Journal of Artificial Intelligence and Consciousness

Technology and Consciousness Workshops (2017): An Introductory Overview

Damien Patrick Williams

Virginia Polytechnic Institute and State University
Blacksburg, Virginia 24061, US
damienw7@vt.edu

John Murray

San José State University
San José, California 95192, USA
jxm@acm.org

Received (Day Month Year)
Revised (Day Month Year)

This report introduces the activities of eight one-week workshops that were held during the summer of 2017 on the topic "Technology and Consciousness." Participants in the series of workshops approached the subject from many different perspectives, with the overall goal of exploring the possibility of machine consciousness, and assess its potential implications. The body of this initial paper summarizes the overview topics and basic introductory themes that were discussed during the early part of the workshop series. Follow-on papers will address the key focus areas that were examined in depth during the course of the full series.

Keywords: Machine consciousness; philosophy of mind; robotics.

1. Background

This paper draws upon some of the activities and findings of the first series of Technology and Consciousness Workshops (T&C Workshops), which were hosted by SRI International and undertaken during the summer of 2017. [Rushby and Sanchez, 2018] Eight, one-week-long workshops were held in various locations, each of which were attended by twelve to twenty participants. A total of fifty research specialists and practitioners participated in one or more of the workshops; their disciplines spanned a variety of interests, including neuroscience, robotics and artificial intelligence, computer science, philosophy of mind, contemporary physics, and cognitive science. Additional perspectives, including eastern and western spiritual and religious traditions, theories from noetic sciences, and psychoactive materials research and related cultures, were also represented.

The mission of the series was to review, present, and debate current state-of-the-art consciousness research, with the goal of informing thinking about the future capabilities of machines that are becoming increasingly intelligent. The participants sought to elicit new insights into critical dimensions of conscious human experience in the context of a multi-disciplinary framework, one that encompasses the numerous specialized insights of the participants. The scope included such ethical and moral considerations as may ultimately emerge and, potentially, their implications upon machine consciousness.

A number of the theories that were discussed characterized consciousness in the form of emergent properties that derive from the organization of, and processes that take place within, a neurobiological substrate. This may be contrasted with a view of consciousness as a specific feature that is explainable by the individual physical particles of which a brain is composed. Other theories that were explored emphasized that the development and content of consciousness is also contingent upon interaction of a physically embodied agent with its environment and culture. Emergence of consciousness can thus be viewed as drawing on evolution, and emphasizing "process" at multiple levels (brain, body, and complex reciprocal action of an organism with its physical environment and society).

The workshop participants were tasked with addressing four pre-specified objectives:

Characterizations of consciousness. The workshop stressed an interdisciplinary dialogue to achieve a common ground definition of consciousness across fields.

Potential mechanistic underpinnings of consciousness. Provided with a definition of consciousness, one can begin to explore the necessary requirements and potential basis for the existence of consciousness.

Metrics of consciousness. What are reasonable, and agreed upon, metrics of consciousness that allow us to assess consciousness in biological and machine agents?

Perspectives on machine consciousness. Consider the (speculative) implications of future machine consciousness, particularly for the safety and welfare of inhabitants of future societies.

The remainder of this paper summarizes the contributions presented during the early part of the T&C Workshop series. These set the framework for the overall series and offered a broad range of overviews of consciousness, as seen from various domains including neuroscience, biology, robotics, and artificial intelligence. Several philosophical perspectives on consciousness were also considered in depth, including phenomenology, qualia, subjective experience, and forms of embodiment.

Additional papers based on the T&C Workshops are planned, which will focus in greater depth on these topics, and the results of the exploration of other domains that were addressed during the series.

2. Multiple Perspectives on Consciousness

2.1. *The Problem of Consciousness:*

As a high-level introduction, David Chalmers surveyed the current state of consciousness research, and the philosophical directions that have been pursued and explored over the many years. He distinguished between phenomenal consciousness (subjective experience) and access consciousness (reasoning, reportable). The latter is the 'easy problem' of consciousness, as AI planning and reasoning have clear functional utility. On the other hand, explaining phenomenological consciousness requires subjective reporting with no compelling argument for its utility. The talk also covered a broad overview of philosophical theories of consciousness (from illusionism to dualism), complications related to measuring consciousness, and potential for machine/AI consciousness.

2.2. *Robot Consciousness:*

Antonio Chella applied Moore's Law to the human brain and suggested that robot consciousness might be realizable by 2029. He distinguished between “weak” and “strong” robot consciousness, whereby weak has already been captured with hard programmed types of intentional consciousness, such as reasoning or planning. One starting set of axioms for determining minimal requirements for artificial consciousness are that the agent can (1) depict, (2) imagine, and (3) attend to components of the external world, and it can (4) plan and (5) emote [Aleksander and Dunmall, 2003]. The ability to generate internal models of itself and the external world is a common requirement for artificial consciousness, and internalization and reflection may be a key component of consciousness [Minsky, 2007]. Computational models of consciousness include information integration theory (IIT) [Tononi, 2007], which enables the integration of complex information inputs.

2.3. *Philosophy of Mind:*

David Sahner presented what he described as “a brief and highly selective” tour of the philosophy of mind. He described how the historical consensus of consciousness and philosophy has shifted away from dualist theories towards those grounded in the underlying physical mechanisms. Even in the metaphysical sense, theories such as functionalism have evolved to explain consciousness in a more grounded way, such that it exists for a purpose. In more modern examinations of consciousness, biology and neuroscience have elucidated tremendous complexity in the human brain, which suggests that modeling efforts may be far more complex than pure “neuron doctrine” adherents would believe.

2.4. *An Android's Dream:*

Robin Zebrowski noted that neuroscience and philosophy have both developed arguments related to how consciousness is a result of inputs and interactions with a physical body, creating a world where consciousness is inseparable from embodiment. Research by Damasio [1994] in neuroscience and the linguistics work of Lakoff and Johnson [1999] provide the body-centric underpinnings of consciousness and cognition that embody our understanding and communication about the world. But studies in prosthetics and sensory substitution demonstrate the plasticity and flexibility of human information processing, suggesting that consciousness may be similarly fluid and dynamic, and extending one's body might translate to changes in one's consciousness.

2.5. *Machine Consciousness:*

Owen Holland identified the tremendous variability in neural structure and function, which makes the search for neural correlates very difficult. Better to focus instead on building a machine consciousness agent that can construct an internal model of itself and the external world. The talk discussed a number of anthropomimetic robots that aim to achieve a human-like physical (or simulated) representation in order to develop a complex self-model (CRONOS, SIMNOS, Ecce Robot).

2.6. *It May Not Feel Like Anything to be an AGI:*

Susan Schneider examined the potential for artificial general intelligence (AGI) and machine super-intelligence. Perhaps consciousness may be orthogonal to highly intelligent machines, in which case consciousness may be a property of AI that humans will have control over in the future. It may be possible to develop intelligent AI that is better served by not having consciousness, such as systems that are used in war. Also, maybe consciousness fundamentally relies on certain underlying physical properties, such as being made from carbon rather than silicon. As we develop more brain-machine interfaces, questions arise regarding how more invasive technologies might affect consciousness of the human host. Also, advanced AI systems could feasibly overtake humans as the next dominant species ("super sapiens"), with or without consciousness.

2.7. *From Friendly Toys to Extreme HRI:*

John Sullins reviewed the broad history of human creations – and the subsequent feelings humans have for these creations – and the depiction of AI in media (e.g., Blade Runner, West World, Iron Giant, Star Trek). The relationship aspect is important because it touches on the affective relationship humans have with robots and AI. While the West has traditionally considered robots to be functional, in that they perform work, Japanese researchers have embraced the idea of social robotics, or affective computing, which manipulate a human user's emotions. Towards understanding the human-robot relationship, there have been committees and initiatives started to explore these ideas. For example, what is the societal impact of sex robots?

2.8. *Time, Consciousness, Nonconsciousness:*

Julia Mossbridge discussed the complexities of empirically measuring consciousness, with particular emphasis on phenomenological consciousness. Neuroscience presumes that non-conscious processes far outweigh conscious processes and that individual conscious awareness is the result of brain activity. This asymmetric relationship is critical, since conscious awareness can only access information that non-conscious processes present. In other words, non-conscious processes handle interactions with the world, and consciousness is merely operating on pre-processed information. Complex, temporally-unconstrained non-conscious processes are forced into an ordered and local representation to support conscious processing. Thus, it is argued that we need to "understand the causative, generally nontemporal, and overwhelmingly global nature of nonconsciousness," because we cannot assume that the mechanisms that produce consciousness are temporally ordered and local.

2.9. *The Irreversibility of Consciousness, Modernized:*

Selmer Bringsjord addressed his previous postulation that, while standard computation is provably reversible, and consciousness apparently isn't, then consciousness cannot be computation. An early version of this argument was first discussed in the pages of Synthese [Bringsjord and Zenzen, 1997]. In his view, it is now time to revisit this model in the light of contemporary intellectual context, as well as accommodating phenomenological investigations of time. The upshot is that any such goal as that of powerful machines via systematic consideration of consciousness should be restricted to specific consideration of human-level cognitive consciousness (HLC-consciousness). There may be merit in exploring how consciousness can be modeled as data compression and an irreversible information processing action.

2.10. *The Troubles with Qualia:*

Mark Bickhard described two logical modeling problems with qualia. Firstly, since qualia are the subjective experience of experience, they are ontologically circular and thus impossible to model. Also, the emergence model means that they are not unitary phenomena, and so modeling must take into account unforeseen complexities. However, a normative, future oriented, pragmatist framework may help address consciousness phenomena as emergent in multiple types of agentive functioning in the world.

Irreversibility, as a key component of consciousness, allows a conscious agent to have a sense of normativity.

2.11. *Trained Phenomenology:*

Hank Barendregt characterized meditation as a form of trained phenomenology, in that an agent models the object, state, and action of reality as it is interpreted through the stream of consciousness. By removing ego, and a sense of self, from an experience, one can understand that feelings or emotions are actually states, rather than qualities of oneself. For example, anger can be portrayed as an object occupying one's consciousness, which needs to be managed, rather than a defining trait of one’s being. Through meditation, it is possible to train one’s ability to interpret feelings, events, and states, in a way that alter interpretations of free will.

2.12. *Automated Ethical Practical Reasoning:*

John Sullins focused on artificial moral agency and phronesis. The potential future danger is that agents will become more autonomous, but without any sense of ethics, leading to a freely-acting system that has no operational or functional morality. Microsoft’s Tay is an example of a fully autonomous learning system that ended up learning anti-social sentiments when left to learn from the internet. Various approaches were presented for how one might achieve an artificial moral or ethical agent, and how at least attempting to incorporate ethical reasoning into a system is a worthwhile endeavor.

2.13. *Where Yat? Epistemic Attacks on the Boundaries of Self:*

Paul Syverson examined the “sense of self” and its role in consciousness as separate from the outside environment. In the future, it may be possible for adversarial attacks to focus on vulnerabilities related to an AI’s sense of self. For instance, it may be possible for an adversary to manipulate the properties or signals that an AI attributes to itself and/or others (beliefs). One way to confuse characteristics or beliefs is for the adversary to determine how the AI's predicates change with context, and then spoof its external influencers and input sensors. Provoking the wrong configuration in an inappropriate circumstance can lead to significant operational vulnerabilities. Information security practitioners need to consider new ways that adversaries can attack the boundaries of an AI's sense of self and knowledge.

2.14. *Mental Qualities Without Consciousness:*

David Rosenthal explored the complexity of empirically testing mental experience, in particular the difficulty of separating conscious and non-conscious contributions to experience. It is possible that many mental qualities are actually due to non-conscious processing and perception. The latter, which is often equated with qualia, is thought to be highly dependent on non-conscious states and properties. Given the complexity of separating mental state from conscious experience, this observation raises questions related to the function of consciousness. If mental qualities can exist outside of conscious awareness, then what does being conscious actually mean?

2.15. *On the Possibility of Synthetic Phenomenology and Intersubjectivity:*

Robin Zebrowski identified the variability resulting from embodiment as a core concept for phenomenological consciousness. Following this argument, conceptualization of machine consciousness may benefit from it having similar "inputs" to humans. As many researchers in philosophy and neuroscience have noted, consciousness is intertwined with the inputs it receives, which links how the physical body shapes conscious experience and extends to the innate needs and motivations that biological beings, possess. Thus, the engineering of human needs and drives may be a fruitful route for exploring machine consciousness. Just as the physical body influences consciousness, through embodiment and needs, physical representations of systems influence interactions and feelings of shared experience. The talk ends with discussions related to anthropomorphic robots/systems and AI that can create art, demonstrating that these experiences can essentially fool humans into a sense of sharing an experience with a machine (that likely has no conscious sense of experience).

2.16. *The Minds of Others: What is Known by and Owed to Nonhuman Persons:*

Damien Williams discussed the complexities of human constructs and their significant impact on human behavior and well-being. Social constructs around race and gender have important implications on policy, but they are also relative and individual-based; for example, a black woman's sense of gender is different than that of a white man. Such constructs are beginning to appear in modern AI systems (intentionally or not), thereby creating questionable artificial morality. If our future intelligent systems are based on our human moral behaviors, it is reasonable to foresee re-creations of ethical atrocities, such as eugenics or the Tuskegee syphilis experiments. Similarly, if we create new kinds of conscious minds, they are likely to be subject to many of the same prejudices and harms that have been and are still levied against living human subjects.

3. Acknowledgements

The authors would like to acknowledge the extensive work of John Rushby and Daniel Sanchez in preparing the SRI Technology & Consciousness Final Report, upon which the content of this paper is based. Credit is also due to Antonio Chella for his encouragement and assistance with the editing and review of this material. Generous funding support by the anonymous sponsors of the T&C Workshop series is recognized on behalf of all the organizers and participants.

References

Rushby, J. and Sanchez, D. [2018] *Technology and Consciousness Workshops: Final Report.* (SRI International) http://www.csl.sri.com/~rushby/abstracts/techconscwks2017

Aleksander, I. and Dunmall, B. [2003] *Axioms and tests for the presence of minimal consciousness in agents I: Preamble*. Journal of Consciousness Studies, Vol. 10(4-5). pp. 7–18.

Minsky, M. [2007] *The emotion machine: Commonsense thinking, artificial intelligence, and the future of the human mind*. (Simon and Schuster).

Tononi, G. [2007]. *The Information Integration Theory of Consciousness.* In Schneider, S. and Velmans, M., eds, The Blackwell Companion to Consciousness. (Blackwell Publishing).

Damasio, A. [1994] *Descartes' Error: Emotion, Reason, and the Human Brain.* (Putnam Publishing).

Lakoff, G. and Johnson, M. [2008] *Metaphors We Live By.* (University of Chicago Press).

Bringsjord, S. and Zenzen, M. [1997] *Cognition is not computation: The argument from irreversibility*. Synthese, Vol. 113(2), pp. 285–320.