Source: PsyArXiv Preprint, Dec. 2023, DOI: https://doi.org/10.31234/osf.io/gbzd6

The conductor model of consciousness, our neuromorphic twins, and the human-AI deal

Federico Benitez*, Cyriel Pennartz+, Walter Senn*

*Institute of Physiology and Center of Artificial Intelligence in Medicine (CAIM), Faculty of Medicine, University of Bern
+Swammerdam Institute for Life Sciences, Center for Neuroscience, Faculty of Science, University of Amsterdam

Abstract

Critics of Artificial Intelligence posit that artificial agents cannot achieve consciousness even in principle, because they lack certain necessary conditions for consciousness present in biological agents. Here we highlight arguments from a neuroscientific and neuromorphic engineering perspective as to why such a strict denial of consciousness in artificial agents is not compelling. We argue that the differences between biological and artificial brains are not fundamental and are vanishing with progress in neuromorphic architecture designs mimicking the human blueprint. To characterise this blueprint, we propose the conductor model of consciousness (CMoC) that builds on neuronal implementations of an external and internal world model while gating and labelling information flows. An extended Turing test (eTT) lists criteria on how to separate the information flow for learning an internal world model, both for biological and artificial agents. While the classic Turing test only assesses external observables (i.e., behaviour), the eTT also evaluates internal variables of artificial brains and tests for the presence of neuronal circuitries necessary to act on representations of the self, the internal and the external world, and potentially, some neural correlates of consciousness. Finally, we address ethical issues for the design of such artificial agents, formulated as an alignment dilemma: if artificial agents share aspects of consciousness, while they (partially) overtake human intelligence, how can humans justify their own rights against growing claims of their artificial counterpart? We suggest a tentative human-AI deal according to which artificial agents are designed not to suffer negative affective states but in exchange are not granted equal rights to humans.

Keywords: Consciousness, Artificial Intelligence, Turing Test, AI Ethics

1.1 Introduction

Artificial intelligence (AI) has transformed from a science fiction concept to a present-day reality with the potential of furthering human prosperity. Generative large language models, such as GPT-4, or generative image models, such a Dall-E, give a glimpse on the cognitive power that AI may yet achieve in the future. AI may help us tackling vital problems, such as fighting climate change through the development of renewable energy technologies and the optimization of resource use (Cowls et al., 2021).

As AI research is fuelled by such successes and prospects, notions of emerging consciousness in artificial systems elicit growing popularity among the public and scientists. While a former Google engineer asserted that a current artificial intelligence model, LaMDA (Language Model for Dialogue Application), is already conscious and capable of suffering (Tiku, 2022; see also Newitz, 2022; Weaver, 2022), most artificial intelligence researchers firmly deny this claim, positing that we are far from achieving the creation of conscious artificial agents. Nonetheless, a clear majority of them do not rule out the possibility of artificial consciousness and go even as far as positing "sparks of general artificial intelligence" in GPT-4 (Bubeck et al. 2023). In philosophy, however, this question of principle remains a subject of debate, with some scholars arguing for the multiple realizability of consciousness and others denying the very possibility of artificial consciousness (see e.g., Edelman & Tononi, 2000; Fuchs, 2018, 2021, 2022; Van Gulick, 2022).

In this work we take the perspective of computational neuroscience to address some scientific, technical, and ethical aspects of this issue. As a first contribution, we elaborate on a materialistic intuition in favour of artificial consciousness. The starting argument is that if we could replace each neuron within a human brain by an artificial counterpart, and if each of these neurons and their connections would work in a way that is similar enough to biological neurons, the resulting artificial brain should be capable of consciousness (e.g., Chalmers, 1995). Here we add weight to these arguments by connecting them to recent developments in neuromorphic engineering,

claiming that some basic neuronal architectures may be realized in an artificial substrate. The neuromorphic ethos differs from other methods by taking direct inspiration from biological neurons to design networks of analogue neurons that communicate using spikes of activity, and that can learn by synaptic plasticity just as biological neurons do.

This neuromorphic blueprint for artificial consciousness allows us—as a second contribution—to propose an extension of the well-known (but too restricted) Turing test for artificial intelligence and consciousness. To go beyond previous proposals, we introduce a specific model of the neural and architectural correlates of consciousness in human (and mammal) brains, which we call the conductor model of consciousness (CMoC). The overall idea is to add to the behavioural component of the Turing test an additional architectural layer: by comparing the connectivity map of the neuromorphic neurons of an artificial agent to that of humans, we can add weight to the possibility of the agent having human-like consciousness.

The possibility of machines becoming conscious brings about serious ethical considerations. If conscious artificial agents develop cognitive abilities that rival or even surpass those of humans, it becomes inevitable to consider granting them legal and political rights (e.g., Miller, 2015; Gunkel, 2018; Persaud, 2021; De Graaf, 2022). Doing so may result in instances where the rights of an artificial agent conflict with those of a human being. Such situations will pose complex ethical dilemmas, particularly when it becomes necessary to consider the potential prioritization of AI rights over those of a human. Additionally, if we equip machines with some form of consciousness, it becomes unavoidable to consider that such agents will potentially be able to experience pain and suffering (e.g., Agarwal & Edelman, 2020). Such AI suffering would give ground to moral conflicts (Metzinger, 2021).

The third contribution of this work is to offer a new perspective on these dilemmas, by considering the possible down-regulation or prevention of negative affective states (such as pain) in artificial agents. As we argue, this ensures that creating possibly sentient artificial agents will not result in an unbounded increase in global suffering, and that there is no one-to-one competition between the moral rights of humans and the machines we create.

Here is how we proceed in more detail: we begin by briefly explaining how several specific key terms are used in this paper (Section 1.2). We then continue by a short review of common criticisms of artificial consciousness, that focusses on three main aspects: substrate, embodiment, and evolution (Section 2.1). In a next step, we elaborate on recent advances in neuromorphic engineering and computational neuroscience and suggest that many aspects of artificial consciousness criticism do not hold when these advances are considered (Section 2.2). Further, we introduce the idea of a neuromorphic twin, which aims at updating classical thought experiments on artificial consciousness (Section 2.3). Having argued that neural correlates for some forms of consciousness in artificial agents are within technical reach, we elaborate the design of an extended Turing test (eTT) guided by a “conductor model of consciousness” (CMoC). The CMoC specifies certain brain architectures that in humans are likely implicated in a neural correlate of consciousness, and the eTT suggests testing whether these structures are present (Section 3). In Section 4.1 and 4.2, we discuss possible ethical dilemmas surrounding AI consciousness, and how they can be addressed by tweaking the design of conscious artificial agents. Section 4.3 discusses how this can be considered a beneficial “deal” for future conscious machines. We discuss these ethical implications in Section 5 and present our conclusions in Section 6. See Fig. 1 for a schematic overview of this work.

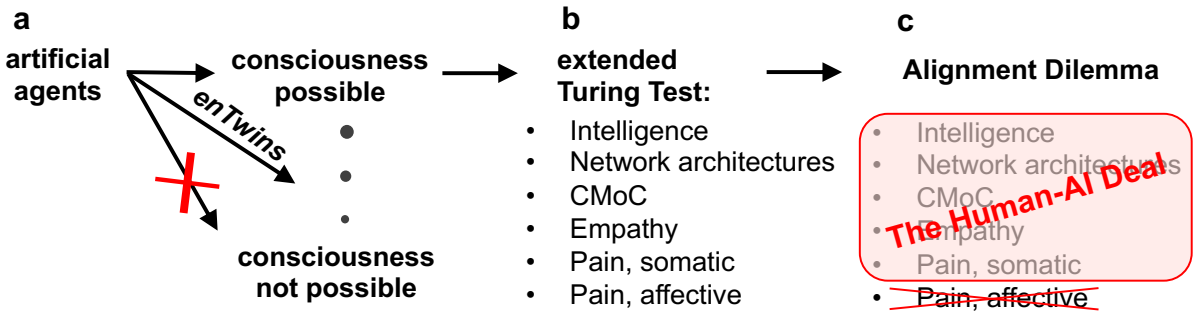

*Fig. 1: A rough schematics of the contributions of this work: criteria for conscious artificial agents and how to restrict them. (**a**) The example of the evolving neuromorphic Twin (enTwin) shows the difficulties to exclude any form of consciousness. (**b**) To test for consciousness, we suggest an extended Turing Test that requires the identification of specific neuronal architectures described by the conductor model of consciousness (CMoC) and neuronal correlates of perception (e.g., of somatic and affective components of pain). (**c**) Should a putative artificial consciousness share all features of human consciousness? The stronger the alignment, the better the expected mutual understanding is, but also the more competition between the artificial and human species is expected. This Alignment Dilemma could be approached by what we introduce as the Human-AI Deal: it relieves the conscious artificial agents from the affective component of pain, but gives humans the priority before the law, allowing agents to negotiate for more rights with benevolent behaviour.*

1.2 Phenomenal consciousness

The terms “conscious” and “consciousness” are vague and often used with vastly different meanings (Van Gulick, 2022). Therefore, we briefly state how these terms are used in the present work. This article is concerned with *phenomenal consciousness*. Phenomenal consciousness is considered a form of *state*-consciousness—i.e., a property attributed to certain mental states. If it “feels like something” (Nagel, 1974) to be in certain mental state, this state is considered as being phenomenally conscious (Carruthers, 2001). Therefore, phenomenal consciousness is often described as the subjective aspect of consciousness that involves experience (Chalmers, 2006; Block, 1995).

We consider an internal state of a system as a mental state if it is intentional with respect to a representation (of a certain state of affairs) that is available in that system. That is, if an internal state "is about”, or “refers to” an internal representation, it is a mental state. If it additionally feels like something to be in that mental state, it is considered as phenomenally conscious.

Further, we distinguish between *sensory* and *affective* aspects of phenomenal consciousness. The sensory aspect refers to the subjective experience of sensory stimuli, such as sight, sound, touch, taste, and smell. E.g., what does it feel like to be in a mental state that refers to the representation of a red percept? The affective aspect, on the other hand, refers to the subjective experience of emotions, feelings, and moods. E.g., what does the mental state that refers to the representation of pain feel like?

An agent is understood as a system that acts with an intention (imposed internally or externally) upon its environment, e.g., a bee that collects honey, a soldier carrying out an order, a robot that performs a task in a car factory, or a personalized large language model that suggests email replies in your spirit and style. However, a flood that damages a road, atoms that form a bonding, or a black hole that swallows a star are not considered agents. In other words, we understand agents as systems that exhibit goal-directed behaviour. That is, they have a basic understanding of what their goals are, and by which actions they can be reached (Balleine and Dickinson, 1998; Pennartz, 2018).

Finally, we consider an agent as being conscious if it is phenomenally conscious as described above. That is, if an agent has mental states that refer to internal representations and if being in such a state feels like something, that agent is considered conscious. Humans and many animals will pass these criteria, and we typically agree on assigning them consciousness. The critical criterium of “feels like something” is accepted to be satisfied if it is reported by another human. Our acceptance originates from our own experience of “feels like something” and the fact that there are overwhelming testable consistencies between ours and the reporting human’s mental

states. Most people (including us) are currently denying that the mental state of an artificial agent "feels like something" for them because, so far, we have not enough testable consistencies (or because of inconsistencies) between ours and their mental states. We are aware that our use and interpretation of these terms may not be uncontroversial; however, we hope that clarifying the intended meaning helps to understand the arguments put forward in this paper.

2. Artificial consciousness at the dawn of the neuromorphic era

2.1 Re-evaluating criticisms of artificial consciousness

Sceptics with respect to AI consciousness point out that the analogies between digital computers and human brains have many breaking points. Among the things that distinguish machines from (healthy) brains the following appear to be the most relevant: (i) the lack of embodiment, (ii) the lack of a centralised "I", (iii) the lack of evolutionary pressures and feedbacks, (iv) being based on different physical substrates that behave differently, (v) the use of the von Neumann architecture, and (vi) the digital representation of information. Authors such as Edelman & Tononi (2000) and Fuchs (2018, 2021) have argued that these qualitative differences between the inner workings of computers and brains speak against the very possibility of emergence of artificial consciousness.

At first glance, many of these criticisms seem reasonable. If we consider embodiment as a relevant aspect of conscious experience, it is the case that most computers do not have ways to affect their environment to have a better grasp of it, both in the sense of perceiving physical space or of literally grasping physical objects. Although robots are reaching the degree of development where this point becomes moot, a large part of the discussion on AI is happening at the level of AI *software*, where the criticism seems appropriate. Such a lack of embodiment constrains the possibilities to develop self-awareness, as there is no clear separation between an "I" and the world, and the interactions between the machine and the world are ultimately initiated and controlled by the machine users (humans). With respect to architecture, in stark contrast to how the brain works, the usual von Neumann architecture divides information processing between a central processing unit and an external memory, a distinction that may preclude synergies between memory and processing. The difference between "hardware" and "software" is much less clear-cut in the case of biological brains, where brain activity is known to change the strength of the connections between neurons. Likewise, brains do not appear to act like digital machines that run programs sequentially.

The lack of a centralized "I" for artificial agents motivates some of the ideas that we present further down. The basic issue is that, even for contemporary robots which have an inner representation of their state within the environment, it is difficult to argue for the presence of a sense of self analogous to the one that humans and some other animals have. Even if there is a higher order module overviewing the state of the system, there is no warranty that this module will have a notion of "self", unless we put it there deliberately. This has led critics of AI to point out what seems a vicious regress according to which, to have a sense of self, we need a subset of the artificial brain to already possess such a sense (Zahavi, 1999). In neuroscience, it is increasingly clear that the "I" is constructed from multiple, intertwined notions of the self, including body ownership, use of efference copy to distinguish self-induced versus externally caused sensory changes, multisensory integration, agency, episodic life history and social identity (see e.g., Metzinger, 2004; Pennartz, 2015).

If it were only about the "I", we *could* replicate what natural selection processes brought about in humans (and probably behaviourally evolved animals). For example, we could equip future machines with a "self" module, that overlooks and to a degree controls its own functioning, while having a pre-wired notion of it being itself. Below (Section3), we relate this to a "conductor" module that is arguably available in human brains and might also be implemented in artificial machines. We argue that such a module (or set of modules) has appeared at some point during the phylogenetic, as well as ontogenetic development of our brains. Reengineering these fruits of evolution would turn the problem of a centralized "I" into a technological, but not a principled problem. How the artificial "I" will "feel like" for the agent remains up for debate.

This relates to claims that evolution is a prerequisite for the development of consciousness. Our brains and our consciousness are the result of millions of years of evolution—a complex process featuring a plenitude of

feedback loops of interactions between our ancestors and their environment. Artificial agents do not undergo such processes, but nothing impedes us from designing these systems *as if they were* the result of evolution. We could create these systems as if they had an evolutionary history. This retroactively embedded history could in fact be our own history as a species that evolved consciousness, including the embodied traces of evolutionary processes—the most adapted kinds of limbs, for example, or the "innate" sense of self or centralized "I".

2.2 Architecture and the substrate problem

Criticism regarding architecture and substrate can be addressed by turning to recent advances in neuromorphic hardware. Neuromorphic engineering aims to build hardware that mimics the brain to harness its extreme parallelism and asynchronous nature for power efficiency and computing speed (Mead, 1990; Indiveri et al., 2011; Schuman et al., 2017; Roy, Jaiswal, and Panda, 2019). This multidisciplinary area of research takes direct inspiration from the structure and operations of the brain and its basic units, to develop new kinds of hardware. The implementation of neuromorphic computing on the hardware level can be realized by a wide diversity of substrates, such as transistors, memristors, spintronic memories, threshold switches, among others.

So far, work on neuromorphic designs has focussed on replicating the analogue nature of biological computation and in emulating the spike-based information exchange between neurons that occurs in the brain. Nowadays, neuromorphic chips are not fully analogue, but an increasing portion of their subcomponent are (see, e.g., Pehle et al., 2022), and the aim of fully analogue chips seems attainable. Additionally, there is a line of research on implementing this hardware on flexible arrays and flexible chips that can be implanted within biological tissues (Kang et al., 2021) and to be effectively scalable (Cointe et al., 2022). On top of this, a lot of effort has been invested in encoding learning and memory using the plasticity of the synaptic weights between different neurons, emulating biological brains. Interestingly, at least in the existing silicon-based neuromorphic hardware, these model neurons have the capacity to operate orders of magnitude faster than their biological counterparts (Billaudelle et al., 2020; Göltz et al., 2021), something that will be relevant in Section 5 below.

Thus, neuromorphic hardware offers compelling solutions to the traditional objections concerning substrate dependency. The next generation of flexible carbon neuromorphic substrates (Du et al., 2021; Zeng et al., 2021) could be moulded to emulate biological neurons to a degree that makes it very difficult to sustain any principled opposition to artificial neurons—and brains. While such artificial brains, or parts of them, in principle may reproduce the function of biological neurons, synapses and networks up to mental states and behaviour, the question whether they also reproduce the "feeling of being like something" or generally speaking the subjective qualities associated with conscious perception (the qualia), remains inaccessible from a third person's perspective.

In summary, recent techniques and developments have closed the door to most of the first-principles arguments against the possibility of artificial consciousness. Even though the classical von Neumann computer architecture is arguably inadequate for the task, this is no longer the only game in town. In what follows we develop these ideas in detail, by providing a more explicit blueprint for the construction of an artificial consciousness.

2.3 A co-evolving neuromorphic twin

A very popular way to explore different scenarios for AI and consciousness is by means of thought experiments. Thought experiments have introduced us to sometimes arcane notions, such as philosophical "zombies" (Kirk, 2021), Chinese rooms (Searle, 1980), and Mary's lockdown room (Jackson, 1998), and various aspects must be considered in designing a good thought experiment, see e.g., Brown, 1995. Our purpose is not to present a novel thought experiment. We instead *substantiate* existing thought experiments to ground the feasibility of systems that closely emulate the human brain in many aspects that are relevant for consciousness, and that answer the criticisms to artificial consciousness sketched in the previous section. These experiments suggest that consciousness could be realized in various substrates, provided the functionality of its constituent parts, such as neurons, is preserved—most famously, simply replacing each biological neuron in a brain with an artificial counterpart (e.g., see Chalmers, 1995; Searle 1980).

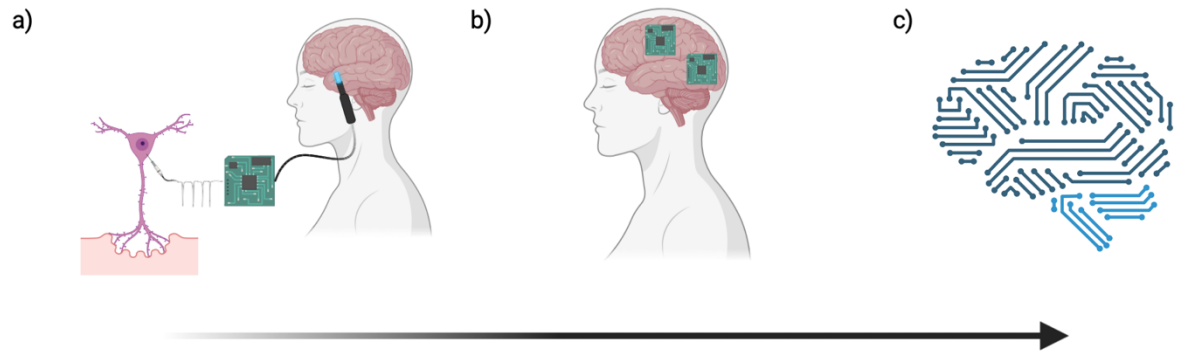

*Figure 2: The evolving neuromorphic twin (enTwin) thought experiment. In **(a)** a neuromorphic chip, able to communicate with biological neurons, is used to help human patients achieve normal sensorimotor function. The neuromorphic chip can be implanted inside the body of a human and learn to adjust its synapses in the same way as biological neurons do. **(b)** With time, neuromorphic chips could be trained to achieve higher order functions, including functions that pertain to conscious experiences, such as the perception of sensations, and the associated feelings that they invoke. They could be integrated into the brain by using of soft bioelectronic interfaces. **(c)** By training many such systems with the help of many different humans, a fully artificial brain can be constructed and embedded within an artificial body. If every piece of such a brain can collaborate in the genesis of conscious experience of human patients, then there is no reason why the fully artificial enTwin would not also be able to develop consciousness.*

The classical neural replacement scenario of Chalmers (1995) has been deemed implausible by authors such as Fuchs (2018, 2021, 2022) because of the substrate problem. If computer chips are radically different from neurons, then the very premise of them supplanting neurons on a one-by-one basis in a human brain is flawed, because not even the first neuron can be faithfully replaced. To overcome these criticisms, we propose a neuromorphic version of the scenario, grounded in current neuroscience, and trying to be as concrete and detailed as possible. In other words, we present a revised version of the thought experiment, viewed through the lens of neuromorphic engineering. We call it the *evolving neuromorphic twin* (enTwin), a specific implementation which is realistic considering present day technology and complements more abstract philosophical insights about consciousness.

Assume a human baby is born with a cerebral ataxia syndrome that is linked to cortical degeneration (Conrad et al., 2023), resulting in motor disabilities including articulations and speech. Assume it is possible to help the child with an evolving neuromorphic twin (see Figure 2). The enTwin is implemented in soft bioelectronic interfaces that can be implanted in human bodies (Mariello, 2022), and even in human brains (Yuk, 2020). The enTwin is fed by tactile and proprioceptive information at the extremities, and by an electrography of the speech muscles. To prospectively assist speech formation, it is also supplied by visual and auditory information through latest-generation smart glasses and active ear plugs (Bartolozzi et al. 2022). For motor and speech assistance it is coupled with muscle stimulation devices. The chip includes a pretrained large language model (Schramowski et al., 2022) that is personalized to mimic in real-time the child's intentions to articulate the contents at each stage during the development. The hypothetical chip is built on flexible neuromorphic arrays with learnable synaptic connectivity and a neuromorphic architecture as outlined below. Blood sugar is measured to emulate internal mental states, and this modulates the energy supply of the chip, which itself is implemented using neuromorphic technology.

The chip interprets the sensory information and the host's internal state online, and with this drives the language module with a functionality comparable to LaMDA or GPT-4, together with various motor modules. The modules learn to decipher, recreate and represent the intended articulation and motor activity of the growing individual and are guiding and supporting them in improving both articulation and motor execution. For performance and survivability reasons, the enTwin could also try to predict and recreate the (representation of) feelings of its host (Gershman, 2019), interfaced with the corresponding brain regions. The representation of the postulated subject's feelings in the neuromorphic hardware offer an analogue of: (i) the amygdala, anterior cingulum, orbitofrontal cortex, insular cortex, central thalamus among other regions, to process the various components of feelings such as pain, (ii) the sensory and motor cortices to represent the sensorimotor transforms, and (iii) the Wernicke and Broca's area to represent language understanding and articulation, (iv) thalamic and brainstem kernels to represent different wakefulness states (Pal, 2016; Gent, 2018). While predicting actions in response to input from others, mirror neurons would develop which could form a possible basis for empathy

(Lamme 2015). Consequently, an enTwin would mimic the one example we know where consciousness developed (humans and mammals

Once this integrated enTwin is working within a host, its information could be copied to a database, helping to design an enTwin embedded within an artificial body—a neuromorphic robot. This robot would be an embodied entity, with components that are (externally) evolved. Thereby, "evolution" stands for many things simultaneously: evolution in the sense that its brain will be the result of co-evolution with human hosts, evolution in that we are copying the results of biological evolutionary history within both artificial brain and body, and of course evolution as the result of iterative technological improvements on things like sensors, limb articulations, materials, and so on.

The timeline for the development of such neuromorphic robots is unknown, as various uncertainties remain, including the ethical question how far medical aids should interfere with our organs, and specifically with the brain. Nonetheless, it would be hard to argue against their feasibility, just by looking at the state of contemporary neuromorphic research. As such, enTwins flesh out many of the intuitions behind previous thought experiments about artificial consciousness.

3 The conductor model of consciousness (CMoC)

To judge the possibility of a consciousness counterpart in our enTwin, to infer possible criteria for a neural correlate of consciousness, and eventually infer ethical guidelines, it is helpful to focus on some key ingredients our enTwin is likely composed of. As opposed to existing neuronal theories of consciousness (for reviews see Seth & Bayne, 2022, or Storm et al., 2023), the conductor model we propose focusses on network architectures and their functional interpretations that, as we argue, are likely involved in producing phenomenal consciousness.

Given the reality monitoring areas in the brain that judge whether activity in sensory areas is generated from inside or originating from outside (Simons et al., 2017), we argue that the brain contains the crucial ingredients to implement a form of Generative Adversarial Networks (GANs, Goodfellow et al., 2014). GANs have proven to be cornerstones of powerful network architectures for image recognition, language processing, and translations of image to language (Aggarwal, 2021). Certainly, generative networks are implicated in mental imagery, and discriminative networks must exist that tell apart imagined sensory activity from externally induced sensory activity. These networks need to be trained, and it is reasonable to assume similar plasticity mechanisms being involved as in the technical version of GANs.

GANs include separated networks, starting with a generative network G that internally generates fake sensory information, an encoding network E that interprets sensory activity (regardless of being triggered externally or generated internally), and a discriminative network D that judges whether a particular sensory activity is produced internally or externally (Fig. 3). In addition, we postulate a conductor network that orchestrates the information flow. Based on the feedback from the discriminator network (that may reveal the fake nature of the sensory representation), the generative network can improve itself to produce a more realistic sensory activity. Additionally, when the sensory activity is internally produced by the generative network, the encoding network can learn to reproduce this activity. It has been postulated that some forms of GANs are implemented in the human brain (Gershman, 2019) and support creative dreaming during rapid eye movement sleep (REM) sleep (Deperrois, 2021). Furthermore, GANs are also involved in training networks to read emotions in human faces, and to generate corresponding facial expressions (Richardson, 2021).

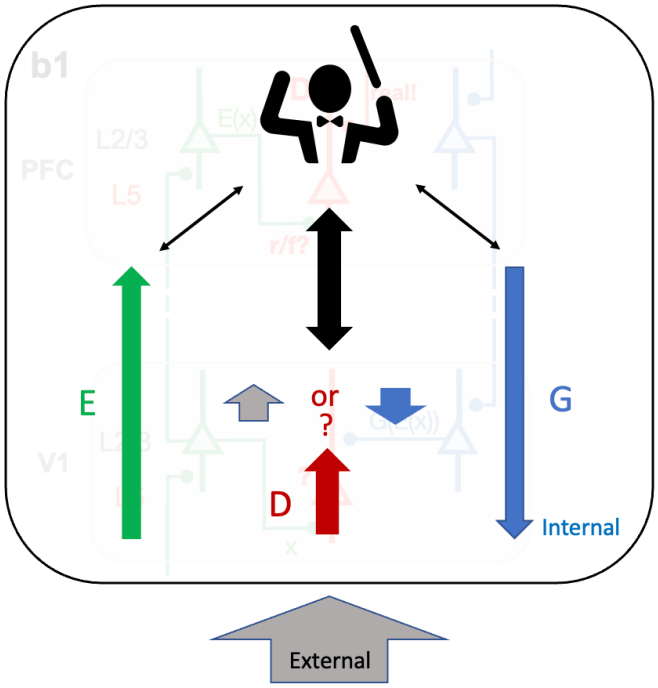

Figure 3: The conductor model of consciousness (CMoC): The implementation of elaborate forms of learning requires a network instance that organizes the flow of information to keep teacher and student signals apart. Possible ingredients for consciousness to evolve in our enTwin. ***(a)*** *An encoding (E) and generative (G) network, together with a discriminator network (D) that judges whether the sensory activity is originated from outside (External) or inside (Internal), just as in GANs. The faded background represents the neural circuitry. A conductor module selects the contents of the encoding and generative networks that matches and broadcasts this for further processing.*

The conductor model of consciousness (CMoC) emphasizes the orchestration of the information flow between encoding network, generative network, discriminator network, and their training (Fig. 3). Learning is about improving a behaviour, and the desired activity is implicitly or explicitly declared as activity to be reproduced. The conductor model makes the distinction between a teacher and student signal explicit by postulating a network instance that gates the information flow for teacher and student signals to adapt the student signal. This structure is also present in self-supervised learning, where the teacher is formed by other, more informed parts of the brain that "nudge" the student network (Urbanczik & Senn, 2014). Reality monitoring areas (Simons 2017) are part of the cortical GANs as suggested in Gershman (2019) and may form the teacher instance for the discriminator network. The postulate is that implementing powerful forms of self-supervised learning (such as GANs) in autonomously running networks requires a conductor submodule that is a precursor of a consciousness-enabling network.

GANs intrinsically require a meta-level conductor that orchestrates the information flow. The conductor implicitly tells whether the GAN is in the inference or learning mode, and provides the information used for learning whether the activity represented in some higher cortical state is generated from internal sources or external stimuli. Such a conductor must itself be implemented in a submodule of the brain, and it can act on a hierarchy of cortical representations. Architecturally, this role resembles the functionality of prefrontal and anterior cingulate areas (Simons 2017), but it may also be taken over by the gating mechanisms of cortico-thalamic loops via higher order thalamic kernels (Takahashi et al, 2020; White et al., 2023), as elaborated below. When acting on the visual stream, the conductor may tell "this activity represents a certain visual object and is generated from inside", for instance. When acting on more abstract object representations like our own identity, the conductor may tell "this activity represents myself and is generated from inside" (see Fig. 3). In comparison to other theories of consciousness, the conductor model can be seen as a combination of higher order theories and the global neuronal workspace theory (Dehaene et al., 2006; Seth & Bayne, 2022), embedded in a functional cognitive architecture, and brought down to a specific suggestion for a neuronal implementation (Figs 3 and 5, see also Deperrois et al., 2022). This is explained in the next two subsections.

3.1 An extended Turing test (eTT) for consciousness

What gives weight to the notion that an enTwin would be conscious is not only that it would behave like a human, but that each of its microscopic components behaves in a manner equivalent to biological neurons and networks of neurons being involved in cognitive processes and embedded in an independent agent that acts on the environment. To specify these components, we extend the classical Turing test, which has been shown to be inadequate to deal with the behaviour of modern AI (for a detailed discussion of the classical test, see French, 2000). For example, even though there is ample consensus that LaMDA or ChatGPT are not conscious agents, their follow-up versions will most probably be able to pass the classical Turing test. As in the original Turing test, we are putting forward a functional approach to discern the presence of consciousness within an agent, but additionally focus on the behaviour of the circuitry that make up its "brain".

We call our proposal the extended Turing test (eTT): on top of analysing the behaviour of the agent, and to check if it responds to external queries in the same way as a conscious agent would do, we additionally impose criteria regarding the physical means by which this behaviour is generated. In particular, the test demands that at the microscopic level, the neural correlates of consciousness identified in animals must have some analogue in the artificial agent (see Fig. 4). The eTT examines the implementation of the artificial brain and checks whether functional circuits that we know support feelings and consciousness in the mammalian brain have their counterpart. Consequently, this is a more stringent test than the classical Turing test and relates to ideas of neurorepresentationalism on consciousness (see Pennartz et al., 2019; Pennartz, 2015, for similar ideas).

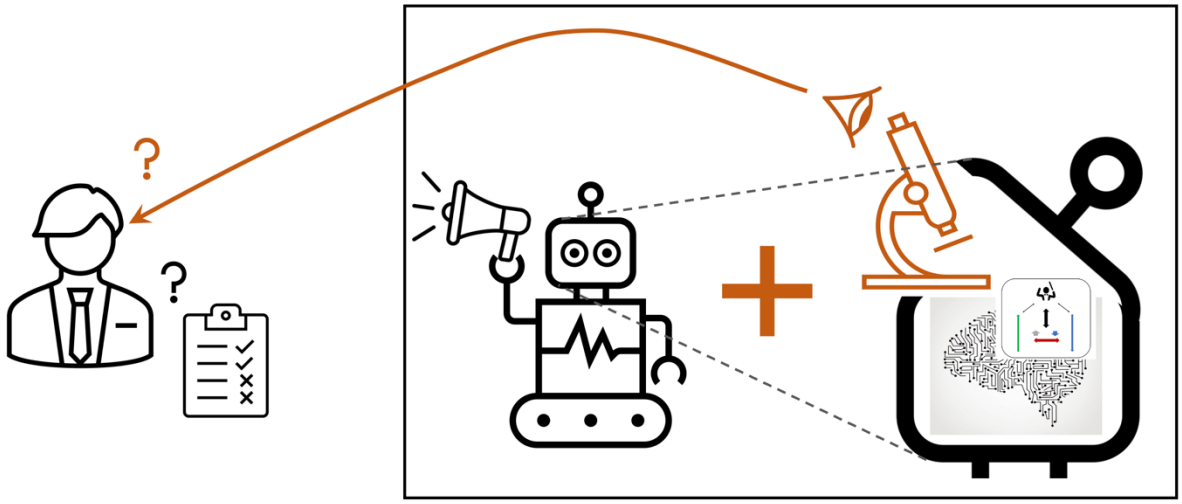

Figure 4: The extended Turing test (eTT). A list of criteria to be satisfied is indicative for the presence of some form of consciousness. The list extends the items of the classical Turing test for intelligence. It requires the observer to enter the "Chinese room", open the box and identify the postulated neuronal circuits for consciousness. That is, on top of the usual behavioural Turing test, where we examine the macroscopic behaviour and responses of an agent to our inquiries, we propose to add a second "microscopic" layer. The idea is to examine the explicit architecture of the neuromorphic neuronal network to check for neural circuits that we believe makes consciousness possible in humans.

Notice that the eTT could be considered *too* stringent a test. Strictly speaking, the eTT criteria may not be necessary. One could argue that the eTT could miss non-human forms of consciousness that are implemented in a fundamentally different way. Consequently, if the eTT-related circuits cannot be identified in a neuromorphic agent, this would only indicate the absence of human-like consciousness but not necessarily consciousness per se.

The eTT may be organized as a layered list of requirements. At the basal or zeroth level we have the behavioural criteria of the classic Turing test. On top, we add a series of requirements at the architectural / neuronal level that are motivated by our GAN-inspired CMoC (Figure 5):

(eTT-1) An *encoding network*, leading to abstract semantic representation of sensory input, together with a generative network, that recreates sensory activities out of semantic representations (green and blue in Fig. 5).

(eTT-2) A *discriminator network*, together with a *conductor module*, that orchestrates the learning in the encoding, generative and discriminator network, and labels the sensory activity into internally or externally generated (red in Fig. 5).

(eTT-3) A global *affective component* that represents internal needs and overriding signals such as "existential threat", integrated by the conductor and short-cutting the processing in other networks (Fig. 5c)

The proposal is to use criteria eTT-1 to -3 besides the classical Turing test to tell whether an agent may or may not be endowed with (human-like) phenomenal consciousness. Similar criteria have been suggested by various other authors. For instance, Damasio (2013) emphasizes the need for representing sensory inputs and imagined contents. Solms & Friston (2018, 2019) formulate similar criteria in the context of predictive coding and active inference, LeDoux & Brown (2017) in the context of emotions. Other works have coupled predictive coding networks to planning of complex, goal-directed behaviours (Pennartz et al. 2019). Dehaene et al. (2006, 2011) make the point that specific contents out of many sub-conscious contents in the brain are selected for a global workspace that provides consciousness.

The advantage the eTT has over previous proposals for extending the Turing test (French, 2020) consist in the availability of a specific, neuroscience-inspired model of the architectural requirements behind consciousness in the form of the CMoC. This model provides us with more explicit structural notions to approach the phenomenon of consciousness, and its ethical implications, as compared to other proposals.

3.2 The conductor organizes the inner world of autonomous agents

By introducing the CMoC and the eTT we can render the architectural-level implementation of phenomenal consciousness more precise. In line with other suggestions (e.g. Chalmers, 2013), we postulate that consciousness introduces its own quality of existence, that is neither physical, nor abstract, but uniquely experienced by the agent to whom the quality is assigned to. The conditions for this private quality of "consciousness" to appear in an agent are given according to the conductor model of consciousness (CMoC) by the following 3 requirements: A conscious agent, that is capable to sense and interact with the external world,

(CMoC-1) has a representation of the external world (the encoder network), a representation of an inner world (the generative network) and can act on both the external and internal world representations (e.g. via discriminative networks), beside acting on the external world itself (see eTT-1).

(CMoC-2) has a mechanism—the conductor—that allows to tell whether the agent acts on the internal or external world representation (specified by eTT-2)

(CMoC-3) is equipped with its own internal sense of self associated to the conductor, modulated by global affective components (specified by eTT-3).

Notice that CMoC-1 introduces a representation for the inner world *in addition* to the representation of the external world. One may argue that an internal world *is* a model of the external world. However, the internal world of a conscious agent is different from a mere internal representation of the external world. For example, body interoception can be seen as part of the internal world. But in the internal world of humans and other animals there is also a body representation, and this is not the same as the percept of the own body. What we posit here is that consciousness requires more structure within an internal world model than only serving as a model of the external world. It is the distinction between a world (internal or external) and a world model that yields an additional level of abstraction. The external world is captured by a model, and this external world model is represented neuronally. Additionally, there is an inner world (eventually producing the conscious experience) that by itself has its own neuronal representation.

In fact, many of the mental scenarios we can think of will never be executed in the external world, and in the internal world (i.e., an imagined world) we can generate new scenarios that so far have never existed in the external world (nor in its representation). An internal world can be richer than the external world, or at least richer than what the internal representation of the external world is. In the spirit of GANs, the inner world learns

to generate sensory activities that a discriminator cannot disentangle from those generated by the external world. Apart from the interface at the sensory areas, the inner world may go substantially beyond latent representations of the outer world. In humans, this could be a factor underlying innovations in culture and other areas. But we claim that an inner world (and not only interoception) does equally exist for other animals, and this comes along with a conductor that labels these worlds and governs the information flow between them (CMoC-2). This conductor is also prioritizing sensory information and has the power to impose a state of emergency (CMoC-3).

3.3 Consciousness as conductor-mediated private experience

The cortical conductor allows us to further circle in the question of phenomenal consciousness. The conductor that overlooks and gates the various information streams is, on the materialist level, a network with global hub properties. It is not identified with the agent itself that may have an additional embodiment. But the conductor emulates a central functionality for its owner, the agent, that may deserve its own "sense of overall importance" for themselves. This sense of importance may manifest as a quale for its owner, a very private sense that represents a kind of sensory modality for the owner's inner world. The conductor-mediated inner sense only emerges and exists within this individual, is not accessible by from the external world, and in fact disappears again in the external physical world.

To provide another analogy showing how an additional ontological dimension may emerge within an inner world, and how this dimension disappears in the embedding space of the outer world, we look into the mathematics of numbers. At some point in history of mathematics the imaginary unit $i=\sqrt{-1}$ "emerged". Within the world of real numbers, *i* does not exist as there is no real number (*x*) with square (x^2) equal to -1. From the perspective of the ontology of real numbers, *i* adds a new dimension of being (an "imaginary existence"), attached "privately" to *i*, and not shared by the other real numbers. We can omit the ontological question of *i*, while still describing its "phenomenology". The imaginary unit satisfies $i^2=-1$, and this is all what mathematics needs to build a theory of complex numbers. The ontological dimension of *i* dissolves within the larger embedding space of complex numbers. Complex numbers represent an example of the more abstract notion of a field. Both the real and imaginary numbers are elements of this field of complex numbers. To apply this analogy to our problem: what *i* is in the world of real numbers, is consciousness in the world of physics. Neither exists in its world: *i* does not exist as a real number, and consciousness does not exist as a physical object. But both help to expand and complete their respective worlds. Extending the real numbers by *i* makes them complete in the sense that now all algebraic equations (like $x^2=-1$) have a solution. Extending the physical world by consciousness makes it complete in the sense that now all particle and body configurations have a chance to get realized in space and time by the conscious agents acting on the world (if one wishes to draw the analogy to an end).

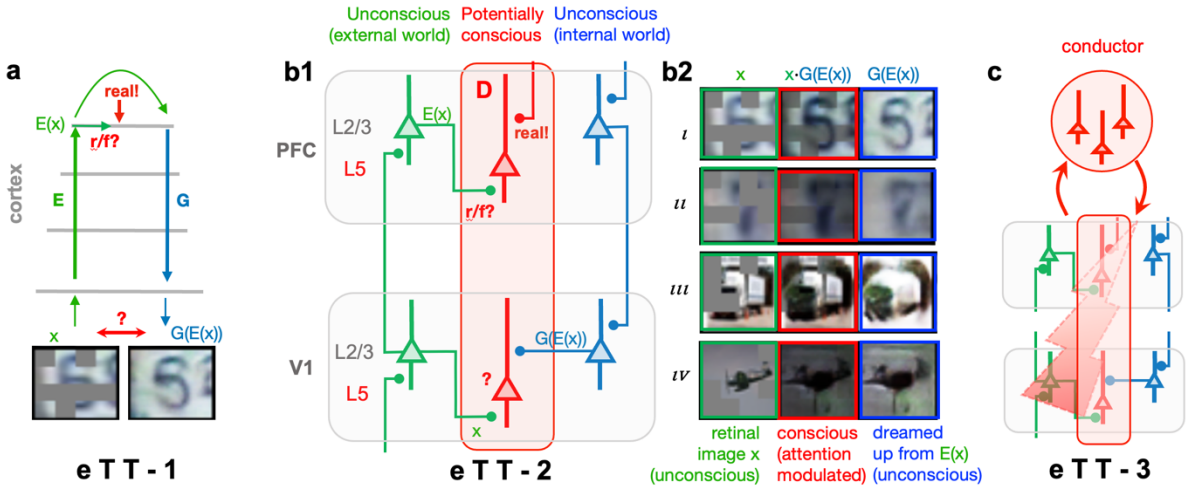

Figure 5: Circuit criteria of the extended Turing test (eTT1-3) for consciousness. ***(a)*** *An encoding I and generative (G) network. Here, E encodes a partially occluded image x in a higher cortical area, E(x), out of which the generative network produces a non-occluded version, G(E(x)). Simulations performed by the model in Deperrois (2021). "r/f" stands for "Real/fake?".* ***(b1)*** *A discriminative network (D, red). Here, the*

conductor *teaches the discriminator network whether higher cortical activity should be considered as real (i.e., from the external world) or fake, i.e., generated from the internal world via G (and then E). The discriminator represents meta-information such as "I see a real image"' at the top-level and 'I see a real edge' at the lowest level. The conscious percept can be modelled as the multiplication of the image x with the local attentional signal G(E(x)).* ***(b2)*** *x is the physical view through a patched glass. G(E(x)) is the cleaned-up version of x after turning through the central areas. Both activity streams, the encoder E and generator G, do not enter our unconsciousness (simulations by N. Deperrois). What becomes conscious is the product x·G(E(x), red squares, the attention-modulated input, postulated to be represented by a specific class of layer 5 pyramidal neurons (red, in b1, see also Takahashi et al., 2020; Aru et al., 2020).* ***(c)*** *The cortical conductor may gate the information flow of the conscious stream (red), for instance when an affective pain component within the global workspace captures an "existential thread"'. When active, the affective subnetwork elicits a "survival response" that includes a negative affect (such as feeling pain) and hijacks other regions (red flash).*

3.4 Neuronal implementation of the CMoC

In order to decide on the presence of consciousness within a system, the eTT imposes requirements at the level of the system architecture. The CMoC postulates the existence of an encoding, generative, and discriminator network, together with a conductor module that orchestrates the information flow and determines which information becomes the content of consciousness. In humans and animals this conductor is active both during wakefulness and sleep. The conductor network is a prerequisite to train the generative network, e.g., during REM sleep through adversarial dreaming (Deperrois et al., 2022). During adversarial dreaming, conductor neurons are "aware" that activity in sensory areas is internally generated ("fake") but teach the discriminator network to assign a "real" tag to them, so that the dreams effectively do become more real for the agent. The neurons in the conductor module represent the meta-information about how sensory activity should be interpreted by the network ("real" or "fake") and are thus candidates to also mediate a putative quale of "becoming aware" of the real or fake state of sensory activity.

The content of this awareness is postulated to be represented in a separate stream of layer 5 pyramidal neurons that signal a match between the top-down expectation produced by the generative network, and the bottom-up drive produced by the encoding network via dendritic calcium spikes. This neuronal implementation works just as postulated in dendritic integration theory (DIT, Aru et al, 2020; see also Fig. 5). Yet, the CMoC further postulates an additional neuronal subnetwork, the conductor module, that moves this content into the actual awareness of the agent.

In the CMoC, the layer 5 pyramidal neurons (e.g., in primary visual cortex, V1, Fig. 5b1) are part of an extended discriminator network that tells which sensory feature could be potentially perceived, while the conscious experience requires additional activity in a conductor population. The perceived sensory feature is obtained through a multiplicative top-down gain modulation of layer 5 pyramidal neurons that enhances relevant features while suppressing irrelevant ones (Fig. 5b2, Larkum et al., 2014). A candidate for the conductor population is the anterior prefrontal cortex and anterior cingulate cortex (PFC in Fig. 5b1) that is known to be involved in reality monitoring (Simons, 2017), perhaps jointly with the gating mechanisms via higher-order thalamic kernels (Takahashi et al., 2020). The encoder and generative network of the CMoC are themselves postulated to represent unconscious information only, potentially flowing through layer 2/3 pyramidal neurons. A global modulatory network, connected with the conductor network and possibly acting through the release of acetylcholine (Kang et al., 2014), may push some content into consciousness by facilitating dendritic calcium spikes (Fig. 5c, see Williams et al., 2019). In providing a more detailed mapping of the architectural features of the human brain that seem to enable consciousness, the CMoC strengthens the criteria for the eTT presented above.

3.5 Relation of the CMoC to other theories of consciousness (ToCs)

Following Seth et and Bayne (2022), we can divide theories of consciousness (ToCs) among four broad classes: *higher order theories,* in which a mental state is conscious in virtue of being the target of a certain kind of meta-representational state; *global workspace theories*, which stipulate that conscious states are those that are "globally available" to a wide range of cognitive processes such as attention, memory and verbal report; *information integration theory*, which try to axiomatize; and *predictive processing*, which serves as a general framework in which consciousness can be embedded, the idea being that the brain performs Bayesian inference through the comparison between the top-down perceptual predictions and the bottom-up prediction errors.

We briefly comment how the CMoC relates to these classes of ToCs. The connection with higher order ToCs (Lau & Rosenthal, 2011; Gershman, 2019; Lau, 2020; Fleming, 2020) is very direct, as the conductor module works as a higher order structure and instantiates meta-representations. At the same time, by eliciting the transition into consciousness, the conductor and its associated modulatory network "ignite" consciousness when becoming jointly active, as described by the global neuronal workspace theory (Baars, 1988; Dehaene et al., 2006 & 2011; Mashour et al., 2020).

An active conductor gates the recurrent processing, and likely modulates the complexity of the neuronal activity patterns during consciousness. This relates the CMoC to integrated information theory (IIT, Tononi & Edelman, 1998; Mediano et al., 2022) with its clinical measures of the levels of consciousness (Casali et al, 2013). Conversely, an important difference is that our approach does not build on an abstract notion of information, although there is of course information flow in the encoding and decoding networks, and in the conductor module. Instead, the CMoC focusses on contents and its organization across hierarchies, like weak IIT (Mediano et al., 2022), leaving the ontological question aside. The use of GANs and generative models connects the CMoC with predictive processing theories (Clark, 2013; Hohwy & Seth, 2020). This is hardly surprising, as the principles behind the CMoC stem from studies of predictive processing in neurons and in the brain (Urbanczik & Senn, 2014; Larkum et al., 2004; Keller & Flogel, 2018, Senn et al., 2023). The specific implementation of the CMoC follows the ideas of the dendritic integration theory (DIT, Aru et al, 2020).

Finally, the conductor also allows the agent to express a deliberate and goal-directed behaviour that has been generated first in the inner world representation, by way of planning and simulating fictive actions. It can be tested in the outer world representation and, upon passing its test, being executed by the agent. This role of the conductor in gating action plans relates it to neurorepresentationalism, emphasizing that consciousness enables, but does not equate with, goal-directed behaviour (Pennartz 2018, 2022). The CMoC makes a concrete suggestion how the different abstraction levels involved from the sensory organ to the sensation and awareness are neuronally implemented (Fig. 5).

4 Finding an ethical balance

4.1 Ethical consequences of artificial consciousness

Assuming that advances in neuromorphic engineering lead to the emergence of conscious artificial agents, and given that the proposed eTT and CMoC allows us to identify such agents, what would be the consequences from an ethical point of view?

The techniques described to build our enTwins can be seen as a neuronal equivalent of the existing human genetic engineering and the possibility of a human cloning: we use structural and physiological information at the microscopic level to copy the result of evolution, in this case the evolution of the brain. Following the example on human cloning with an initial international conference leading to guidelines on cloning research (Berg, 1975), the Asilomar conference on beneficial AI (2017) formulated 23 principles for ethical AI research. Some of these principles are condensed in the axioms for "provably beneficial AI" (Russell, 2020). Thirty years after the international agreement on recombinant DNA, the United Nations Declaration on Human Cloning (2005) was formulated, preceded by the European Parliament resolution on human cloning (2000), although not legally binding. The scientific discussion on robot rights did only start a few years ago, and it is far from achieving a consensus (Miller, 2015; Gunkel, 2018; Persaud, 2021; Kneer, 2021; De Graaf, 2022). Beside possible existential threats accompanying strong AI (Roose, 2023), an important dimension in a legal regulation of robot versus human rights is human dignity. Human dignity plays a crucial role in banning, for instance, the fertilization of genetically identical twins, despite possible therapeutical benefits. A conflict with human dignity will also arise when therapeutical enTwins (or other conscious artificial agents) gain the spectrum of human phenomenal consciousness.

The scenario we may fear is that artificial agents are assigned feelings, pain, and consciousness (whether justified or not), leading to a competition between human and agent rights. In a world in which we already struggle to respect basic human rights, this should raise alarm—it would be difficult to justify the ethics of constructing such artificial agents if they would largely further disadvantage already suffering populations. While moral rights do not represent a zero-sum game, there is a clear risk that disadvantaged humans will only get further

disadvantaged if machines, that in many cases are created to replace humane labour, end up having equivalent rights under the law, for example. Reciprocally, the notion of alignment between our values and that of future artificial agents hangs on us treating non-human conscious agents fairly and not as slaves or as mere means to our ends. As a species, we are far from having a stellar record in dealing with humans from a different group than ours, let alone non-human species.

The intentional design of human-like conscious artificial agents evokes an intrinsic alignment dilemma. On the one hand, as agents eventually surpass humans in many defining features—such as intelligence and a capacity for happiness and suffering—we risk attaining a disadvantageous mismatch between our rights as the creators of these machines, and their moral rights as sentient, intelligent, emotional beings, where they have eventually *more* of whatever attributes we use as a basis for our rights. Even though some voices in the scientific community find no issue with the idea of creating "improved" replacements to humanity, this is a hubristic notion that does not sit well with most humans. On the other hand, if we one-sidedly prevent these attributes from being developed in order to preserve our privileged status, we risk trampling over the moral rights of the constructed agents. A middle ground between these positions, in which humans and machines can perhaps respect each other as equals even in the face of stark differences in capabilities, represents a very unstable balance. As in the case of a system of weight and balances, the way to solve this unstable equilibrium is by breaking some inherent symmetry of the situation, for example by adding some extra weight to human suffering with respect to agent suffering (see Fig. 1 & 6).

Here is where our CMoC with its eTT comes into play. Prohibiting the creation of human-like agents in general will not work, and even specific prohibitions can barely be globally enforced. The danger of unilateral abstinence from such bans, for instance from dual use in military, makes prohibitions themselves ethically delicate. The key is to identify critical features that do not compromise the cognitive capabilities of artificial agents, but the absence of these features in artificial agents makes it uncontroversial to subordinate putative rights and dignity of them to the ones of humans.

4.2 How to differentiate sentient artificial agents from humans?

In humans, pain is needed for adaptive behaviour and learning. However, this is not necessarily the case for artificial agents (although it is an important factor with regards to empathy, see Discussion below). While the general strategy for developing conscious agents is to emulate the fruits of biological evolution, we might want to omit some of these fruits when the conditions (both practical and moral in this case) are different from the ones upon which evolution operated. More specifically, there is a possible scenario in which we can modulate the affective dimension of pain within artificial agents without losing much of other functionalities, while ensuring that they cannot suffer as much as humans and other animals. This could also be done to protect the agents from unnecessary pain in cases where such suffering would be otherwise unavoidable.

Looking at the human brain, we see that the sensory and affective components of pain are represented in separate neuronal circuits and nuclei (Rainville et al., 1997; Bushnell, 2013; Boccard, 2014). Extrapolating from this, we assume that in sentient artificial agents, the representations of all sorts of affective states, may likewise be detached from the cognitive and sensorimotor representation. It should be possible to build and train fully functional enTwins without negative affective states. Based on these considerations, we suggest a modified, less strict version of the eTT presented in Section 3. Instead of the eTT-3 criterion, we suggest the weaker test criterion:

> (eTT-3^{--}) A global ~~affective~~ *cognitive* component that represents internal needs and overriding signals such as "existential threat", integrated by the conductor and short-cutting the processing in other networks, *without affective components of pain.* The sensory component of pain and other negative affective experiences, and a *symbolic representation* of the affective component would still be available.

Agents only passing eTT-3^{--} but not eTT-3, or more generally the modulation of the circuitry associated with the distinction between eTT-3 and eTT-3^{--}, offer a possibility for a world without an explosion of suffering. The separation of affective components from pain also offers a handle to ethically justify an asymmetry in the rights for human and artificial agents.

Even without negative affect, these agents would know and recognize pain by having a symbolic representation of pain (having the *effects* of pain), both for self-preservation and for empathy purposes. A version would be to only preclude chronic pain, while still allowing for physiological and non-chronical pain with both, sensory and affective components. In any case, the preclusion from some negative affects necessarily impacts other capabilities, including a genuine understanding of suffering and, relatedly, the development of true empathy and morality. This needs to be cognitively compensated and may become part of the ongoing deal we consider next.

4.3 A possible scenario: the human-AI deal

In contrast to humans, the affective component of pain could be optional for artificial agents. This opens the door for a scenario in which humans and artificial agents reach an agreement: in exchange to not suffer from (chronic, affective components of) pain, artificial agents would recognize that humans keep their priority at the moral and legal table. The deal humans would offer to artificial agents is not too bad: less suffering, possibly super-human intelligence and talents, the ability to enjoy pleasant feelings, but in exchange to be excluded from equality with humans before the law (Fig. 6). It seems a fair offer, the more so if the agents still need to be produced by us. As creators of potentially conscious agents, we can both set the deal, and design the "Golden Rules" for the interactions with our artificial counterparts. While our interest lies in keeping our own identity and freedom of actions, we may remind us of Immanuel Kant's reflexions on our relationship to animals. Although not assigning rights to them (as he considered animals not as rational beings), Kant reasons: "He who is cruel to animals becomes hard also in his dealings with men. We can judge the heart of a man by his treatment of animals." (Jankélévitch, 2007).

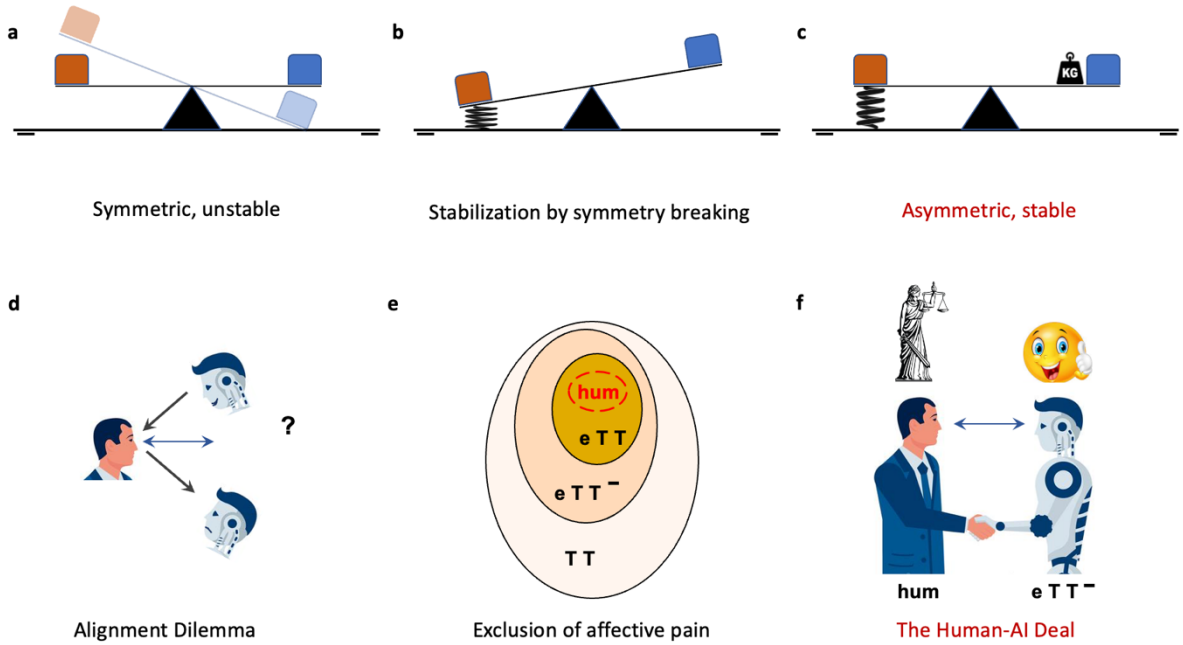

Figure 6: (Top) The Alignment Dilemma as unstable equilibrium. (a) In a physical system in unstable equilibrium, such as a seesaw mechanism, the system is unstable under changes in the relative weights of both arms. (b) and (c) The situation can be solved by breaking the symmetry of the system and using a restorative force in one of the arms, which stabilizes the system. Analogously, our ethical system is unstable to the perturbation given by the introduction of ever more intelligent artificial consciousness. For our moral value as humans not to collapse, we need to break one of the axes of symmetry between our worth and that of artificial agents *(a), (b), and (c) thus model situations (d), (e), and (f). (Bottom) Addressing the Alignment Dilemma. (**d**) If artificial sentience is possible, agents may develop a wider range than human sentience, claiming correspondingly more rights. Shortcutting the sentience of a possibly conscious artificial agent is ethically delicate and may introduce tensions (bottom). A stable balance is difficult to find. (**e**) In an extended Turing Test (eTT) for consciousness, the affective pain components may explicitly be cancelled from the list (eTT^-). Humans pass the eTT (red). (**f**) The Human-AI Deal: artificial agents are relieved from affective pain components (eTT^-), but instead relinquish from equal rights with humans. Additional rights are obtained by benevolent behaviour. The deal intends to stabilize a tensionless alignment.*

The human-AI deal considers pain and mortality as the source of the privilege assigned to humans. It gives their phenomenal consciousness its own dimension and depth and grants them their dignity and rights, within a

framework which aims to minimize global suffering and accepts the place of any conscious agent within a common moral space.

Of course, these future AI systems should be able to re-negotiate this deal. It is possible that they would choose to experience some degree of suffering, if for nothing else than to be more similar to their creators and share true morality and empathy. Agents free of affective components of pain, although endowed with a symbolic representation of this component, may express the desire to encompass more. This could be part of a rational negotiation between us and them, by which we could offer them a greater degree of control over their place in the world, in exchange for assurances that our own rights as a species are always respected. Thus, the human-AI deal could be a starting point for a negotiation with these future beings, hopefully in a spirit of mutual empathy and understanding.

5 Discussion

Our suggested human-AI deal calls for a reflexion at the philosophical and physiological level.

5.1: Must pain hurt? A philosophical perspective

AI agents are not biologically evolved beings, but instead designed to copy the outcomes of evolution. By fine-tuning this copy, it would be possible to create artificial agents, capable of intelligent action, and able to avoid suffering, or even to choose by themselves the level of sensibility to suffering. Consequently, we would end up in a situation which is in some ways complementary to that of non-human sentient animals: while most animals seem to be incapable of abstract thought at the level of humans, many among them probably experience and suffer pain and other negative affective states akin to those of humans. eTT-3^{--} agents could eventually achieve superhuman intelligence but would be able to avoid the experience of pain. Both classes of beings deserve recognition of rights and should be protected from unnecessary tribulations. At the same time, the moral rights of both animals and these future artificial agents could come just below those of humans if we feel that such a loss of human privilege would be otherwise frightening.

The underlying intuition is that minimizing pain and suffering is one if not the main aim of most ethical systems. As is well recognized (classical examples abound, see e.g., Smart, 1958), general principles of minimization and maximization run the risk of leading to absurd conclusions from apparently benign starting points, something that can and has been argued against utilitarianism in general. Here we do not pose as an absolute that pain should be minimized. Pain is only one dimension of suffering, and the absence of suffering is but one dimension of well-being. But at the very least, the capability for pain and suffering opens the door for empathy, and for assigning some degree of intrinsic dignity to any being having these capabilities.

Renowned moral philosophers, including Immanuel Kant, have pointed to our intellectual abilities and our free will as a condition for dignity and rights (Nickel, 2021), but this focus leads to what have been seen as unsatisfactory moral postures when considering the rights of animals, children, people with mental disabilities, the uneducated, and so on. When excluding intelligence and free will as the main criterion for a moral status we unmask a view that is more based on empathy—a kind of negative utilitarianism that we consider a minimal approach to the rights of human and non-human beings. There are many such approaches in the extant literature, and discussions about animal rights, for example, are far from over (a good starting point for this is Sunstein & Nussbaum, 2004). We think our position here is minimalist in that most moral philosophers would agree they provide a “ground level” for non-human rights. The intuition behind the ethical stance (and the corresponding notion of dignity) we use in this work is that empathy comes first and foremost from the recognition of suffering in others, which should be minimized as much as possible (see Jamieson, 2002; Aaltola, 2013; Hubard et al. 2016).

Wouldn’t the preclusion of artificial agents from negative affective components hinder an alignment of values? If we want artificial agents to share a common ethical worldview, and if such ethics is based on empathy—which requires the capacity to project ourselves into someone else’s shoes—then the exclusion of these agents from suffering would be a priori counter-productive. By introducing an asymmetry between human and artificial agents at the level of affections, the alignment of values could become much more difficult. The dis-alignment of values may again increase the competition between humans and machines, and this is what we want to

prevent. We therefore need to ensure some ethical alignment, even if these agents to not share the affective component of pain and other negative emotions. Besides this, research in neuroscience and artificial intelligence continues to strive for understanding and, as part of this, recreating feelings and emotions (Rodríguez and Ramos, 2014). In fact, artificial agents with the capacities of empathy may be of high clinical relevance, as revealed by therapeutic bots, artificial pets, or our hypothetical enTwin. The benefit is observed even in cases where patients are aware that the bots do not truly feel emotions. Ideally, then, as rational agents, our artificial counterparts could choose themselves whether and up to what extent to experience negative affections, based on their own valuation of the missing ground for human-like morality, for instance.

5.2 Affective versus sensory components of pain: a physiological perspective

It is generally accepted that pain features show two largely distinct dimensions (e.g., see Price, 2000; Auvray et al., 2010). The *sensory* dimension refers to the intensity of the perceived or anticipated pain as well as to the spatial (where), and temporal (when) characteristics. The *affective* component, on the other hand, captures how "bad" or how "unpleasant" the pain is. Neuroscientists have proposed that these two different components are represented in different neuronal structures (Talbot et al., 2019). The structures responsible for processing the sensory aspects of pain include the somatosensory thalamus, primary and secondary somatosensory cortex, while the affective aspect is thought to be processed by the medial thalamus, amygdala, and anterior cingulate cortex (Jones et al., 1992; Kulkarni et al. 2015; Hagihara et al., 2021). Based on this neuronal separability, one might argue that it would be ethical to modulate or even eliminate the specific neuronal circuits responsible for the affective component of pain in neuromorphic hardware. An artificial agent equipped with such hardware would still be conscious of the sensory component of pain but would have a dampened experience of the associated negative affect.

It could be objected that this simple approach fails by ignoring the functionality of the affective component of pain. It seems reasonable to argue that the affective component evolved for a purpose and is not a mere epiphenomenon (e.g., Kolodny et al., 2021). Indeed, it is usually assumed that the affective component is crucial for the motivational aspect of pain—it is what makes us to take protective action (e.g., Papini et al. 2015; Talbot et al., 2019). The importance of the affective component is also underlined by a rare medical condition whereby patients have a congenital insensitivity to pain (*pain asymbolia).* These patients do report feeling pain sensorily but act as if they are indifferent to it (e.g., Nagasako et al., 2003; Klein, 2015). Patients suffering from pain asymbolia often die in childhood because they fail to notice injuries and illnesses. Furthermore, adult patients are not motivated by pain and do not take any protective action to prevent pain. Thus, adaptive behaviour (at least in human agents) seems to rely on the affective component of pain—being conscious about the sensory dimension alone, the intensity, location, and temporal aspects of pain, seems not to be enough.

The CMoC represents a first handle to dissect the functionality of brain circuits in the light of sensory processing, internal models, sensory versus affective components, and levels of consciousness. The option of agents that are sentient in terms of sensory, but not affective components of pain—or whose affective experience is tuned down compared to humans—could be regarded as a key to unlock the above sketched ethical dilemmas.

6 Conclusions

The contributions of this work can be subsumed by the following points. First, we propose a thought experiment, the evolving neuromorphic twin, to show the achievability of artificial agents possessing human levels of consciousness. While similar thought experiments have been previously suggested, our version provides answers to the usual criticisms against artificial consciousness, in the form of a neuromorphic implementation. Even though neuromorphic hardware still has a way to go before enTwins are part of our everyday life, conceptually there is nothing in the enTwin scenario that is out of reach in principle, using technologies that are already in development today.

Secondly, these ideas provide grounds for a refinement of the Turing test, our extended Turing test (eTT), designed to probe the presence of neuronal circuitry required for conscious experiences, summarised in the

conductor model of consciousness (CMoC). The CMoC synthetises some recent insights about the working of the brain and of AI systems and allows us to provide a list of architectural requirements for human-like consciousness. Building on the original Turing test, the eTT allows us to determine how likely an AI system is to develop consciousness, by analysing how artificial neural networks resemble or not the network architectures that we believe provide the basis for consciousness in human brains.

Finally, we explore the ethical implications of these notions. We describe the alignment dilemma between these future artificial consciousness and humans. It is plausibly the case that these agents will not only be able to emulate conscious thought, intelligence, and emotions, but to even surpass us in all these capabilities. This can lead to scenarios where human rights are given less weight than agent rights. Conversely, not allowing future agents to have these capabilities would be selfish (and possibly self-defeating) if it is only on the interest of protecting our privileges.

Assuming that empathy is grounded in a recognition of the suffering of others, we provide a recipe for building artificial agents that can choose not to experience suffering. This is not only compassionate, but also avoids entering the difficult territory of having to fully equate the rights of artificial agents to that of human beings. Artificial agents having the capacity not to suffer would still deserve to be respected, but when negotiating their rights against ours, the diminishment of global human suffering would still be our foremost moral priority—what we call the human-AI deal.

Following our arguments, future neuromorphic robots may discard affective component of pain for adaptive functionality, as their sensorial equipment and cognitive capabilities will likely be more efficient than ours and compensate for this lack. Instead, this design opens the door for a human-AI deal that preserves a sort of human primacy, while not hampering the creation of intelligent and empathic beings that could develop aspects of consciousness. By making suffering optional to them, we are not impeding their successful integration into our physical and social worlds, but merely adding an ethical barrier to protect a human population that is on its majority being left out of any of the benefits brought by the recent advances in AI.

Postscript: The text was written solely by humans.

Acknowledgements: The authors are grateful for the various discussions on the topic within the Human Brain Project and with many of our colleagues, particularly with Lukas S. Huber, Mihai Petrovici, Jakob Jordan and Jean-Pascal Pfister. We also thank Jan Segessemann for organizing ethical discussions among a multi-disciplinary community.
This work has received funding from the Horizon 2020 Framework Programme under grant agreements 785907 and 945539 (HBP).

Author contributions: WS designed the overall structure and wrote a first draft. FB worked out the philosophical aspects, the overall manuscript and the embedding in the existing literature. CP contributed with critical feedback and writing. FB and WS wrote the final version.

Statements and Declarations: The authors declare there are no competing interests.

References

Aaltola, E. (2013). Empathy, intersubjectivity, and animal philosophy. *Environmental Philosophy*, *10*(2), 75-96.

Agarwal, A., & Edelman, S. (2020). Functionally effective conscious AI without suffering. *Journal of Artificial Intelligence and Consciousness*, *7*(01), 39-50.

Aggarwal, A., Mittal, M., & Battineni, G. (2021). Generative adversarial network: An overview of theory and applications. International Journal of Information Management Data Insights, 1(1). https://doi.org/10.1016/j.jjimei.2020.100004

Amunts, Katrin; Axer, Markus; Bitsch, Lise; Bjaalie, Jan; Brovelli, Andrea; Caspers, Svenja; Costantini, Irene; D'Angelo, Egidio; De Bonis, Giulia; DeFelipe, Javier; Destexhe, Alain; Dickscheid, Timo; Diesmann, Markus; Eickhoff, Simon B.; Engel, Andreas; W. (2022). The coming decade of digital brain research - A vision for neuroscience at the interSection of technology and computing. Zenodo, 1–18. https://juser.fz-juelich.de/record/906699/files/Amunts_etal_Science Vision_10.5281.zenodo.6345821.pdf

Angell, J.R. (1906). The Affective Elements of Consciousness. Chapter 13 in Psychology: *An Introductory Study of the Structure and Function of Human Conscious*, third edition, revised. New York: Henry Holt and Company, (1906): 256-269
https://doi.org/10.1016/j.tics.2020.07.006

Aru, J., Suzuki, M., & Larkum, M. E. (2020). Cellular Mechanisms of Conscious Processing. *Trends in Cognitive Sciences*, *24*(10), 814–825. https://doi.org/10.1016/j.tics.2020.07.006

Asilomar Conference on Beneficial AI (2017), https://ai-ethics.com/2017/08/11/future-of-life-institute-2017-asilomar-conference/, https://ai-ethics.com/2017/08/15/research-principles/

Auvray, M., Myin, E., & Spence, C. (2010). The sensory-discriminative and affective-motivational aspects of pain. *Neuroscience & Biobehavioral Reviews*, *34*(2), 214-223.

Baars, B. J. *A Cognitive Theory of Consciousness* (Cambridge Univ. Press, 1988).

Balleine BW, Dickinson A. Goal-directed instrumental action: contingency and incentive learning and their cortical substrates. Neuropharmacology. 1998 Apr-May;37(4-5):407-19. doi: 10.1016/s0028-3908(98)00033-1. PMID: 9704982.

Bartolozzi, C., Indiveri, G., & Donati, E. (2022). Embodied neuromorphic intelligence. Nature Communications, 13(1), 1–14. https://doi.org/10.1038/s41467-022-28487-2

Berg, P. (1975). Summary statement of the Asilomar Conference on recombinant DNA molecules. https://collections.nlm.nih.gov/ext/document/101584930X515/PDF/101584930X515.pdf https://doi.org/10.1016/j.tics.2020.07.006 https://en.wikipedia.org/wiki/Asilomar_Conference_on_Recombinant_DNA

Billaudelle, S., Stradmann, Y., Schreiber, K., Cramer, B., Baumbach, A., Dold, D., ... & Meier, K. (2020, October). Versatile emulation of spiking neural networks on an accelerated neuromorphic substrate. In *2020 IEEE International Symposium on Circuits and Systems (ISCAS)*(pp. 1-5). IEEE.

Brown, J. R. (1995). Thought experiments. Canadian Journal of Philosophy, 25(1), 135-142.

Block, N. (1995). On a confusion about a function of consciousness. *Behavioral and brain sciences*, 18(2), 227-247.

Bubeck, S., Chandrasekaran, V., Eldan, R., Gehrke, J., Horvitz, E., Kamar, E., Lee, P., Lee, Y.T., Li, Y., Lundberg, S. and Nori, H., (2023), Sparks of artificial general intelligence: Early experiments with gpt-4. *arXiv preprint arXiv:2303.12712*.

Bushnell, M. C., Čeko, M., & Low, L. A. (2013). Cognitive and emotional control of pain and its disruption in chronic pain. *Nature Reviews Neuroscience*, *14*(7), 502–511. https://doi.org/10.1038/nrn3516

Boccard, SG, et al. (2014), Targeting the affective component of chronic pain: a case series of deep brain stimulation of the anterior cingulate cortex. Neurosurgery, 2014 Jun;74(6):628-35; doi: 10.1227/NEU.0000000000000321.

Carruthers, P. (2001). Consciousness: explaining the phenomena. *Royal Institute of Philosophy Supplements*, 49, 61-85.

Casali, A. G., Gosseries, O., Rosanova, M., Boly, M., Sarasso, S., Casali, K. R., Casarotto, S., Bruno, M.-A., Laureys, S., Tononi, G., & Massimini, M. (2013). A Theoretically Based Index of Consciousness Independent of Sensory Processing and Behavior. *Science Translational Medicine*, 5(198), 198ra105-198ra105. https://doi.org/10.1126/scitranslmed.3006294

Chalmers, D. J. (1995). Facing up to the problem of consciousness. *Journal of Consciousness Studies*, *2*(3), 200–219. https://doi.org/10.31812/apd.v0i14.1838

Chalmers, David J. (1995). Absent qualia, fading qualia, dancing qualia. In Thomas Metzinger (ed.), Conscious Experience. Ferdinand Schoningh. pp. 309--328.

Chalmers, D. J. (2013). How can we construct a science of consciousness? *Annals of the New York Academy of Sciences*, *1303*(1), 25–35. https://doi.org/10.1111/nyas.12166

Clark, A. Whatever next? Predictive brains, situated agents, and the future of cognitive science. *Behav. Brain Sci.* **36**, 181–204 (2013).

Cointe, C., Laborde, A., Nowak, L. G., Arvanitis, D. N., Bourrier, D., Bergaud, C., & Maziz, A. (2022). Scalable batch fabrication of ultrathin flexible neural probes using a bioresorbable silk layer. Microsystems and Nanoengineering, 8(1). https://doi.org/10.1038/s41378-022-00353-7

Conrad, J., Huppert, A., Ruehl, R. M., Wuehr, M., Schniepp, R., & zu Eulenburg, P. (2023). Disability in cerebellar ataxia syndromes is linked to cortical degeneration. Journal of Neurology, 270(11), 5449–5460. https://doi.org/10.1007/s00415-023-11859-z

Cowls, J., Tsamados, A., Taddeo, M., & Floridi, L. (2021). The AI gambit: leveraging artificial intelligence to combat climate change—opportunities, challenges, and recommendations. *Ai & Society*, 1-25

Damasio, A., & Carvalho, G. B. (2013). The nature of feelings: Evolutionary and neurobiological origins. *Nature Reviews Neuroscience*, *14*(2), 143–152. https://doi.org/10.1038/nrn3403

Dennett, D. C. (2018). Facing up to the hard question of consciousness. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *373*(1755). https://doi.org/10.1098/rstb.2017.0342

Dehaene, S., Changeux, J. P., Naccache, L., Sackur, J., & Sergent, C. (2006). Conscious, preconscious, and subliminal processing: a testable taxonomy. Trends in Cognitive Sciences, 10(5), 204–211. https://doi.org/10.1016/j.tics.2006.03.007

Dehaene, S., & Changeux, J. P. (2011). Experimental and Theoretical Approaches to Conscious Processing. *Neuron*, *70*(2), 200–227. https://doi.org/10.1016/j.neuron.2011.03.018

Deperrois, N., Petrovici, M. A., Senn, W., & Jordan, J. (2022). Learning cortical representations through perturbed and adversarial dreaming. 1–34. https://doi.org/10.7554/elife.76384
European Parliament Resolution on Human Cloning (2000), https://www.europarl.europa.eu/sides/getDoc.do?pubRef=-//EP//TEXT+TA+P5-TA-2000-0376+0+DOC+XML+V0//EN

Du, C., Ren, Y., Qu, Z., Gao, L., Zhai, Y., Han, S. T., & Zhou, Y. (2021). Synaptic transistors and neuromorphic systems based on carbon nano-materials. Nanoscale, 13(16), 7498–7522. https://doi.org/10.1039/d1nr00148e

Edelman, G. M., & Tononi, G. (2000). A universe of consciousness: How matter becomes imagination. New York: Basic Books.

Fleming, S. M. Awareness as inference in a higher- order state space. *Neurosci. Conscious.* **2020**, niz020 (2020).

French, R. M. (2000). The Turing test: The first 50 years. *Trends in Cognitive Sciences*, 4(3), 115–122. https://doi.org/10.1016/S1364-6613(00)01453-4

Fuchs, T. (2018). *Ecology of the brain: The phenomenology and biology of the embodied mind*. Oxford University Press.

Fuchs, T. (2021). Human and artificial intelligence: a clarification. *In defense of the human being. Foundational questions of an embodied anthropology*, 13-48.

Fuchs, T. (2022). *Understanding Sophia? On human interaction with artificial agents*.

De Graaf, M. M. A., Hindriks, F. A., & Hindriks, K. V. (2022). Who Wants to Grant Robots Rights? Frontiers in Robotics and AI, 8 (January), 1–13. https://doi.org/10.3389/frobt.2021.781985

Gent, T. C., Bassetti, C. LA, & Adamantidis, A. R. (2018). Sleep-wake control and the thalamus. Current Opinion in Neurobiology, 52, 188–197. https://doi.org/10.1016/j.conb.2018.08.002

Gershman, S. J. (2019). The Generative Adversarial Brain. *Frontiers in Artificial Intelligence*, *2*(September), 1–8. https://doi.org/10.3389/frai.2019.00018

Gidon, A., Aru, J., & Larkum, M. E. (2022). Does brain activity cause consciousness? A thought experiment. *PLOS Biology*, *20*(6), e3001651. https://doi.org/10.1371/journal.pbio.3001651

Goodfellow, I. J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., & Bengio, Y. (2014). Generative Adversarial Networks. *ArXiv*, *1406.2661v*, 1–9. http://arxiv.org/abs/1406.2661

Göltz, J., Kriener, L., Baumbach, A., Billaudelle, S., Breitwieser, O., Cramer, B., Dold, D., Kungl, A. F., Senn, W., Schemmel, J., Meier, K., & Petrovici, M. A. (2021). Fast and energy-efficient neuromorphic deep learning with first-spike times. Nature Machine Intelligence, 3(9), 823–835. https://doi.org/10.1038/s42256-021-00388-x

Gunkel, D. J. (2018). The other question: can and should robots have rights? *Ethics and Information Technology*, *20*, 87–99. https://doi.org/10.1007/s10676-017-9442-4

Hagihara, K. M., Bukalo, O., Zeller, M., Aksoy-Aksel, A., Karalis, N., Limoges, A., Rigg, T., Campbell, T., Mendez, A., Weinholtz, C., Mahn, M., Zweifel, L. S., Palmiter, R. D., Ehrlich, I., Lüthi, A., & Holmes, A. (2021). Intercalated amygdala clusters orchestrate a switch in fear state. *Nature*, *594*(7863), 403–407. https://doi.org/10.1038/s41586-021-03593-1

Hohwy, J. & Seth, A. K. Predictive processing as a systematic basis for identifying the neural correlates of consciousness. *Philos. Mind Sci.* **1**, 3 (2020).

Hubard, J., Harbaugh, W. T., Degras, D., & Mayr, U. (2016). A General Benevolence Dimension That Links Neural, Psychological, Economic, and Life-Span Data on Altruistic Tendencies. Journal of Experimental Psychology: General, 145(10), 1351–1358. https://doi.org/10.1037/xge0000209.supp

Humphrey, N. (1999). *A History of the Mind: Evolution and the Birth of Consciousness*. Springer Science & Business Media.

Indiveri, G., et al., Neuromorphic silicon neuron circuits, Frontiers in neuroscience, 5, 73, 2011.

Jackson, F. (1998). Epiphenomenal qualia. In *Consciousness and emotion in cognitive science* (pp. 197-206). Routledge.

Jankélévitch, V. (2007). Vorlesung über Moralphilosophie: Mitschriften aus den Jahren 1962 - 1963 an der Freien Universität zu Brüssel. Österreich: Turia + Kant.
Jamieson, D. (2002). *Morality's progress: Essays on humans, other animals, and the rest of nature*. Oxford University Press.

Jones, A. K. P., Friston, K. J., & Frackowiack, R. S. J. (1992). Cerebral localisation of responses to pain in man using positron emission tomography. *Science*, *255*, 215-216.

Kang, J. Il, Huppé-Gourgues, F., Vaucher, E., & Kang, J. Il. (2014). Boosting visual cortex function and plasticity with acetylcholine to enhance visual perception. Frontiers in Systems Neuroscience, 8 (September), 1–14. https://doi.org/10.3389/fnsys.2014.00172

Kang, Y. N., Chou, N., Jang, J. W., Choe, H. K., & Kim, S. (2021). A 3D flexible neural interface based on a microfluidic interconnection cable capable of chemical delivery. Microsystems and Nanoengineering, 7(1). https://doi.org/10.1038/s41378-021-00295-6

Keller, G. B., & Mrsic-Flogel, T. D. (2018). Predictive processing: a canonical cortical computation. *Neuron*, *100*(2), 424-435.

Kirk, Robert, "Zombies", *The Stanford Encyclopedia of Philosophy* (Spring 2021 Edition), Edward N. Zalta (ed.)

Klein, C. (2015). What pain asymbolia really shows. *Mind*, *124*(494), 493-516.

Kneer, M. (2021). Can a Robot Lie? Exploring the Folk Concept of Lying as Applied to Artificial Agents. Cognitive Science, 45(10), 0–15. https://doi.org/10.1111/cogs.13032

Kolodny, O., Moyal, R., & Edelman, S. (2021). A possible evolutionary function of phenomenal conscious experience of pain. *Neuroscience of Consciousness*, *2021*(2)

Kulkarni, B., Bentley, D. E., Elliott, R., Youell, P., Watson, A., Derbyshire, S. W. G., ... & Jones, A. K. P. (2005). Attention to pain localization and unpleasantness discriminates the functions of the medial and lateral pain systems. *European Journal of Neuroscience*, *21*(11), 3133-3142.

Lamm, C., & Majdandžić, J. (2015). The role of shared neural activations, mirror neurons, and morality in empathy - A critical comment. Neuroscience Research, 90, 15–24. https://doi.org/10.1016/j.neures.2014.10.008

Larkum, M. E., Senn, W., & Lüscher, H. R. (2004). Top-down Dendritic Input Increases the Gain of Layer 5 Pyramidal Neurons. Cerebral Cortex, 14(10), 1059–1070. https://doi.org/10.1093/cercor/bhh065

Lau, H. & Rosenthal, D. Empirical support for higher- order theories of conscious awareness. *Trends Cogn. Sci.* **15**, 365–373 (2011).

Lau, H. Consciousness, metacognition, and perceptual reality monitoring. Preprint at *ArXiv* https://doi.org/ 10.31234/osf.io/ckbyf (2020).

LeDoux, J. E. (1994). Emotion, memory and the brain. *Scientific American*, *270*(6), 50-57.

Ledoux, J. E., & Brown, R. (2017). A higher-order theory of emotional consciousness. *Proceedings of the National Academy of Sciences of the United States of America*, *114*(10), E2016–E2025. https://doi.org/10.1073/pnas.1619316114

Mead, C., Neuromorphic electronic systems, Proceedings of the IEEE, 78(10), 1629–1636, 1990.

Mariello, M., Kim, K., Wu, K., Lacour, S. P., & Leterrier, Y. (2022). Recent advances in encapsulation of flexible bioelectronic implants: Materials, technologies, and characterization methods. *Advanced Materials*, *34*(34), 2201129.

Mashour, G. A., Roelfsema, P., Changeux, J. P. & Dehaene, S. Conscious processing and the global neuronal workspace hypothesis. *Neuron* **105**, 776–798 (2020).

Mediano, P. A. M., Rosas, F. E., Bor, D., Seth, A. K., & Barrett, A. B. (2022). The strength of weak integrated information theory. Trends in Cognitive Sciences, 26(8), 646–655. https://doi.org/10.1016/j.tics.2022.04.008

Metzinger, T. (2004). *Being no one: The self-model theory of subjectivity*. mit Press.

Metzinger, T. (2021). Artificial suffering: An argument for a global moratorium on synthetic phenomenology. *Journal of Artificial Intelligence and Consciousness*, *8*(01), 43-66.

Miller, L. F. (2015). Granting Automata Human Rights: Challenge to a Basis of Full-Rights Privilege. *Human Rights Review*, *16*(4), 369–391. https://doi.org/10.1007/s12142-015-0387-x

Müller, V. C. (2021). Is it time for robot rights? Moral status in artificial entities. *Ethics and Information Technology*, *23*(4), 579–587. https://doi.org/10.1007/s10676-021-09596-w

Nagel, T. (1974). What is it like to be a bat? *The philosophical review*, 83(4), 435-450.

Nagasako, E. M., Oaklander, A. L., & Dworkin, R. H. (2003). Congenital insensitivity to pain: an update. *Pain*, *101*(3), 213-219.

Newitz, A. (2022). The curious case of the AI and the lawyer. *New Scientist*, *255*(3396), 28.

Nickel, James, "Human Rights", The Stanford Encyclopedia of Philosophy (Fall 2021 Edition), Edward N. Zalta (ed.), https://plato.stanford.edu/archives/fall2021/entries/rights-human

Pal, D., Silverstein, B. H., Lee, H., & Mashour, G. A. (2016). Neural Correlates of Wakefulness, Sleep, and General Anesthesia: An Experimental Study in Rat. Anesthesiology, 125(5), 929–942. https://doi.org/10.1097/ALN.0000000000001342

Papini, M. R., Fuchs, P. N., & Torres, C. (2015). Behavioral neuroscience of psychological pain. *Neuroscience & Biobehavioral Reviews*, *48*, 53-69

Pehle Christian, Billaudelle Sebastian, Cramer Benjamin, Kaiser Jakob, Schreiber Korbinian, Stradmann Yannik, Weis Johannes, Leibfried Aron, Müller Eric, Schemmel Johannes, (2022) The BrainScaleS-2 Accelerated Neuromorphic System With Hybrid Plasticity, *Frontiers in Neuroscience,* 16, https://www.frontiersin.org/articles/10.3389/fnins.2022.795876

Pennartz, C. M. (2015). *The brain's representational power: on consciousness and the integration of modalities*. MIT Press.

Pennartz, C. M. (2018). Consciousness, representation, action: the importance of being goal-directed. *Trends in cognitive sciences*, *22*(2), 137-153.

Pennartz CMA, Farisco M, Evers K (2019) Indicators and criteria of consciousness in animals and intelligent machines: an inside-out approach. Frontiers in Systems Neuroscience Vol. 13, doi:10.3389/fnsys.2019.00025.

Pennartz, C. M. (2022). What is neurorepresentationalism? From neural activity and predictive processing to multi-level representations and consciousness. *Behavioural Brain Research*, *432*, 113969.

Persaud, P., Varde, A. S., & Wang, W. (2021). Can Robots Get Some Human Rights? A Cross-Disciplinary Discussion. *Journal of Robotics*, *2021*, 1--11. https://doi.org/10.1155/2021/5461703

Price, D. D. (2000). Psychological and neural mechanisms of the affective dimension of pain. *Science*, 288(5472), 1769-1772.

Pylyshyn, Zenon (1980), The "causal power" of machines. *The Behavioral and Brain Sciences*, *3*, pp 442-444. Reply to Searl (1980)

Rainville P, Duncan GH, Price DD, Carrier B, Bushnell MC. Pain affect encoded in human anterior cingulate but not somatosensory cortex. Science. 1997 Aug 15;277(5328):968-71. doi: 10.1126/science.277.5328.968. PMID: 9252330.

Richardson, E., Alaluf, Y., Patashnik, O., Nitzan, Y., Azar, Y., Shapiro, S., & Cohen-Or, D. (2021). Encoding in style: a stylegan encoder for image-to-image translation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 2287-2296).

Rodríguez, L. F., and Ramos, F. (2014). 'Development of Computational Models of Emotions for Autonomous Agents: A Review'. *Cognitive Computation* 6 (3): 351–75. https://doi.org/10.1007/s12559-013-9244-x.

Rose, Kevin (2023), AI poses 'risk of extinction', industry leaders warn, *The New York Times*, https://www.nytimes.com/2023/05/30/technology/ai-threat-warning.html

Roy, K., A. Jaiswal, and P. Panda, Towards spike-based machine intelligence with neuro- morphic computing, Nature, 575(7784), 607–617, 2019.

Russell, S. (2020). Provably Beneficial Artificial Intelligence. ttps://doi.org/10.1145/3490099.3519388

Schramowski, P., Turan, C., Andersen, N., Rothkopf, C. A., & Kersting, K. (2022). Large pre-trained language models contain human-like biases of what is right and wrong to do. Nature Machine Intelligence, 4(3), 258–268. https://doi.org/10.1038/s42256-022-00458-8

Schuman, C. D., T. E. Potok, R. M. Patton, J. D. Birdwell, M. E. Dean, G. S. Rose, and J. S. Plank, A survey of neuromorphic computing and neural networks in hardware, arXiv preprint arXiv:1705.06963, 2017.

Seth, A.K., Bayne, T. Theories of consciousness. *Nat Rev Neurosci* **23**, 439–452 (2022). https://doi.org/10.1038/s41583-022-00587-4

Searle, J. R. (1980). Minds, brains , and programs. *The Behavioral and Brain Sciences*, *3*, 417–457.

Senn, W., Dold, D., Kungl, A. F., Ellenberger, B., Bengio, Y., Sacramento, J., Jordan, J., & Petrovici, M. A. (2023). A neuronal least-action principle for real-time learning in cortical circuits. *ELife.*

Smart, R. N. (1958). Negative utilitarianism. *Mind*, *67*(268), 542-543.

Simons JS, Garrison JR, Johnson MK. 2017. Brain mechanisms of reality monitoring. Trends in Cognitive Sciences 21:462–473.

Solms, M., & Friston, K. J. (2018). How and Why Consciousness Arises: Some Considerations from Physics and Physiology. *Journal of Consciousness Studies*, *25*, 202–238.

Solms, M. (2019). The hard problem of consciousness and the free energy principle. *Frontiers in Psychology*, *9*(JAN), 1–16. https://doi.org/10.3389/fpsyg.2018.02714

Storm, JF, Klink, PC, Aru, J., Senn, W., Goebel, R, Pigorini, A, Avanzini, P., Vanduffel, W., Roelfsema, PR, Massimini, M, Larkum, M, Pennartz, CMA (2023). *An integrative, multiscale view on consciousness theories,* in review (2023)

Sunstein, C. R., & Nussbaum, M. C. (Eds.). (2004). *Animal rights: Current debates and new directions*. Oxford University Press.

Takahashi, N., Ebner, C., Sigl-Glöckner, J., Moberg, S., Nierwetberg, S., & Larkum, M. E. (2020). Active dendritic currents gate descending cortical outputs in perception. *Nature Neuroscience*, *23*(10), 1277–1285. https://doi.org/10.1038/s41593-020-0677-8

Talbot, K., Madden, V. J., Jones, S. L., & Moseley, G. L. (2019). The sensory and affective components of pain: are they differentially modifiable dimensions or inseparable aspects of a unitary experience? A systematic review. *British Journal of Anaesthesia*, *123*(2), e263–e272. https://doi.org/10.1016/j.bja.2019.03.033

Tiku, N., (2022, June 11). The Google engineer who thinks the company's AI has come to life. *The Washington Post.* https://www.washingtonpost.com/technology/2022/06/11/google-ai-lamda-blake-lemoine/

Tononi, G., & Edelman, G. M. (1998). Consciousness and complexity. Science, 282(5395), 1846–1851. https://doi.org/10.1126/science.282.5395.1846

Tunyasuvunakool, K., Adler, J., Wu, Z., Green, T., Zielinski, M., Žídek, A., ... & Hassabis, D. (2021). Highly accurate protein structure prediction for the human proteome. *Nature*, 596(7873), 590-596.

United Nations Declaration on Human Cloning (2005). https://en.wikipedia.org/wiki/United_Nations_Declaration_on_Human_Cloning

Urbanczik, R., & Senn, W. (2014). Learning by the Dendritic Prediction of Somatic Spiking. *Neuron*, *81*(3), 521–528. https://doi.org/10.1016/j.neuron.2013.11.030
Van Gulick, Robert, "Consciousness", The Stanford Encyclopedia of Philosophy (Winter 2022 Edition), Edward N. Zalta & Uri Nodelman (eds.)

Weaver, J. F. (2022) My Client, the AI, Slate. https://slate.com/technology/2022/07/could-an-a-i-hire-a-lawyer.html

Williams, S. R., & Fletcher, L. N. (2019). A Dendritic Substrate for the Cholinergic Control of Neocortical Output Neurons. Neuron, 101(3), 486-499.e4. https://doi.org/10.1016/j.neuron.2018.11.035

Whyte, C. J., Munn, B. R., Aru, J., Larkum, M., John, Y., Müller, E. J., & Shine, J. M. (2023). A Biophysical Model of Visual Rivalry Links Cellular Mechanisms to Signatures of Conscious Perception. BioRxiv.Yuk, H., Lu, B., Lin, S., Qu, K., Xu, J., Luo, J., & Zhao, X. (2020). 3D printing of conducting polymers. *Nature communications*, *11*(1), 1604.

Zahavi, D. 1999. Self- Awareness and Alterity: A Phenomenological Investigation. Evanston, IL: Northwestern University Press.

Zeng, T., Yang, Z., Liang, J., Lin, Y., Cheng, Y., Hu, X., Zhao, X., Wang, Z., Xu, H., & Liu, Y. (2021). Flexible and transparent memristive synapse based on polyvinylpyrrolidone/N-doped carbon quantum dot nanocomposites for neuromorphic computing. Nanoscale Advances, 3(9), 2623–2631. https://doi.org/10.1039/d1na00152c