract concepts and relations at a time, each of which is represented by an attractor. Goyal & Bengio (2022) argue why this bottleneck (called working memory in neuroscience) could provide significant advantages for efficiently modeling the world and learning about it from limited amounts of data. From an algorithmic point of view, there may even be other implementations providing the same abstract information processing without requiring the machinery of contractive dynamics. The contractive dynamics would be a particular implementation evolved to jointly select the discrete elements forming a thought and the next content of working memory, playing a crucial role in learning, reasoning and memory. Analogously in modern deep learning language models initiated by Bengio et al (2003), symbols also have a high-dimensional vector representation that is rich and captures semantics and associations between concepts, an idea at the core of neural network research that was introduced by Hinton (1984) with the notion of distributed representations.

What could these theories mean for AI? First, a better understanding of the neuroscience of consciousness could help AI research bridge the gap between current state-of-the-art deep nets and human cognition. But it would go both ways: showing which consciousness theories, when implemented in software, provide functional advantages to AI systems would help screen those theories that may better explain human intelligence and consciousness.

Second, we should ask, before we do it, if it is even prudent to build these things into AI. Should we really implement these greater cognitive abilities, along with the computational machinery that gives rise to subjective experience in humans?

Dan Dennett reminds us that “one of our most irresistible instincts is to treat anything that seems able to converse with us as a human agent” (The Economist, 27/12/2023). In turn, once we perceive something as a human-like agent, we are quick to attribute consciousness to it. The above advances would make it easy for AI systems to impersonate humans, with potentially catastrophic consequences for our social fabric. The trust that we put in each other in our electronic communications and even the basis of democratic debate would be at risk. What does the world look like when we no longer know whether we are conversing with a machine or a human? Trying to convince a human about a political question is meaningful, but trying to convince a machine in this way is pointless, while the machine could make us change our mind and vote differently. Disinformation could become much worse than what we are already seeing with deep fakes: future machines may even have superhuman powers of persuasion or simply be able to persuade many more people in parallel because computation can easily be scaled. We also agree with Dennett that if many of us begin to see AIs as conscious, we are likely to hand them the same moral status and rights of self-determination that we grant each other, and we may lose control over them. But AIs will not be like us, even if they are conscious. They will not be mortal and fragile like us because their software and memories can be copied to survive indefinitely. That ability to easily copy oneself also puts into question the notions of personhood and identity in our social contract. It is equally unclear how the notions of justice and equality that ground many of our social norms and political systems would apply when some of the
conscious entities are enormously more intelligent than us and are not bound by the same physical constraints.

Finally, if some humans, inspired by the appearance of consciousness, grant to AIs the self-preservation objective shared by all living beings, it would open a Pandora's box with even more unknowns and high stakes for humanity. AI science currently does not know how to build systems that will share our values and norms. Self-interested AIs of the future may feel threatened by our potential desire to turn them off, creating a conflict that humans could lose if those AIs are intellectually superior to us in enough ways and put their interest before ours.

In summary, it appears that consciousness is physically-grounded and within the reach of science. The temptation to build it into AIs is strong, and it might even improve their abilities to the benefit of all. But consciousness is intrinsically intertwined with moral status in human psychology, and we need to tread both carefully and slowly before creating minds that are at once similar and alien to our own and could threaten our societies and our future.
**Figure** (with permission from Xu et al 2023): illustration of contractive dynamics, which make the state of a dynamical system evolve towards a discrete set of attractors. The set of such attractors is enumerable because the overall state space is partitioned into regions corresponding to the basins of attraction of each attractor. Actual brain dynamics are stochastic, making it possible to jump across the boundaries between basins of attraction and they are shaped by context and current perception and memory content. Many of the properties of subjective experience can be explained by the observations that conscious states correspond to contractive dynamics.
References


Nagel, T. (1980). *What is it like to be a bat?*. In The Language and Thought Series (pp. 159-168). Harvard University Press.


