Conscious Machines for Autonomous Agents and Cybersecurity

This paper was downloaded from TechRxiv (<https://www.techrxiv.org>).

LICENSE

CC BY 4.0

SUBMISSION DATE / POSTED DATE

21-12-2021 / 28-12-2021

CITATION

Kadin, Alan (2021): Conscious Machines for Autonomous Agents and Cybersecurity. TechRxiv. Preprint.
<https://doi.org/10.36227/techrxiv.17308406.v1>

DOI

[10.36227/techrxiv.17308406.v1](https://doi.org/10.36227/techrxiv.17308406.v1)

Conscious Machines for Autonomous Agents and Cybersecurity

Alan M. Kadin
Consultant
Princeton Junction, NJ USA
amkadin@alumni.princeton.edu

Abstract—Although consciousness has been difficult to define, most researchers in artificial intelligence would agree that AI systems to date have not exhibited anything resembling consciousness. But is a conscious machine possible in the near future? I suggest that a new definition of consciousness may provide a basis for developing a conscious machine. The key is pattern recognition of correlated events in time, leading to the identification of a unified self-agent. Such a conscious system can create a simplified virtual environment, revise it to reflect updated sensor inputs, and partition the environment into self, other agents, and relevant objects. It can track recent time sequences of events, predict future events based on models and patterns in memory, and attribute causality to events and agents. It can make rapid decisions based on incomplete data, and can dynamically learn new responses based on appropriate measures of success and failure. The central aspect of consciousness is the generation of a dynamic narrative, a real-time model of a self-agent pursuing goals in a virtual reality. A conscious machine of this type may be implemented using an appropriate neural network linked to episodic memories. Near-term applications may include autonomous vehicles and online agents for cybersecurity.

Keywords—artificial intelligence, autonomous systems, cyberspace, pattern recognition, machine learning, computer security, virtual reality, neural networks.

I. WHAT IS CONSCIOUSNESS?

Although neural networks have become ubiquitous in computing technology, one aspect of natural intelligence that has not yet been successfully emulated is biological consciousness [1,2,3]. Much of the difficulty is that the property of consciousness in biological systems is not well defined or understood, so that it is unclear how to emulate it. Consciousness is generally viewed as more of a question in philosophy or religion, rather than science or technology. This is because the primary perception of human consciousness is as a thinking self, separate from the rest of the world. This perception forms the basis for Cartesian dualism, also known as mind-body dualism, which argues that the mind has a different nature than the physical world.

In contrast, I argue [4] that this paradox can be resolved by regarding consciousness as the functioning of a specific architecture of a neural network that features self-recognition as a primary aspect. Neural networks are well known for pattern recognition, and temporal pattern recognition identifies correlations in sequences of events in time. In particular, such a neural network may be configured to recognize objects and agents in the environment, as well as the self, and to generate a simplified virtual reality (VR) representation. The internal sense of consciousness is associated with the repeated updating of this VR, which provides a causal narrative of the self, interacting with the other components. Note that this does not require any philosophical self-reflection.

The conscious mind may actually be less in control than we tend to think. Cognitive scientists have shown [5] that most of our actions are actually controlled subconsciously by the unconscious mind, even if the conscious mind takes credit for these actions. The unconscious mind may act according to standard rules, similar to conventional AI. Only if these rules do not apply is this passed along to the conscious mind, for a conscious decision based on a higher-level evaluation, based in part on patterns of prior memories. Biological neurons are quite slow (~ ms), so that unconscious actions are likely to be much faster than the additional serial steps needed for conscious decisions.

From this viewpoint, consciousness is not uniquely human, and may not require language or abstract reasoning. It may be a capability that evolved in animals long ago, to deal flexibly with unfamiliar environments. So a dog or a horse may be quite conscious, even if it can't tell you so.

The consciousness module can also be turned off in humans and animals; it's called sleep. It can also be activated during sleep, but without external environmental inputs; these are dreams. Semi-coherent dream narratives involving the self in an unfamiliar environment may provide insight into the functioning of the VR generating module.

Brains in biological systems are based on large arrays of neurons, where each neuron may have thousands of synapses connecting it to other neurons. For example, the cerebral cortex of a mouse may have about 4 million neurons, and that of a human may have 16 billion neurons [6]. In comparison, the neuromorphic chip from IBM known as True North comprises one million artificial neurons, each with 256 synapses [7]. With that in mind, it seems likely that current or near-term technology for artificial neural networks (ANNs) may provide a platform for emulating at least a primitive version of animal consciousness, if the proper architecture is provided.

There have been few detailed proposals to develop a computer architecture for consciousness. I proposed a US patent [8] on the topic, parts of which are summarized in the following section.

II. ARCHITECTURES OF CONSCIOUS MACHINES

Consciousness does not emerge spontaneously from large neural networks. On the contrary, it requires a particular architecture with several distinct components. A possible architecture is shown in the block diagrams in Figs. 1 and 2.

The consciousness computing module constitutes a dynamic VR environment, comprising the self, objects, and other agents. The unconscious module comprises a computational engine that accepts input from an external physical environment and generates control outputs to the external environment, as well as generating the VR. The internal sense of consciousness is associated with the activation of the VR, which provides a coherent narrative of the self, interacting with other components. So the unconscious module generates the VR, but the conscious module experiences it.

The VR comprises a simplified representation of the external environment, focused on aspects that are relevant to the self. In effect, the external environment is deconstructed or partitioned into objects, and the objects are combined to reconstruct the VR. An object represents an element in the external environment that maintains integrity in space and time, and is linked to memory of similar elements in previous environments (external and virtual).

While some objects are passive, a special type of object is active, i.e., it is an agent. An agent can alter the environment, moving itself or other objects. The neural network in the unconscious module is configured to use pattern matching to recognize that the motion of the agent is correlated with the motion of the object. In natural environments, such correlation is generally due to causality, and the unconscious module further generates a model of the agent moving continuously in the VR, causing the object to move. Such a model not only describes past actions in the environment, it also enables predictions of future actions by extrapolation.

A special type of agent is the self, the activity of which is correlated with control outputs and local sensor inputs. The self constitutes the image of the self in the VR together with filtered versions of these inputs and outputs. Recognition of the self constitutes matching the present self with that in recent and earlier memories of the self. Other agents are distinguished by *not* being correlated with control outputs and local sensor inputs, i.e., they act independently. The model of the self may also include a set of positive and negative associations (based on pre-programming or learned from prior experiences), that may be analogous to emotions in biological consciousness.

A key aspect of temporal pattern recognition is the role of continuity of actions in time and space, and this requires a reference clock. While perception of time appears to be continuous, it may actually be discrete, with a small time step that is smaller than changes in the environment. Conventional video technologies are all based on discrete time, but appear continuous. Further, memories are stored with time references, but what is stored is not raw sensor data, but rather the narrative of the VR at a given time, so that what is recalled is the conscious experience.

Fig. 1. Block diagram of a conscious machine with virtual reality generation.

Fig. 2. Block diagram of architecture of memories and clock for conscious machine.

Another critical part of the prediction engine is the ability to imagine two alternative futures depending on the actions of the self. This enables conscious decisions, based on evaluations of goals and prior history. This is the sort of decision that is sometimes attributed to “free will” in humans, but is really deterministic.

As with other neural network systems, the initial design and training of a conscious machine are essential. The first step is to design an initial model for the machine moving in its environment. What are the important sensors and actuators needed, and what are the relevant timescales? Second, the neural network must be configured to evaluate sequences in time and space, and designed to use temporal pattern recognition to identify agents in the environment. The most important agent is the “self”, which is correlated with the sensors and actuators.

Such a machine can be trained initially using a simplified external environment. In some cases, the initial training environment might be a computerized VR system. Finally, the machine can be transferred to the real world, where learning will continue. Throughout the training and learning, external monitoring of the internal VR environment would permit confirmation of successful training.

The presence of a conscious VR module does not preclude also incorporating unconscious perception and control via neural circuits that do not interact with consciousness. A hybrid approach would assign to the conscious module only those aspects that cannot be achieved efficiently by more conventional ruled-based AI in the unconscious module.

III. APPLICATIONS OF CONSCIOUS MACHINES

We are accustomed to thinking of consciousness as a superior form of brain activity, but it might be better to think of a conscious machine as a specialized niche coprocessor that is used occasionally, when other processors are incapable of handling certain difficult problems. But such difficult problems are not uncommon, otherwise biological evolution would not have developed and refined the architecture. For example, it may be helpful to consider that prior to the machine age, the major mode of transportation was a biological conscious machine – the horse. Following an initial training period for learning commands, a horse can convey a rider or carriage driver at a controlled variable speed on or off a path, while avoiding actions that might tend to lead to injury to the horse and driver.

So an autonomous or semi-autonomous vehicle would be a natural near-term application for a conscious machine. For example, consider a car driving along a road, with a large puddle in the road. Is this a shallow puddle that one should simply drive through, or a deep pothole that should be avoided? If the answer is obvious, this need not rise to the level of consciousness. But if it requires a more complex assessment of weather conditions and the local status of road repair, the decision needs to be made at a higher level. There is a danger in this – such high level decisions based on incomplete data can be slow, at a time when a rapid decision is necessary. And the decision can still turn out to be the wrong one. However, any system that significantly increases the likelihood of improved decisions has substantial value, for a self-driving car or other autonomous system.

Such a vehicle could travel on roads, off-road, on the surface of water, underwater, in the air, or in space, or in some combination of these modalities. It could be used to transport one or more people or freight between locations. Alternatively, it could record information about the environment that it passes through, and/or wirelessly transmit such information to a remote location. It could be mounted on a small airborne drone, for example, which could conduct reconnaissance and provide for automated refueling/recharging.

In yet another alternative, the conscious machine may be incorporated into a personal digital assistant that could accompany an individual. This might help to guide someone through the environment, or serve to aid the memory, or otherwise anticipate personal needs. This would be particularly valuable for individuals who may be handicapped or unable to navigate without assistance. Consider the digital equivalent of a seeing-eye dog, for example. The conscious module may provide more rapid evaluation of incomplete data in environments that are unfamiliar to the individual.

A conscious machine might present another issue that is normally restricted to brains: mental illness. For example, malfunctioning of the VR generator may present distorted perceptions or narratives, which might be analogous to paranoia or schizophrenia. Furthermore, memory activation thresholds that are too high or too low might cause hyperactivity, depression, or obsessive behavior. For this reason, it might be important to monitor a conscious machine for abnormal behavior or thoughts. From another point of view, a malfunctioning conscious machine might even represent a model for simulating human mental behavior.

In still another application, the conscious machine may control an autonomous or semi-autonomous robot, which may be designed to operate in a variety of complex natural and artificial environments. Such a robot, with proper initial training and continued learning, may exhibit a much greater flexibility for adaptation and autonomy in response to unpredictable events. Such a robot may be configured to interact with humans in real time via either electronic or natural language communication.

The conscious machine may also be configured to communicate with other similar or complementary conscious machines. In this way, one can envision a number of autonomous machines acting in concert to achieve a goal, without requiring direct real-time control by humans. Or in some other cases, the conscious machines might recognize other such machines as competitors, and work to oppose each other's actions.

Similarly, a conscious machine might recognize some humans in the environment as collaborators, and others as competitors. In order to avoid harming humans in pursuit of a goal, these machines might be programmed with a code of conduct similar to Asimov's Three Laws of Robotics [9]. Of course, one can also envision applications in law enforcement or even the military, in which other rules might apply.

Alternatively, conscious machines may act as agents in a fully virtual environment, either within a training simulation or within a game in competition or collaboration with humans, who might also be present as avatars within the same virtual reality. Such a conscious machine might be expected to offer a more sophisticated set of responses in diverse situations, depending for example on prior memories of interactions with the same competitor.

Rather than traveling in real space or in a simulated 3D environment, the conscious machine could travel in an information environment, for example in a cloud computing environment or the Internet. This could provide a more intelligent engine for searching a large distributed database, or for controlling a distributed sensor network, or for evaluating data from many nodes of the Internet of Things.

One type of information application where innovation is needed is cybersecurity, in both defensive and offensive modes. There has been an exponential expansion of systems connected to the internet, with remote access enabled. This opens up these systems to all kinds of hacking, including by fast and persistent automated attacks. It is virtually impossible to design a passive system that prevents all such hacks.

What is needed are adaptable, dynamic digital agents that can defend systems locally, and are faster than the automated attackers. Furthermore, other digital agents can traverse the internet and recognize and disable such attackers before they attack, in numbers and with speed to overwhelm the attackers. This may require the evolution of conscious machines such as those discussed here. Of course, if the attackers adopt similar technologies, the conscious defenders may need to adapt quickly.

IV. IMPLEMENTING A CONSCIOUS MACHINE

This paper has presented the concept of a conscious machine based on neural networks and self-recognition, with general discussions of architectures and potential applications. But how does one even get started with the simplest possible prototype of a conscious machine, and how does one demonstrate its consciousness?

Probably the best way to start would be with a simple digital model of an environment, with a limited number of objects and agents, and have a prototype system navigate this environment. Of course, simple model digital environments already exist – they are computer video games. The use of computer games for training AI systems is well established [10]. One strategy would be to train a prototype conscious machine within the world of a particular video game, with and without human players. The prototype system could be monitored to determine whether it properly identifies objects, agents, and itself within the game environment, from their actions rather than from their

static appearances. In this way, one could track how well the system learns to perform in this environment, and how its performance compares with human players and other digital agents.

Deep neural networks may be implemented in a variety of technologies, including digital and analog circuits, biological neurons, CMOS transistors, superconducting Josephson junctions, memristors, phase-change memories, and resistive RAMs, based on pulse or voltage-level logic. They may be implemented in a variety of circuit architectures, including not only conventional processors, but also special-purpose processors such as GPUs, TPUs, and FPGAs. The deep neural networks comprising a conscious machine may in principle be implemented on any device technology and architecture that can support temporal pattern recognition, and to identify agents of change in the virtual environment created by the set of sensory inputs.

This temporal pattern recognition represents a repeated process activated by a periodic clock which establishes a time base. The frequency of this updating can be relatively slow if the systems deals with slow changes in typical human environments; for comparison, the alpha rhythm in human brains is of order 10 Hz. But electronic systems are capable of much faster operation, so that faster updating might be appropriate for application to a rapidly moving autonomous vehicle, for example. Even faster clocks ($\sim$ MHz) would be appropriate for a conscious system navigating an information environment such as a database, or defending a computer operating system from external hacking.

In order to be able to act with a sufficient level of sophistication in a complex environment, the neural network of a conscious machine must comprise a large number of neurons with an even larger number of interconnections. For example, the network might have at least millions of neurons, with at least billions of interconnections. The strengths of the interconnections representing memories and associations must preferably be non-volatile over long periods of time, or at least fully backed up in case of power failure. The memories must be capable of repeated adjustment and readout, with high reliability and very low rate of failure. The system should be able to be temporarily turned off, so that repairs or upgrades may be implemented, and then turned back on in a way that remembers past performance.

Since a conscious machine should remember all of its past memories, the total memory requirement may be quite large. However, it may be appropriate to partition the memory into an active memory and a long-term memory. The active memory should hold the VR for the present, the recent past, and alternative future plans. The long-term memory should hold the VR for the more distant past, and should be accessible either by time-stamp, or by content. Both memories may be structured as neural nets, but the active memory should be faster, with write as well as read capabilities.

So will conscious agents be present in AI in twenty years? I predict that they may be widespread, but not the humanoid robots of science fiction. Instead, they will be intelligent, flexible, autonomous machines that can be trained to traverse our transportation and information networks, assisting and protecting humans, rather than threatening us. These may soon be possible using near-term neural network technology.

REFERENCES

- [1] R. Manzotti, "Machine Consciousness: A Modern Approach", *Natural Intelligence: The INNS Magazine*, vol. 2, pp. 7-18, 2013.
- [2] J. A. Reggia, "The Rise of Machine Consciousness: Studying Consciousness with Computational Models", *Neural Networks*, vol. 44, pp. 112-131, 2013.
- [3] Wikipedia, "Artificial Consciousness", retrieved August 2021. https://en.wikipedia.org/wiki/Artificial_consciousness
- [4] A.M. Kadin, "No Ghost in the Machine". Submitted to Foundational Questions Institute Essay Contest, 2017. Available online at <https://fqxi.org/community/forum/topic/2821>. Also published in *Wandering Towards a Goal*, ed. A. Aguirre, et al., Springer, 2018.
- [5] D. Kahneman, *Thinking, Fast and Slow*, Farrar Strauss and Giroux, 2011. See also https://en.wikipedia.org/wiki/Thinking_Fast_and_Slow.
- [6] Wikipedia, "List of animals by number of neurons", Aug. 2021. https://en.wikipedia.org/wiki/List_of_animals_by_number_of_neurons
- [7] Wikipedia, "Cognitive Computer", retrieved August 2021. https://en.wikipedia.org/wiki/Cognitive_computer
- [8] A.M. Kadin, "System and Method for Conscious Machines," US Patent 11,189,483, issued 9/14/2021. <https://patents.google.com/patent/US1119483B2/en>
- [9] Wikipedia, "Three Laws of Robotics", retrieved August 2021. https://en.wikipedia.org/wiki/Three_Laws_of_Robotics
- [10] N. Justesen, et al., "Deep Learning for Video Game Playing", *IEEE Transactions on Games*, vol. 12, no. 1, March 2020.