Towards Artificial Phronesis: Some Potential First Steps Along the Road to Moral Agency Using Case Studies

We will survey recent AI regulatory and policy activities from several perspectives, to understand their potential for helping the development of artificial moral agency, or Phronesis.
Consciousness (ICOMTC) and associated related consciousness theories (CTM, consciousness measures as applied to ICOM systems and their uses in context including defining of the basic assumptions for their functioning of cognitive systems, it remains indirectly relevant through axiology since we have reasons to care whether other beings have or lack their first-person feel of the world (the Church-Turing Lovers argument). In this paper I sketch out a deflationary theory of non-reductive first-person consciousness. Reflection on machine consciousness helps define such motions; it is also helped by it since non-reductive consciousness be- comes a bit more of a technical term employable in investigating future AI.

Four Preconditions for Solving MC4 Machine Consciousness

A machine is MC4 conscious if it has phenomenal experiences that are comparable to human conscious experiences. From an ethical point of view it is important to know whether we have created MC4 consciousness in a machine. MC4 consciousness research can also contribute to the development of general theories of human consciousness. This paper discusses four problems that have to be solved before we will be able to address MC4 machine consciousness in a systematic way. We need more clarity about the measurement of consciousness, we need better ways of describing the physical world and consciousness, and we need to reach agreement about the final form that a theory of consciousness should take. When these problems have been addressed we will be able to develop scientific theories of consciousness that can make accurate believable predictions about MC4 consciousness in machines.
A novel framework for symbol grounding in artificial agents is presented, which relies on the key idea that concepts are emerging clusters in multiple conceptual spaces. Symbols will be learned from color, texture, shape, and position that in turn are the properties of the objects populating the agent's environment. Objects are represented in a suitable object conceptual space where all their features are composed together again using clustering in such spaces. Such spaces describe percepts such as color, texture, shape, and position that in turn are the properties of the objects populating the agent's environment. Objects are represented in a suitable object conceptual space where all their features are composed together again using clustering in such spaces. Symbols will be learned from such a tensor space. A detailed description of both the framework and its theoretical foundations are reported and discussed in this work.

---

**04:30pm - 04:45pm Jeffrey L. Krichmar**  
A Neurobiologically Inspired Plan Towards Cognitive Machines

Despite incredible recent progress in artificial intelligence, current systems fall short of what we would consider to be intelligent, thinking machines. This paper presents a neurobiologically inspired path towards creating cognitive machines. It suggests that incorporating aspects found in biological organisms, such as flexible learning, efficient processing, embodiment, value systems, and predictive coding could lead to systems that are truly cognitive.

**04:45pm - 05:00pm David J. Gunkel**  
No Brainer: Why Consciousness is Neither a Necessary nor Sufficient Condition for AI Ethics

The question concerning the moral and/or legal status of others is typically decided on the basis of pre-existing ontological properties, e.g., whether the entity in question possesses consciousness or sentience or has the capacity to experience suffering. In what follows, I contest this standard operating procedure by identifying three philosophical problems with the properties approach (i.e. substantive, terminological, and epistemological complications), and I propose an alternative method for defining and deciding moral status that is more empirical and less speculative in its formulation. This alternative shifts the emphasis from internal, ontological properties to extrinsic social relations, and can, therefore, be called a "relational turn" in AI ethics.

---

**05:00pm - 05:15pm Henry Shevin**  
To Build Conscious Machines, Focus on General Intelligence: A Framework for the Assessment of Consciousness in Biological and Artificial Systems

**05:15pm - 05:30pm Discussion**
David Chalmers (among others) is fond of saying that consciousness has no function; it can be there or not - it makes no difference to behavior. In that sense, it supposedly is not like a pumping heart that helps keep one alive. Here we argue to the contrary: that consciousness has a critical function, and one that AI will be forced to deal with as a practical matter, as we probe more deeply into realtime commonsense reasoning. We will draw on a broad range of work - philosophical and otherwise - in making our argument.

02:30pm - 02:45pm Ricardo Sanz, Julita Bermejo-Alonso
Consciousness and Understanding in Autonomous Systems
This position paper will highlight the importance of having a formal notion of understanding as one of the cornerstones in the construction of conscious AIs. It will show that the capability of understanding both the perceptual and the action flows is critical for the correct operation of situated autonomous systems. An assessment is also made on the contribution of the machine learning domain towards this direction.

02:45pm - 03:00pm Lorijn Zaadnoordijk, Tarek R. Besold
Artificial Phenomenology for Human-Level Artificial Intelligence
For human cognizers, phenomenal experiences take up a central role in the daily interaction with the world. In this paper, we argue in favor of shifting phenomenal experiences into the focus of human-level AI (HL-AI) research and development. Instead of aiming to make artificial systems feel in the same way humans do, we focus on the possibilities of engineering capacities that are functionally equivalent to phenomenal experiences. These capacities can provide a different quality of input, enabling a cognitive system to self-evaluate its state in the world more efficiently and with more generality than current methods allow. We ground our general argument using the example of the sense of agency. A key issue within the area of putatively conscious AI systems is that of whether a wholly virtual system can...

03:00pm - 03:15pm Christian Balkenius, Trond A. Tjøstheim, Birger Johansson
Arousal and Awareness in a Humanoid Robot
We describe how an arousal system that controls the levels of awareness can be implemented in a robot. The different levels of awareness correspond to different states of consciousness and we argue that an artificial arousal system modeled after its biological counterpart has a useful function in controlling the cognitive processing of a brain-like cognitive architecture. The level of awareness depends on arousal that in turn is controlled by novel or emotionally charged stimuli as well as by a circadian clock. Arousal is also modulated during cognitive tasks to control the randomness of decision processes and to select between exploration and exploitation.

03:15pm - 03:30pm Discussion
03:30pm - 04:00pm Break
Tuesday, March 26 Afternoon Session 2
Session Chair: Jonathan Pfautz

04:00pm - 04:15pm Selmer Bringsjord, Naveen Sundar Govindarajulu
Introducing Λ for Measuring Cognitive Consciousness
In this brief work-in-progress document we introduce Λ, a new and novel framework for measuring cognitive consciousness that stands in dramatic contrast with the longstanding Φ of Tononi (2012), which gives a measure of phenomenal consciousness. Our plan herein is straightforward: First, we quickly distinguish between these two radically different types of consciousness (§2). Because Λ is erected atop a distinctive foundation developed by us, we rapidly describe three salient parts of this foundation (§3). In §4 we give a glimpse of the technical side of Λ, and rely there on an example of moral cognitive consciousness. The penultimate section, §5, lists a few distinctive properties (corresponding to underlying theorems) of the Λ framework; and then we wrap up with a concluding remark.

04:15pm - 04:30pm Naveen Sundar Govindarajulu, Selmer Bringsjord
Towards a Computable & Harnessable Model of Consciousness
We present a computable model of consciousness that is a modification of an existing model of universal computation. This modification is partially motivated by two existing, and non harnessable, models of consciousness. We say a model of consciousness is harnessable if the following statement holds: if the model predicts that a system v is more conscious than another system u, then we should, in general, find v more useful than u in a wide range of tasks. While there are no domain-general definitions of what makes a system harnessable, we give a preliminary proposal here and assess our model against this yardstick.

04:30pm - 04:50pm Owen Holland
Can a Virtual Entity Support Real Consciousness, and How Might This Lead to Conscious Robots?
A key issue within the area of putatively conscious AI systems is that of whether a wholly virtual system can ever be capable of supporting real consciousness. This paper first considers the theoretical implications if such systems using digital technologies could in fact exist, and then explores the consequent practical implications for the creation of virtual and real conscious systems. It is argued that the best strategy is probably to first create a virtual consciousness with components analogous to a virtual embodiment, and then to migrate the conscious core by degrees into a physical embodiment.

04:50pm - 05:00pm Joscha Bach
Phenomenal Experience and the Perceptual Binding State
How can a computational model of cognition account for the hard problem of consciousness? This contribution addresses some of our intuitions about the nature of phenomenal experience and the first person perspective, and suggests avenues for their realization in a cognitive architecture.

05:00pm - 05:15pm Susmit Jha, John Rushby
Inferring and Conveying Intentionality: Beyond Numerical Rewards to Logical Intentions
Shared intentionality is a critical component in developing conscious AI agents capable of collaboration, self-reflection, deliberation, and reasoning. We formulate inference of shared intentionality as an inverse reinforcement learning problem with logical reward specifications. We show how the approach can infer task descriptions from demonstrations. We also extend our approach to actively convey intentionality. We demonstrate the approach on a simple grid-world example.
05:15pm - 05:30pm Discussion

06:00pm - 07:30pm Plenary Session: Jordan Hall (420-040)
Plenary talk by Owen Holland for "Towards Conscious AI Systems."

Wednesday, March 27 - Morning Session 1
Session Chair: David Gamez

09:00am - 09:15am Johnathan Charles Flowers
Reconsidering the “Artificial,” the “Intelligent,” and the “Conscious” in Artificial Intelligence and Machine Consciousness through American Pragmatism

Abstract. This paper aims to reconsider the “artificial” nature of AI not as a result of the simulation of the human organism, but as the result of a disconnection from experience. Experience, in this context, is defined as the ways in which an organism is in nature through culture or has a world which is defined by a universe of discourse. To this end, an artificial intelligence is only “artificial” insofar as it is disconnected from the ways in which the world is experienced as mediated by culture. Put simply, artificial intelligence will cease to be “artificial” the moment it is implicated in nature through culture, at which point it will become “intelligent” or “conscious.”

09:15am - 09:30am Johnathan Charles Flowers
Strong and Weak AI: Deweyan Considerations

Work in artificial intelligence and machine consciousness is often discussed using Searle’s (1980) distinction between Strong and Weak AI. Weak AI presents AI as a tool for solving problems, whereas Strong AI is the generation of an “actual” mind. This paper will reconsider the possibilities of Strong and Weak AI in the context of John Dewey’s naturalistic pragmatism to recast our understandings of the qualities of “weak” and “strong” AI, and ultimately present the two as in continuity with one another.

09:30am - 09:45am Julia Mossbridge
Time and (Un)conscious Processes - Predictive Anticipatory Activity and Potential Applications

Predictive Anticipatory Activity (PAA) is the physiological and behavioral activity in an organism related to gathering accurate information about future events not through the usual senses, inference, or directly causing the events themselves to occur. It has been demonstrated and replicated in multiple independent and controlled laboratory experiments examining human behavior and physiology and in two animal experimental systems. Aside from the versions of PAA explicitly developed through conscious training, spontaneous PAA may represent an unconscious attempt to prepare organisms for future events. The mechanisms underlying PAA are unknown, and it is not clear that physical laws actually forbid it. It is thus possible that information about physical events is time symmetric in nature, and that conscious experience generally only presents us with a unidirectional flow that we call the “arrow of time.” Based on these ideas, in this position paper the argument is made that PAA can be thought of as a glimpse into physical reality, not as it is presented to us via the mechanisms that create our conscious experiences, but as it is “beyond the veil” of the conscious mind. Potential research and practical applications for PAA, especially with regard to consciousness in AI systems, are briefly discussed.

09:45am - 10:30am Discussion on Future Directions in Machine Consciousness Research

10:30am - 11:00am Break

Wednesday, March 27 - Morning Session 2
Session Chairs: Antonio Chella, David Gamez, Patrick Lincoln, Jonathan Pfautz

11:00am - 12:00pm General discussion
12:00am - 12:30am Conclusions and wrap-up