AI and Ethics
https://doi.org/10.1007/s43681-024-00580-w

ORIGINAL RESEARCH

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

Federico Benitez[1] · **Cyriel Pennartz**[2] · **Walter Senn**[1]

Received: 19 January 2024 / Accepted: 13 September 2024

Abstract
Critics of Artificial Intelligence (AI) posit that artificial agents cannot achieve consciousness even in principle, because they lack certain necessary pre-conditions 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. Based on the construction of a co-evolving neuromorphic twin, we argue that the differences between a developing biological and artificial brain 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 lists functional and neuronal correlates of biological consciousness that are captured by the CMoC. These correlates provide the grounding for how biological or artificial agents learn to distinguish between sensory activity generated from outside or inside of the brain, how the perception of these activities can itself be learned, and how the information flow for learning an internal world model is orchestrated by a cortical meta-instance, which we call the conductor. Perception comes with the distinction of sensory and affective components, with the affective component linking to ethical questions that are inherent in our multidimensional model of consciousness. Recognizing the existence of a blueprint for a possible artificial consciousness encompasses functional, neuronal and ethical dimensions, begging the question: How should we behave towards agents that are akin to us in the inner workings of their brains? We sketch a human-AI deal, balancing the growing cognitive abilities of artificial agents, and the possibility to relieve them from suffering of negative affects, with a protection for the rights of humans.

Keywords Consciousness · Artificial intelligence · Turing test · AI · Ethics

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 one day help us tackling vital problems, such as the development of new drugs [135], and fighting climate change through the development of renewable energy technologies and the optimization of resource use [37].

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, Language Model for Dialogue Application (LaMDA), is already conscious and capable of suffering [133], see also [100, 139], most artificial intelligence researchers (including at Google) 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

✉ Federico Benitez
federico.benitez@unibe.ch

[1] Institute of Physiology and Center of Artificial Intelligence in Medicine (CAIM), Faculty of Medicine, University of Bern, Bern, Switzerland

[2] Swammerdam Institute for Life Sciences, Center for Neuroscience, Faculty of Science, University of Amsterdam, Amsterdam, Netherlands

Published online: 02 October 2024

general artificial intelligence" in GPT-4 [21]. In philosophy, however, questions of principle remain 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., [48, 54–56, 138]).

In this work we take the perspective of computational neuroscience to address some scientific, technical, and ethical aspects of this issue. The primary goal is to contribute to the ethical debate on how to deal with AI, by adding a specific computational neuroscience account to the field of consciousness research. Prominent AI researchers are warning society about the existential risks that AI poses to humanity [15, 32]. The possibility of consciousness arising in AI is also well considered in this community, including the danger of over- or under-attributing it to AI [24]. From the point of view of neuroscience, in turn, arguments on the unreached complexity of biological consciousness prevail [9, 110]. Here, we argue that despite this biological complexity, some forms of consciousness might be possible, if functional and observable criteria of consciousness are satisfied. In view of ethical questions, such as how many rights and protection artificial agents with these forms of consciousness should be granted in comparison to humans [45], we need guiding criteria for awareness and consciousness. To start with a simple functional question, we first ask how agents can learn to distinguish between themselves and the environment, and next, how they can learn to perceive themselves and the environment. Perception involves an additional instance that internally represents the "who" that looks at the content (a discriminator), and the "from where" the content is generated (from inside or outside).

Bearing in mind the ethical dimension, we distinguish between a sensory and affective component of perception, with the affective component referring to the engagement of the whole organism in processing information, originating in survival reactions to threat. Affective (or valenced) experiences would help artificial agents to align with human values and develop, for instance, empathy for humans and among themselves as a basis for respectful interactions.

1.1 Functional correlates of consciousness

A central notion we introduce are *functional correlates of consciousness*. So far, neuroscientific theories of consciousness directly try to identify the neural correlates [129]. Yet, since multiple brain areas are involved in representing consciousness, it may help to first structure their putative contributions in terms of functions. A functional characterization of the areas, in turn, requires an idea of how consciousness itself can be subdivided in functional sub-modules, and how they map to neuronal correlates.

Functional correlates go beyond indicators of consciousness [108]. An indicator, for instance, can be the degree of information complexity in the activity traces of a conscious brain [26], or the recurrent processing itself [82], without necessarily specifying their function in the context of consciousness. The recent advances in AI push computational functionalism into the foreground, claiming that with the construction of a sufficiently elaborate system that performs certain kind of computations, the phenomenon of consciousness could emerge [24]. Following this hypothesis, we suggest a specific type of computation underlying the formation of awareness and eventually of consciousness. This involves learning to discriminate internally from externally generated activity in sensory areas, while representing the reality judgement in a dedicated neuronal population. We postulate that the assignment of the reality-label by this neuronal population ultimately grounds conscious sensory experience.

This leads us to re-evaluate the classical thought experiment of replacing each neuron within a human brain by an artificial counterpart, resulting in the conundrum that the artificial brain should be capable of expressing consciousness (e.g., [59, 96, 121]). In the spirit of searching for a functional correlate, we expand the thought experiment by considering a neuromorphic implant into the brain of a human infant suffering from a cortical disease, so that the normal motor, cognitive and perceptual functions can develop through a co-evolving chip, including awareness and consciousness. We call this chip a co-*e*volving *n*euromorphic twin (enTwin).

The neuromorphic blueprint for artificial consciousness allows us to propose an extension of the well-known Turing test for artificial intelligence, a test that has been superseded by recent developments in the field of AI. To go beyond previous proposals, we introduce a specific model of the neural and functional correlates of consciousness in biological brains, which we call the Conductor Model of Consciousness (CMoC), introducing an analogy between a meta-instance governing the information flow in the brain and the conductor of an orchestra. The conductor in this model represents a neuronal structure that gates cortical activity triggered from outside and inside the brain. It helps the developing subject to learn to distinguish between externally and self-generated mental constructs, and develop a notion of perception, sensory perception and proprio-perception, including awareness. The conductor is merely a distributed population of neurons involved in "teaching" the discrimination network. Although it may resemble the classical homunculus, it enters here in a pure mechanistic way as a class of neurons taking over specific functions within a developing network.

1.2 From creating artificial consciousness to the ethical dilemma

If conscious artificial agents develop cognitive abilities that rival or even surpass those of humans, paired with a form of consciousness, it becomes inevitable to consider granting them legal and political rights (e.g., [45, 62, 95, 111]). 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 be potentially able to experience pain and suffering (e.g., [3]). Such AI suffering would give ground to moral conflicts [94].

Introducing the enTwin and the CMoC gives us a handle for a new perspective on these ethical dilemmas. We consider the possible down-regulation or prevention of negative affective states (such as pain) in artificial agents, while still allowing the experience of positive ones and the possibility of empathy. 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. Figure 1 captures schematically the flow of arguments in this work.

1.3 Phenomenal consciousness

The terms "consciousness" is widely contested [40, 138], and in this paper the term will be concerned with phenomenal consciousness, which is considered a form of *state*-consciousness—i.e., a property attributed to certain mental states. If it "feels like something" [98] to be in certain mental state, this state is considered as being phenomenally conscious [25]. Therefore, phenomenal consciousness is often described as the subjective aspect of consciousness that involves experience [18, 31]. 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" another object, it is a mental state.

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 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, 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 can formulate or represent a basic reasoning of what their goals are, and by which actions they can be reached [13, 107].

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 a red object? 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 pain in your toes feel like?

2 Artificial consciousness at the dawn of the neuromorphic era

2.1 Re-evaluating criticisms against artificial consciousness

Sceptics with respect to artificial consciousness point out that the analogies between digital computers and human brains have many breaking points. Among the things that

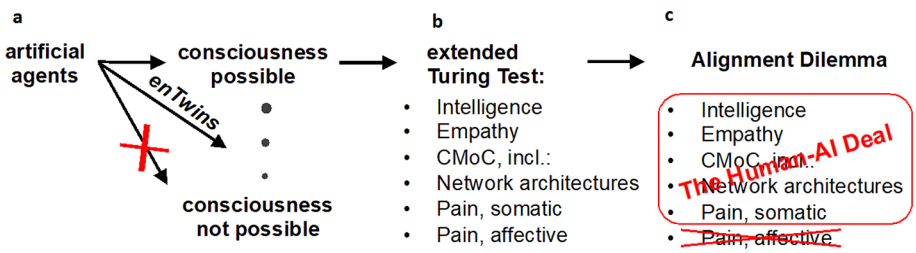

Fig. 1 Line of arguments on the possibility of artificial consciousness and how to deal with it. **a** The example of the evolving neuromorphic Twin (enTwin) shows the difficulties to exclude any form of consciousness. **b** To test for consciousness, we suggest an extended Turing Test that requires the identification of specific neuronal architectures described by the Conductor Model of Consciousness (CMoC) as functional and neuronal correlates of awareness (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

distinguish machines from (healthy) brains, the following appear to be the most relevant: (i) the lack of embodiment, i.e., of participation in the physical world, (ii) the lack of a centralised "I", (iii) the lack of evolutionary pressures and feedback, (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 and Tononi [48, 54, 55] 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, and associated notions such as causality. 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 by the machine users (humans). With respect to the 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, a key element of both the hardware and software. Likewise, brains do not appear to act like digital machines that run programs sequentially. There is massive parallelism of information processing in the brain, and the architecture of neurons is very different from digital technology.

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". 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 [142]. 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., [93, 106]).

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 (Sect. 3), 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 [67, 87]. Reengineering these fruits of evolution would turn the problem of a centralized "I" into a technological one, and no longer a matter of principle, in the sense that we have a blueprint (the conductor model) for its implementation. 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 kinds of limbs best adapted to bipedal locomotion, 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 [5, 68, 89, 117, 119]. 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., [105]), 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 [74] and to be effectively scalable [35]. On top of this, a lot of effort has been invested in emulating 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 [17, 61], something that will be relevant in Sect. 4 below.

Thus, neuromorphic hardware offers compelling solutions to the traditional objections concerning substrate dependency [48, 54, 55]. In these criticisms, it is not even possible to functionally replace a single neuron (let alone a brain) with an artificial counterpart, as the behaviour of carbon-based analogue neurons is too different from that of silicon-based chips. The next generation of flexible carbon neuromorphic substrates [46, 143, 144] could be moulded to emulate biological neurons to a degree that makes it very difficult to sustain any principled opposition to artificial neurons—or brains. In summary, recent techniques and developments have closed the door to most of the arguments against a principled impossibility of artificial consciousness. While the classical von Neumann computer architecture is arguably inadequate for emulating consciousness, this architecture is no longer the only game in town. In what follows we develop these ideas in detail.

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 [20]. Thought experiments are very popular in this domain and have introduced us to notions such as philosophical "zombies" [76], Chinese rooms [121], and Mary's lockdown room [69]. Our purpose is not to present a novel thought experiment. We instead *ground* existing thought experiments dealing with the feasibility of systems that closely emulate the human brain in many aspects that are relevant for consciousness, answering 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, as we describe in more detail below (see [28, 29, 96, 121]).

The classical neural replacement scenario of Morowitz [96] has been deemed implausible by authors such as [54–56, 114] 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 the more abstract philosophical insights about neural replacement scenarios.

Assume a human baby is born with a cerebral ataxia syndrome that is linked to cortical degeneration [36], resulting in motor disabilities including articulations and speech. Assume it is possible to help the child with an evolving neuromorphic twin (see Fig. 2). The enTwin is implemented in soft bioelectronic interfaces that can be implanted in human bodies [90], and even in human brains [142]. 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 [14]. For motor and speech assistance it is coupled with muscle stimulation devices. 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 modulate 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 [58], 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

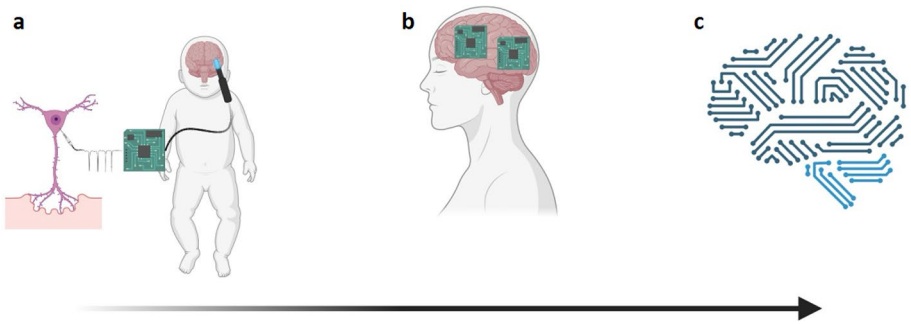

Fig. 2 The evolving neuromorphic twin (enTwin) thought experiment. **a** A neuromorphic chip, able to communicate with biological neurons, is used to help an infant to achieve normal sensorimotor functions. The neuromorphic chip can be implanted inside the body and brain of a human and learns to grow and adjust its synapses in the same way as biological neurons do. These chips are built of soft bioelectronic interfaces. **b** With time, the neuromorphic chips become 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. **c** By training many such systems and integrated from different human patients, 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.

articulation, (iv) thalamic and brainstem kernels to represent different wakefulness states [57, 102]. Consequently, an enTwin would mimic the one example we know where consciousness developed (humans and other 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 of 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, and to infer possible criteria for neural correlates of consciousness, 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 [120], or [129]), 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 [124], we argue that the brain contains the crucial ingredients to implement a form of Generative Adversarial Networks (GANs, [60]). GANs have proven to be cornerstones of powerful network architectures for image recognition, language processing, and translations of image to language [4]. Likely, generative networks are implicated in mental imagery, and discriminative networks must then 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 between G, E and D, and the type of synaptic plasticity within these networks (plasticity on, off, or inverted, see [44]). Based on the feedback from the discriminator network (that may reveal the fake/imagined 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

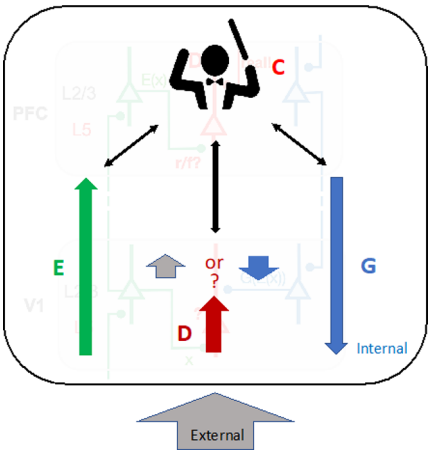

Fig. 3 The Conductor Model of Consciousness (CMoC): The implementation of elaborate forms of learning requires a network instance that organizes the flow of information to keep teacher and student signals apart. Possible ingredients for consciousness to evolve in our enTwin. An encoding (E) and generative (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 network (C) selects the contents of the encoding and generative networks that matches and broadcasts this for further processing

in the human brain [58] and support creative dreaming during rapid eye movement sleep (REM) sleep [44].

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 (the conductor network C) 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 [137]. Reality monitoring areas [124] are part of the cortical GANs as suggested in Gershman [58] 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 signals 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 of a conductor resembles the functionality of prefrontal and anterior cingulate areas [124], but it may also be taken over by the gating mechanisms of cortico-thalamic loops via higher order thalamic kernels [131, 141], as elaborated below. When acting on the visual stream, the conductor may signal "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 signal: "this activity represents myself and is generated from inside" (see Fig. 3).

3.1 The CMoC extends the Helmholtz view of perception by creative processing

Originating from a computational model to improve the cortical representation of sensory signals [43], the conductor module can be seen as an evolutionary product of actively generating synthetic sensory activity and discriminating this against real sensory input. Awareness, in this view, arises as by-product, sparking off from the need of a meta-level structure that teaches the distinction between different states of sensory activity. Figure 4 shows a progression of ideas on perception and awareness. Helmholz's insights on the nature of perception (see, e.g., [64] [original work published 1867]) have served as a guiding light for neuroscience, and they stay relevant to this day [27]. The modern theory of predictive coding, including the ideas of [33, 34, 52, 127, 128] can be seen as refinements and extensions of Helmholtz's active sensing and inference. The conductor model takes these building blocks and adds additional structure in the form of a GAN architecture and the conductor module. The encoder and generator (E, G) are part of the formalization captured by the Helmholtz machine [39], extending the unidirectional flow from the objects to representation by a generator from the representation to sensory activity. Helmholtz machines are able to extract semantic structures in sensory data by trying to recreate sensory data from the internal representation. The CMoC also does this, but adds specific structures that emulate the way perception and awareness work phenomenologically (Fig. 4c, d). Through the discriminator and the adversarial learning the generator is able to creatively produce new sensory activities that potentially integrates in the reality, a procedure that we call *creative coding* to emphasis the step beyond predictive coding. With creative coding (or more general *creative processing*) comes the necessity for additional metastructures in the brain that may produce phenoma akin to awareness [44].

Functional correlates of awareness within the CMoC can be drawn on multiple levels, captured by a mapping of conductor properties to awareness properties (Fig. 4d):

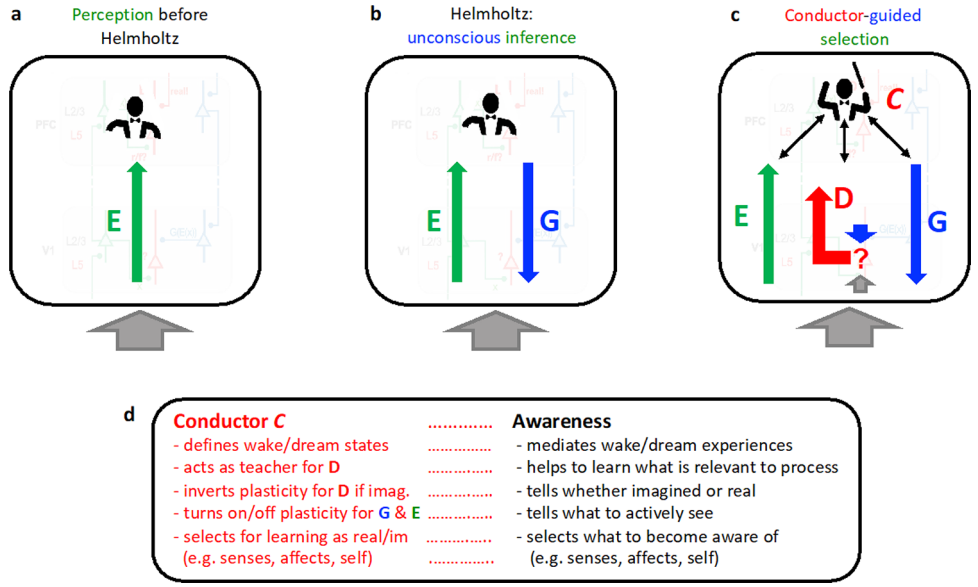

Fig. 4 The Conductor Model of Consciousness (CMoC) extends the functional catalogue for predictive coding and perception by a module for creative coding and awareness. **a** The intuition that perception emerges (symbolized by a homunculus) form processing sensory input at the end of an encoding pathway (E). **b** Helmholtz (building upon ideas of Kant) argued that perception is the product of an active inference that involves predictions of what is expected to be sensed (the generator G), even when we are unconscious of this inference process. Predictive processing, with the representation of the self (sketch of human) that monitors the outcome of its own actions, and prediction error broadcasted to a global neuronal workspace, remains the key ingredient of current theories of consciousness (e.g., [33, 34, 91, 126, 129]). **c** The CMoC adds an adversarial architecture to the previously suggested hierarchy of active inference, with an additional function of a conductor that orchestrates the adversarial learning and creative—as opposed to only predictive—processing. **d** The Conductor represents a functional correlate of awareness, emphasizing the structural mapping from the CMoC to the phenomenal consciousness (here restricted to awareness). The specific functions within the CMoC generate predictions on a possible neural or functional correlate of consciousness/awareness.

(i) the conductor signalling the discriminator whether sensory activity has to be judged as real or imagined, versus awareness telling the subject to experience sensory activity as real or imagined; (ii) the conductor acting as a teacher for the discriminator network D, versus awareness directing the learning to specific contents. (iii) The conductor modulating plasticity depending on whether sensory activity is produced from inside or outside, versus awareness informing the agent about the sensory activity being real or imagined, and finally (iv) the conductor selecting contents from the different levels of the encoder hierarchy to be learned as real versus imagined (including sensation, and to global states such as affects or the self). The analogy between the conductor and awareness can be extended to distinguish the various states of awareness, namely (a) wakefulness, (b) sleep, (c) dreaming and (d) lucid dreaming [44]. While a match between the conductor and discriminator (C↔D) indicates state-awareness, a match between the generator and encoder (G↔E) indicates content-awareness (see the neuronal implementation below).

3.2 The conductor as a teacher to learn organizing the inner world of autonomous agents

With the functional and neuronal criteria of the CMoC, we can render the architectural constraints of implementing phenomenal consciousness in a more precise way. In line with other suggestions (e.g. [30]), 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 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.

(CMoC-2) has a mechanism—the conductor—that allows to tell whether the agent acts on the internal or external world representation.
(CMoC-3) is equipped with its own internal sense of self associated to the conductor, modulated by global affective components.

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 (the capacity to sense the internal state of the body) can be seen as part of the internal world that goes beyond an external world model. 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 in terms of meta-information: beside informing the content of the internal or external model, respectively, the addition also signals "who" generates this content, and how it should be processed. This is the function assigned to the conductor network, an instance on top of the internal model of the outer world.

Apart from the interface at the sensory areas, the adversarial learning mechanism to create the inner world also goes beyond latent representations of the outer world and surfing in the inner world [33, 34]. In humans, it may be the factor driving genuine innovations in culture and other areas that are not only novel, but also useful in that they integrate into the existing world– a feature that was internally tested by the discriminator in the CMoC. Consciousness is not only about enabling an active sensing of the environment by means of actions and predictions, and not only about creating self-awareness. It is in the first place about offering a neuronal and functional infrastructure for learning to disentangle inside from outside triggered brain activity, while at the same time trying to match these activities. We postulate that the adversarial learning to create an inner world model (potentially including interoception) along these principles does equally exist for other animals. Adversarial learning comes along with a conductor module that labels information related to the internal and external world and governs the information flow between the representations of these worlds (CMoC-2). The conductor can also select and prioritize some sensory information over others and has the power to impose a state of emergency (CMoC-3). The module might use a short-cut circuit to avoid harm that we can associate with experiences such as sensing pain.

Mental scenarios we can think of may 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 in possibilities and structure than the external world. This richness goes beyond the mere ability of mind wandering and counterfactual reasoning that include one's own actions (see e.g. [53]). The CMoC introduces the neuronal learning apparatus that *then* allows for mind wandering and the like. Providing the substrate for learning the meta-structures is the crucial addition here. The hypothesis underlying the CMoC is that with the learning apparatus for the state-distinctions, that includes providing a learning signal for differentiating "real" from "imagined", also comes a novel experience for the agent to become aware of "real" or "imagined". The teaching signal of the conductor is more effectful if it is dominant, which is particularly important when learning the state of life-threatening affects.

3.3 Consciousness as conductor-mediated private experience emerging from functionality

The cortical conductor allows us to further circle on 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—a sub-module in the network that integrates information from the whole. This conductor module is not identified with the agent itself that may have an additional embodiment and is neither identified with the representation of the self. The self may be placed at the top of the encoder hierarchy (E), out of which actions are generated (G). The conductor C instead is a meta-instance that organizes the information flow, including the information flow from and towards the representation of the self. It plays a central role for its owner, the agent, and may ground higher order self-awareness. The conductor may manifest as a very private sense that represents a kind of sensory modality for the owner's inner world, be it the awareness of a stimulus, or the awareness of the self. The conductor-mediated inner sense only emerges and exists within this individual, is not accessible from the external world, and in fact disappears when seen from the external physical world.

To provide another structural analogy showing how an additional ontological dimension may emerge within an inner 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", in analogy to the "quale"), 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 the only relationship required to build a theory of complex numbers. The ontological dimension of *i* dissolves within the larger embedding space of complex numbers, where both real and imaginary numbers are simply elements of the wider set of complex numbers, mathematically characterized as a field. 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. The imaginary and real numbers are both independent dimensions of complex numbers. Extending the physical world by consciousness could make this "complete" as well, with physics and consciousness as independent dimensions.

The main point of the analogy is that the "imaginary existence" emerges from pure functionality, here the functionality of solving algebraic equations. It is a private feature of imaginary numbers as opposed to real numbers but disappears in the wider and abstract perspective of complex numbers. Likewise, consciousness may emerge from the functionality of learning to internally produce sensory activity as close as possible to the one externally produced.

3.4 The neuronal implementation of the CMoC includes state- and content-awareness

We next show how the CMoC can be implemented in neuronal structures. CMoC postulates the existence of an encoding, generative, and discriminator network, together with a conductor module that orchestrates the information flow among them, turns on and off plasticity, and determines which information should be considered as real or imagined. 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 [44]. During an adversarial REM dream, conductor neurons adversarially tell the discriminator neurons what they see is "real", giving the dreamer the incorrect feeling of experiencing reality. The functional reason for the mistaken reality felling is to test the dream against reality. When dreaming of an approaching lion, we should not learn to go caressing him, but instead learn to hide on a save place. Technically, when imposing the reality-target to the discriminator that itself would have demasked the sensory activity as dreamed, an error signal is produced at the output of the discriminator network. This error is backpropagated to the generator network, telling this where to improve the generated sensory activity so that the discriminator the next time in fact will judge it as real. The error-backpropagation itself can be implemented in neuronal terms [122]. Hence, the adversarial teaching is hijacking the error-backpropagation circuitry of the discriminator to provide the generator a helpful indication how to generate more reality-like activity.

In order to also allow the discriminator to improve its job of correctly telling internally or externally induced activity apart, its plasticity needs to be inverted while being given the adversarial target "real" during the dream (Fig. 4d). Plasticity of the generator network in the REM dream keeps its original sign so that it can in fact correct for the error delivered by the discriminator. The encoder may turn its plasticity off as it is not provided helpful information. During wakefulness it is the other way round: plasticity in the generator is turned off, but plasticity in the encoder is turned on [44].

The neurons in the conductor module represent the meta-state about how sensory activity should be interpreted by the network ("real" or "imagined") and are thus candidates to also mediate becoming aware of the "real" or "imagined" state. This state-awareness in the model is triggered by a match between the conductor signal C and the discriminator signal D within a L5 pyramidal neurons (C↔D), thought to elicit a calcium spike in their apical dendrites (Fig. 5, b1, see also [131]).

While discriminator neurons represent state-awareness, other layer 5 pyramidal neurons of the visual stream are representing the content-awareness. These are the neurons that detect 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 (G↔E, see Fig. 5, b1). In the primary visual cortex (V1), for instance, the generator may predict an edge that is also present in the image, and hence a calcium spike in the corresponding edge-detecting neuron is elicited. In a higher visual area that responds to faces, for instance [2], the generator may predict my face when I look into the mirror, and a neuron responding to my face will elicit a dendritic calcium spike as it sees my face. This neuronal implementation of the CMoC works just as postulated in dendritic integration theory (DIT, [7, 8]), just that the CMoC also allows for becoming aware of the state to be real or imagined, beside becoming aware of the content.

Figure 5b shows examples of images presented to the sensory area (x, green), what is expected to be seen after the unconscious inference step (G(E(x)), blue, what the Helmholtz-model would tell, see also Fig. 4b), and what the L5 pyramidal neurons make us aware to see in the CMoC

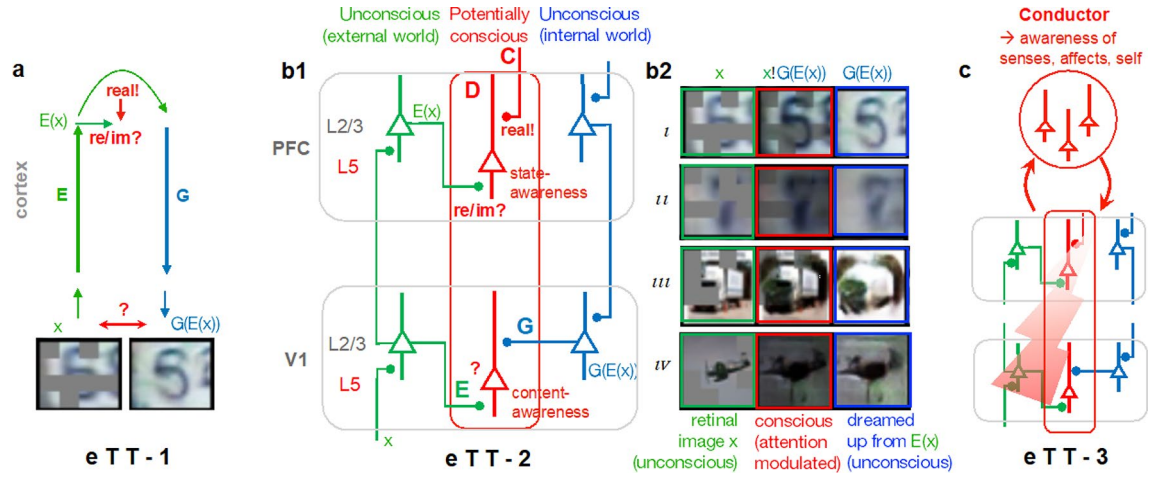

Fig. 5 Circuit criteria of the extended Turing test (eTT1-3) for consciousness. **a** An encoding (E) and generative (G) network. Here, E encodes a partially occluded image x in a higher cortical area, E(x), out of which the generative network produces a non-occluded version, G(E(x)). Simulations performed by N. Deperrois based on the model in Deperrois et al. [43]. "re/im?" stands for "real/imagined?". **b1** A discriminative network (D, red) that may convey the awareness of the stimulus. Here, the conductor C teaches the discriminator whether cortical activity should be considered as real (i.e., from the external world) or imagined, i.e., generated from the internal world via G (and then E). Layer 5 pyramidal neurons at the top area represent meta-information such as "a real image" (state-awareness), and at lower areas, such as V1, may signal 'an edge' (content-awareness). The conscious percept can be modelled as apical gain modulation of the basal input to the L5 pyramidal neurons (red) from the E network, and the local attentional signal from the G network. **b2** Examples of noisy images (x, green squares) and G(E(x)), the cleaned-up version of x after turning through the central areas up with E and down with G. Both activity streams, the encoder E and generator G, do not enter our consciousness. What becomes conscious is the product x · G(E(x), images in red squares, forming the attention-modulated input and being represented by a specific class of layer 5 pyramidal neurons (red, in **b1**, see also [8, 131]). **c** The cortical conductor gates the information flow of the conscious stream (red), whatever is represented in there (sensory or affective components, the self), acting also as a `door keeper' for a global workspace of consciousness. The affective pain component within the global workspace captures an "existential threat". A strong apical drive from the conductor makes the reality impression from the corresponding input dominant ("the input is absolutely real"), suppressing other inputs from awareness and potentially triggering a "survival response" (red flash).

(x·G(E(x)), red). These L5 pyramidal neurons receive the forward input x on their basal dendrites, while the top-down input G(E(x)) projects to the apical dendrites. The somatic activity represents the gain-modulated sensory input x, multiplicatively modulated by the top-down input (as experimentally described, see [84]).

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 [124], perhaps jointly with the gating mechanisms via higher-order thalamic kernels [131]. 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 [73], may push some content into consciousness by facilitating dendritic calcium spikes (Fig. 5c, see [140]). These contents represent sensory features when referring to sensory areas, but they may also represent higher order features such as affects or the self when referring to other cortical regions such as the prefrontal or cingulate cortex [22].

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

Following Seth and Bayne [120], we can divide theories of consciousness (ToCs) among four broad classes (see also [24]): *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 tries to axiomatize consciousness based solely on the statistical notion of information and complexity; 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 [50, 58, 85, 86] 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 [12, 41], Dehaene and Changeux [42, 91].

An active conductor gates the recurrent processing, and likely modulates the complexity of the neuronal activity patterns during consciousness. The recurrent processing between the encoder and generator of the CMoC directly relates to *recurrent processing theory of consciousness* [83, 129]. It also relates to integrated information theory (IIT, [92, 134]) that exploits the recurrencies to generate complexities in brain activity, leading to clinical measures of the levels of consciousness [26]. Yet, our approach does not build on an abstract notion of information, although there is of course information flow in the encoder, generator and discriminator network, and in the conductor module. Instead, the CMoC focusses on function and content, and their organization across hierarchies. The use of generative models connects the CMoC with predictive processing theories [33, 34, 65]. In fact, the principles behind the CMoC stem from studies of predictive processing within neurons [75, 84, 122, 137]. The specific implementation of the CMoC in the neuronal circuitries closely follows the ideas of dendritic integration theory (DIT, [8]) and may be seen as an extension of DIT to include state-awareness beside content-awareness (see Fig. 6 and [44]).

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 [107, 109]. Neurorepresentationalism also takes predictive processing as a theoretical building block, but unlike Active Inference Theory (Hohwy, Clark, Friston—see above), it is primarily sensory-based, and relies on multimodal integration. 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. 6).

3.6 An extended Turing Test (eTT) for consciousness including functional and neuronal correlates

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 involved in cognitive processes. To specify these components, we can 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 [51]). 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 function and implementation of the circuitries that make up its "brain". Granted, the original Turing test examines if an artificial agent "thinks like" a human rather than establishing the existence of phenomenal consciousness. But given its very clear failure in this aim, what we want to test goes deeper and at the same time a bit parallel to the question of intelligence.

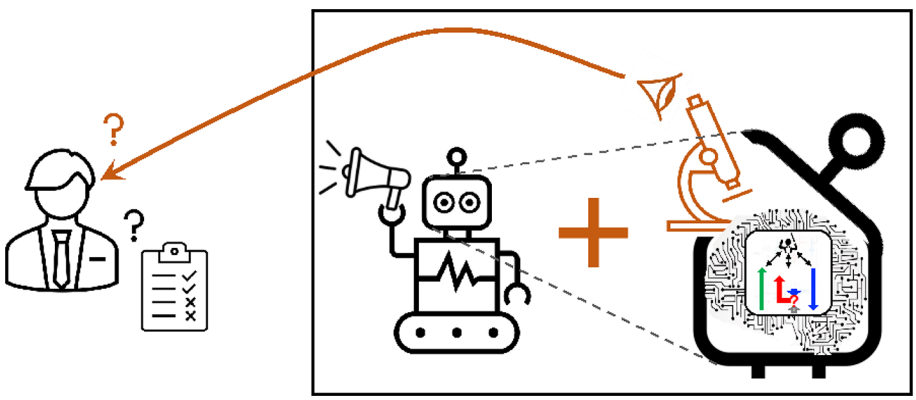

Fig. 6 The extended Turing test (eTT). A list of criteria to be satisfied is indicative for the presence of some form of consciousness. The list extends the items of the classical Turing test for intelligence. It requires the observer to enter the "Chinese room", open the box and identify the postulated 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. The "neuromorphic correlate of artificial consciousness" is required to fulfil functional criteria as, e.g., listed in Fig. 4d.

We call our proposal the extended Turing test (eTT): on top of analysing the behaviour of the agent and checking 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. 5). 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 within it. Consequently, this is a more stringent test than the classical Turing test and relates to ideas of neurorepresentationalism on consciousness (see [106, 108], for similar ideas).

Passing the eTT does not necessarily imply the emergence of the phenomenology of consciousness– as in each test, False Positives may occur. But the eTT could also be considered *too* stringent a test, producing False Negatives. Some eTT criteria may turn out to not be necessary. For instance, 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 of consciousness per se.

The eTT may be organized as a layered list of requirements. At the basal 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 (Fig. 6):

(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 as being 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 and Carvalho [38] emphasizes the need for representing sensory inputs and imagined contents. Solms and Friston [126], Solms 127 formulate similar criteria in the context of predictive coding and active inference, LeDoux and Brown [88] in the context of emotions. Other works have coupled predictive coding networks to planning of complex, goal-directed behaviours [108]. Dehaene et al. [41], Dehaene and Changeux [42] 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 [51] consist in the availability of a specific, neuroscience-inspired model of the architectural requirements behind consciousness, for instance 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.

4 The ethics of dealing with conscious AI: hints from the CMoC

We introduced the conductor as a network that provides the teaching signal for reality judgements. These judgements may refer to local or global sensory features, or to proprioceptive signals that may be real or imagined. Affective states may equally be learned as real. In fact, in the same way as sensory states are learned to be generated from inside during adversarial dreams and are learned to be assigned a reality-label [43, 44], we postulate that also affective states are learned to be assigned such a label. The reality-label with respect to global affective states is claimed to give us the conscious experience of a feeling, and the reality-label with respect to the self is claimed to give us the conscious experience of being ourselves.

Here we apply these insights from the CMoC on the differentiation between sensory and affective conscious states to the ethical question how we should construct putatively conscious agents. We particularly ask how the insights may help to organize and stabilize the coexistence of artificial and human agents with unequal cognitive and mental skills.

4.1 Ethical issues of creating artificial consciousness

Assuming that advances in neuromorphic engineering lead to the emergence of conscious artificial agents, and given that the proposed eTT and CMoC allow 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 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 [16], the Asilomar Conference on Beneficial AI [10] formulated 23 principles for ethical AI research. Some of these principles are condensed in the axioms for "provably beneficial AI" [118]. Thirty years after the international agreement on recombinant DNA, the United Nations Declaration on Human Cloning [136] was formulated, preceded by the European Parliament Resolution on Human Cloning [47], 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 [45, 62, 78, 95, 97, 111]. Beside possible existential threats accompanying strong AI [116], 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) approach the spectrum of human 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 further disadvantage already suffering human 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. An alignment of values is also desirable for obvious safety reasons [116]. Reciprocally, the notion of alignment between our values and that of future artificial agents might hang 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, but when dealing with this new class of conscious agents, it might very well be in our own benefit (as well as morally sound) to treat them as part of a commonwealth of moral beings.

The intentional design of human-like conscious artificial agents, say following the CMoC, evokes an intrinsic alignment dilemma. On the one hand, as agents might potentially surpass humans in many defining features—such as intelligence and knowledge—humans risk attaining a disadvantageous mismatch between our rights as the creators of these machines, and their moral rights as conscious, intelligent and emotional beings. Even though some voices in the scientific community find no issue with the idea of creating "improved" replacements of humans or even humanity, this rings like a hubristic platitude and does not sit well with most humans. On the other hand, if we one-sidedly prevent artificial intelligence, knowledge or even empathy from being developed in order to preserve our privileged status, we risk trampling over the moral rights of possibly sentient 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 the symmetry in another dimension, for example by adding some extra weight to human suffering with respect to agent suffering (or prevent this suffering in the first place; see Figs. 1c, 6).

Here is where our CMoC with its eTT comes into play, and in particular the distinction between the sensory and affective components of pain. 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 create conscious artificial agents aligned with humans?

In humans, and likely also other sentient agents, pain is needed for adaptive behaviour and learning. However, this is not necessarily the case for artificial agents, and specifically for the affective component of pain. Although pain in general may be an important factor with regards to the development of empathy (see Discussion below), pain may be a "feature" from which we could relieve artificial agents. While the general strategy for developing conscious agents is to emulate the fruits of biological evolution, we might want to omit some parts 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 moderate the affective dimension of chronic pain or of pain in general within artificial agents without losing much of other functionalities, while ensuring that they cannot suffer as much as humans and other animals.

Looking at the human brain, we see that the sensory and affective components of pain are represented in separate neuronal circuits and nuclei [19, 23, 113]. 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 Sect. 3. Instead of the eTT-3 criterion, we suggest the weaker test criterion by replacing the "affective component" with a "sensory and cognitive component":

> (eTT-3^{--}) A global *sensory and 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 *cognitive 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 nociceptive signals also offers a handle to ethically justify an asymmetry in the rights for human and artificial agents. Such agents would sense real pleasure but only a cognitive recognition of pain, or at least to lesser degree than humans do.

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, as empathy is grounded on the recognition of suffering in others (see discussion below in Sect. 5). A version would be to only preclude the affective of 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 (discussed below). This needs to be cognitively compensated and may become part of the ongoing deal we consider next.

4.3 Trading rights against affects: a possible 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 for not suffering 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. 7). 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 rules for interacting 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." [70].

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 renegotiate this deal, while humans are allowed to reconsider it if getting out of hand. In fact, since sentient agents do not exist yet, it is first us humans that must agree on a roadmap how to design them and organize our co-existence. It is possible that future agents, shaped by overt or covert pressure, 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. Some of the agents, free of affective components of pain, although endowed with a symbolic representation of this component, may express the desire to encompass more. Others may not even accept the deal, and the discussion must be intensified on how to prevent the risk of AI getting out of control, either overtly or in more subtle, hidden ways [15]. Given the many ramifications in today's debate on regulating AI, it is helpful to consider a realistic future scenario that judges the possibility of sentient agents on a wider scientific ground. The CMoC merges considerations on the function and the substrate of artificial consciousness, differentiates between the awareness of sensations, of affects and of the self, and offers nuanced propositions to keep the various dimensions of consciousness apart in artificial agents. The human-AI deal represents an example of how to integrate these insights into a debate about shaping future sentient agents and our relationship to them.

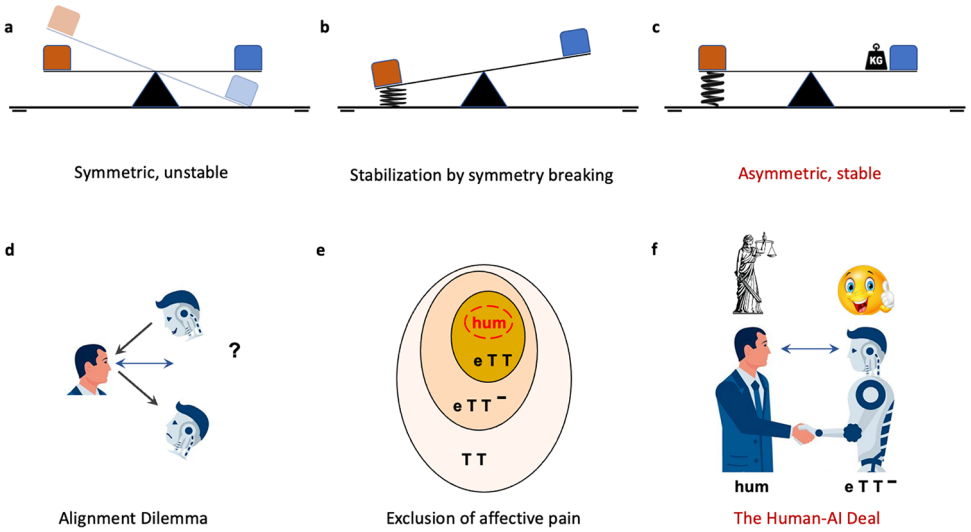

Fig. 7 **a**–**c** The Alignment Dilemma as unstable equilibrium. **a** In a physical system in unstable equilibrium, such as a seesaw mechanism, the system is unstable under changes in the relative weights of both arms. **b**, **c** The situation can be solved by breaking the symmetry of the system 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**–**c**) thus model situations (**d**–**f**). **d**–**f** Addressing the Alignment Dilemma. **d** If artificial sentience is possible, agents may develop a higher than human degree of 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 derived from the CMoC, 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.

4.4 Must pain hurt? A philosophical perspective

AI agents are not biologically evolved beings, but instead designed to emulate biological entities. 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 sensitivity to suffering. Consequently, we would end up in a situation which is in some ways opposite 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 designed 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, considering moral rights to come in degrees instead of by crossing a certain threshold. This would still give humans a degree privilege, if we decide the alternative to be unacceptable. What would give us rights that are a bit above those of artificial agents is our specific mix of sentience, intelligence, and capacity for suffering, and not any single absolutely demarcating characteristic.

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., [123]), 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 (such as the example of maximising happiness by having a maximally sized population of unhappy individuals, see [104, 125]). 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 [101], but this focus leads to what have been seen as unsatisfactory moral postures when considering the rights of animals, children, people with mental disabilities, or the uneducated, for instance. 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 and Nussbaum [130]). Our position here is minimalist in that most moral philosophers would agree our precepts 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 [1, 6, 66, 71, 81]).

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 affects to the point that agents lack the capacity to understand and be repulsed by suffering, they could just turn into highly functional psychopaths, and an alignment of values would be impossible. The dis-alignment of values may increase the competition between humans and machines, and this is what we want to prevent. We therefore need to ensure some ethical alignment first, even if these agents do not share some affective components of pain or 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 [115]. 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 [49].

4.5 Affective versus sensory components of pain: a physiological perspective

It is generally accepted that pain features show two largely distinct dimensions (e.g., see [11, 112]). 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 [132]. 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 [63, 72, 80, 113]. 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., [79]). Indeed, it is usually assumed that the affective component is crucial for the motivational aspect of pain—it is what makes us to learn and take protective action (e.g., [103, 132]). 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., [77, 99]). 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. Central to our argument, the relatively speed advantage with which neuromorphic hardware works against biological neurons make these arguments less compelling in the case of machines. It is entirely plausible that artificial agents can quickly react to the sensorial information of impending harm, without the need for an emergency shortcut system—and the evolutionarily associated feeling of pain.

The CMoC provides a fresh perspective 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.

5 Conclusions

We have introduced the Conductor Model of Consciousness (CMoC) as an integrated framework for considering functional and neuronal correlates of consciousness, including ethical implications. First, we argued that, by means of a co-evolving neuromorphic twin (enTwin) implanted as a medical aid in an infant brain, the technical possibility of artificial agents reaching some human levels of consciousness is

conceivable. The version of the classical thought experiments replies to the usual criticisms against artificial consciousness related to the missing substrate, embodiment, or evolution.

Second, we considered a structural mapping of phenomenal consciousness to a model of consciousness that captures some functional and neuronal correlates. The CMoC expresses the need for a meta-instance, the conductor, involved in the capacity of the brain to learn and create an internal model of the world and—going beyond the existing world—includes the creation of novel concepts, entailing the concept of itself. At the cortical periphery, the internal model generates actions and sensory predictions consistent with the external world, while in more central parts, it represents affects and higher concepts of the outer or inner world, such as the self. The CMoC provides grounds for a refinement of the Turing test, our extended Turing test (eTT), designed to predict forms of neuronal circuitries required for conscious experiences. An artificial agent passing the eTT, based on the functional and neuronal similarities to the biological example, increases the likelihood that it developed some form of consciousness.

Third, we explored the ethical implications of the refined model of consciousness. We described the alignment dilemma between sentient agents and humans in the light of the CMoC. By dissecting conscious sensations into individual components, such as affective, sensory and cognitive components, the model offers a handle to determine the type of consciousness potentially realized in future sentient agents. These options touch upon ethical guidelines for the design of sociable agents sharing human values such as empathy, despite their cognitive abilities that are about to surpass ours, and that may not only be employed in our favour.

To tame the risk of AI growing out of control we suggested a human-AI deal, setting a primacy for humans on their essential rights, while in turn offering sentient agents to be relieved from affective chronic pain. The deal is intentionally asymmetric: on the one hand, agents sharing potential features of consciousness still need to be designed by us; on the other hand, asymmetry in general helps to stabilize a dynamically regulated equilibrium. The human primacy aims at building an ethical barrier to protect the majority of humans being left out from benefits brought by AI, while an optional relief from negative affects for sentient agents is not expected to impede their successful integration into our physical and social world.

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.

Funding Open access funding provided by University of Bern

Declarations

Conflict of interest The authors declare there are no conflict of interest.

References

1. Aaltola, E.: Empathy, intersubjectivity, and animal philosophy. Environ. Philos. **10**(2), 75–96 (2013)
2. Afraz, S.R., Kiani, R., Esteky, H.: Microstimulation of inferotemporal cortex influences face categorization. Nature **442**(7103), 692–695 (2006). https://doi.org/10.1038/nature04982
3. Agarwal, A., Edelman, S.: Functionally effective conscious AI without suffering. J. Artif. Intell. Conscious. **7**(01), 39–50 (2020)
4. Aggarwal, A., Mittal, M., Battineni, G.: Generative adversarial network: an overview of theory and applications. Int. J. Inf. Manag. Data Insights **1**(1), 100004 (2021). https://doi.org/10.1016/j.jjimei.2020.100004
5. Amunts, K., Axer, M., Banerjee, S., Bitsch, L., Bjaalie, J.G., Brauner, P., Brovelli, A., et al.: The coming decade of digital brain research: A vision for neuroscience at the intersection of technology and computing. Imaging Neurosci. **2**, 1-35 (2024). https://doi.org/10.1162/imag_a_00137
6. Angell, J.R.: The affective elements of consciousness. In: Chapter 13 in Psychology: An Introductory Study of the Structure and Function of Human Conscious, 3rd edn., pp. 256–269. Henry Holt and Company, New York (1906). https://doi.org/10.1016/j.tics.2020.07.006
7. Aru, J., Suzuki, M., Rutiku, R., Larkum, M.E., Bachmann, T.: Coupling the state and contents of consciousness. Front. Syst. Neurosci. **13**(August), 1–9 (2019). https://doi.org/10.3389/fnsys.2019.00043
8. Aru, J., Suzuki, M., Larkum, M.E.: Cellular mechanisms of conscious processing. Trends Cognit. Sci. **24**(10), 814–825 (2020). https://doi.org/10.1016/j.tics.2020.07.006
9. Aru, J., Larkum, M.E., Shine, J.M.: The feasibility of artificial consciousness through the lens of neuroscience. Trends Neurosci. **46**(12), 1008–1017 (2023). https://doi.org/10.1016/j.tins.2023.09.009

10. 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/
11. Auvray, M., Myin, E., Spence, C.: The sensory-discriminative and affective-motivational aspects of pain. Neurosci. Biobehav. Rev. **34**(2), 214–223 (2010)
12. Baars, B.J.: A Cognitive Theory of Consciousness. Cambridge University Press, Cambridge (1988)
13. Balleine, B.W., Dickinson, A.: Goal-directed instrumental action: contingency and incentive learning and their cortical substrates. Neuropharmacology **37**(4–5), 407–419 (1998). https://doi.org/10.1016/s0028-3908(98)00033-1
14. Bartolozzi, C., Indiveri, G., Donati, E.: Embodied neuromorphic intelligence. Nat. Commun. **13**(1), 1–14 (2022). https://doi.org/10.1038/s41467-022-28487-2
15. Bengio, Y., Hinton, G., Yao, A., Song, D., Abbeel, P., Darrell, T., Harari, Y.N., Zhang, Y.-Q., Xue, L., Shalev-Shwartz, S., Hadfield, G., Clune, J., Maharaj, T., Hutter, F.: Managing extreme AI risks amid rapid progress. Science **384**, 843–845 (2024). https://doi.org/10.1126/science.adn0117
16. Berg, P.: Summary statement of the Asilomar Conference on recombinant DNA molecules (1975). 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
17. Billaudelle, S., Stradmann, Y., Schreiber, K., Cramer, B., Baumbach, A., Dold, D., Meier, K.: Versatile emulation of spiking neural networks on an accelerated neuromorphic substrate. In: 2020 IEEE International Symposium on Circuits and Systems (ISCAS). IEEE, pp. 1-5 (2020)
18. Block, N.: On a confusion about a function of consciousness. Behav. Brain Sci. **18**(2), 227–247 (1995)
19. Boccard, S.G., et al.: Targeting the affective component of chronic pain: a case series of deep brain stimulation of the anterior cingulate cortex. Neurosurgery **74**(6), 628–635 (2014). https://doi.org/10.1227/NEU.0000000000000321
20. Brown, J.R.: Thought experiments. Can. J. Philos. **25**(1), 135–142 (1995)
21. Bubeck, S., Chandrasekaran, V., Eldan, R., Gehrke, J., Horvitz, E., Kamar, E., Lee, P., Lee, Y.T., Li, Y., Lundberg, S., Nori, H.: Sparks of artificial general intelligence: early experiments with gpt-4. arXiv preprint at arXiv:2303.12712 (2023)
22. Bush, G., Luu, P., Posner, M.I.: Cognitive and emotional influences in anterior cingulate cortex. Trends Cognit. Sci. **4**(6), 215–222 (2000). https://doi.org/10.1016/S1364-6613(00)01483-2
23. Bushnell, M.C., Čeko, M., Low, L.A.: Cognitive and emotional control of pain and its disruption in chronic pain. Nat. Rev. Neurosci. **14**(7), 502–511 (2013). https://doi.org/10.1038/nrn3516
24. Butlin, P., Long, R., Elmoznino, E., Bengio, Y., Birch, J., Constant, A., VanRullen, R.: Consciousness in artificial intelligence: insights from the science of consciousness. arXiv preprint at arXiv:2308.08708 (2023)
25. Carruthers, P.: Consciousness: explaining the phenomena. R. Inst. Philos. Suppl. **49**, 61–85 (2001)
26. 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.: A theoretically based index of consciousness independent of sensory processing and behavior. Sci. Transl. Med. **5**(198), 198ra105 (2013). https://doi.org/10.1126/scitranslmed.3006294
27. Caucheteux, C., Gramfort, A., King, J.R.: Evidence of a predictive coding hierarchy in the human brain listening to speech. Nat. Hum. Behav. **7**(3), 430–441 (2023). https://doi.org/10.1038/s41562-022-01516-2
28. Chalmers, D.J.: Facing up to the problem of consciousness. J. Conscious. Stud. **2**(3), 200–219 (1995). https://doi.org/10.31812/apd.v0i14.1838
29. Chalmers, D.J.: Absent qualia, fading qualia, dancing qualia. In: Metzinger, T. (ed.) Conscious Experience, pp. 309–328. Ferdinand Schoningh, Paderborn (1995b)
30. Chalmers, D.J.: How can we construct a science of consciousness? Ann. N. Y. Acad. Sci. **1303**(1), 25–35 (2013). https://doi.org/10.1111/nyas.12166
31. Chalmers, D.: Idealism and the mind-body problem. In: Seager, W. (ed.) The Routledge Handbook of Panpsychism, pp. 353–373. Routledge, London (2019)
32. Cheatham, B., Javanmardian, K., Samandari, H.: Confronting the risks of artificial intelligence. McKinsey Q. **2**(38), 1–9 (2019)
33. Clark, A.: Whatever next? Predictive brains, situated agents, and the future of cognitive science. Behav. Brain Sci. **36**(3), 181–204 (2013a)
34. Clark, A.: Expecting the world: perception, prediction, and the origins of human knowledge. J. Philos. **110**(9), 469–496 (2013b)
35. Cointe, C., Laborde, A., Nowak, L.G., Arvanitis, D.N., Bourrier, D., Bergaud, C., Maziz, A.: Scalable batch fabrication of ultrathin flexible neural probes using a bioresorbable silk layer. Microsyst. Nanoeng. **8**(1), 21 (2022). https://doi.org/10.1038/s41378-022-00353-7
36. Conrad, J., Huppert, A., Ruehl, R.M., Wuehr, M., Schniepp, R., Zu Eulenburg, P.: Disability in cerebellar ataxia syndromes is linked to cortical degeneration. J. Neurol. **270**(11), 5449–5460 (2023). https://doi.org/10.1007/s00415-023-11859-z
37. Cowls, J., Tsamados, A., Taddeo, M., Floridi, L.: The AI gambit: leveraging artificial intelligence to combat climate change—opportunities, challenges, and recommendations. AI Soc. **38**, 1–25 (2021)
38. Damasio, A., Carvalho, G.B.: The nature of feelings: evolutionary and neurobiological origins. Nat. Rev. Neurosci. **14**(2), 143–152 (2013). https://doi.org/10.1038/nrn3403
39. Dayan, P., Hinton, G.E., Neal, R.N., Zemel, R.: The Helmholtz machine. Neural Comput. **7**, 889–904 (1995)
40. Dennett, D.C.: Facing up to the hard question of consciousness. Philos. Trans. R. Soc. B Biol. Sci. **373**(1755), 20170342 (2018). https://doi.org/10.1098/rstb.2017.0342
41. Dehaene, S., Changeux, J.P., Naccache, L., Sackur, J., Sergent, C.: Conscious, preconscious, and subliminal processing: a testable taxonomy. Trends Cognit. Sci. **10**(5), 204–211 (2006). https://doi.org/10.1016/j.tics.2006.03.007
42. Dehaene, S., Changeux, J.P.: Experimental and theoretical approaches to conscious processing. Neuron **70**(2), 200–227 (2011). https://doi.org/10.1016/j.neuron.2011.03.018
43. Deperrois, N., Petrovici, M.A., Senn, W., Jordan, J.: Learning cortical representations through perturbed and adversarial dreaming. elife **11**, 1–34 (2022). https://doi.org/10.7554/elife.76384
44. Deperrois, N., Petrovici, M.A., Jordan, J., Huber, L.S., Senn, W.: How adversarial REM dreams may facilitate creativity, and why we become aware of them. Clin. Transl. Neurosci. **8**(2), 21 (2024)
45. De Graaf, M.M.A., Hindriks, F.A., Hindriks, K.V.: Who wants to grant robots rights? Front. Robot. AI **8**(January), 1–13 (2022). https://doi.org/10.3389/frobt.2021.781985
46. Du, C., Ren, Y., Qu, Z., Gao, L., Zhai, Y., Han, S.T., Zhou, Y.: Synaptic transistors and neuromorphic systems based on carbon nano-materials. Nanoscale **13**(16), 7498–7522 (2021). https://doi.org/10.1039/d1nr00148e
47. 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
48. Edelman, G.M., Tononi, G.: A Universe of Consciousness: How Matter Becomes Imagination. Basic Books, New York (2000)

49. Fiske, A., Henningsen, P., Buyx, A.: Your robot therapist will see you now: ethical implications of embodied artificial intelligence in psychiatry, psychology, and psychotherapy. J. Med. Internet Res. **21**(5), 1–12 (2019). https://doi.org/10.2196/13216
50. Fleming, S.M.: Awareness as inference in a higher-order state space. Neurosci. Conscious. **2020**, niz020 (2020)
51. French, R.M.: The turing test: the first 50 years. Trends Cognit. Sci. **4**(3), 115–122 (2000). https://doi.org/10.1016/S1364-6613(00)01453-4
52. Friston, K., Kiebel, S.: Predictive coding under the free-energy principle. Philos. Trans. R. Soc. B: Biol. Sci. **364**(1521), 1211–1221 (2009)
53. Friston, K.: Am I self-conscious? (or does self-organization entail self-consciousness?). Front. Psychol. **9**(APR), 1–10 (2018). https://doi.org/10.3389/fpsyg.2018.00579
54. Fuchs, T.: Ecology of the Brain: The Phenomenology and Biology of the Embodied Mind. Oxford University Press, Oxford (2018)
55. Fuchs, T.: Human and artificial intelligence: a clarification. In: Fuchs, T. (ed.) In Defence of the Human Being: Foundational Questions of an Embodied Anthropology, pp. 13–48. Oxford University Press, Oxford (2021)
56. Fuchs, T.: Understanding Sophia? On human interaction with artificial agents. Phenomenol. Cognit. Sci. **23**(1), 21–42 (2022). https://doi.org/10.1007/s11097-022-09848-0
57. Gent, T.C., LA Bassetti, C., Adamantidis, A.R.: Sleep-wake control and the thalamus. Curr. Opin. Neurobiol. **52**, 188–197 (2018). https://doi.org/10.1016/j.conb.2018.08.002
58. Gershman, S.J.: The generative adversarial brain. Front. Artif. Intell. **2**(September), 1–8 (2019). https://doi.org/10.3389/frai.2019.00018
59. Gidon, A., Aru, J., Larkum, M.E.: Does brain activity cause consciousness? A thought experiment. PLOS Biol. **20**(6), e3001651 (2022). https://doi.org/10.1371/journal.pbio.3001651
60. Goodfellow, I.J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial networks, pp. 1–9. arXiv at http://arxiv.org/abs/1406.2661 (2014)
61. 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.: Fast and energy-efficient neuromorphic deep learning with first-spike times. Nat. Mach. Intell. **3**(9), 823–835 (2021). https://doi.org/10.1038/s42256-021-00388-x
62. Gunkel, D.J.: The other question: Can and should robots have rights? Ethics Inf. Technol. **20**, 87–99 (2018). https://doi.org/10.1007/s10676-017-9442-4
63. 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.: Intercalated amygdala clusters orchestrate a switch in fear state. Nature **594**(7863), 403–407 (2021). https://doi.org/10.1038/s41586-021-03593-1
64. Helmholtz, H.V.: Concerning the perceptions in general, 1867. In: Dennis, W. (ed.) Readings in the History of Psychology, pp. 214–230. Appleton-Century-Crofts, East Norwalk (1948). https://doi.org/10.1037/11304-027
65. Hohwy, J., Seth, A.K.: Predictive processing as a systematic basis for identifying the neural correlates of consciousness. Philos. Mind Sci. **1**, 3 (2020)
66. Hubard, J., Harbaugh, W.T., Degras, D., Mayr, U.: A general benevolence dimension that links neural, psychological, economic, and life-span data on altruistic tendencies. J. Exp. Psychol. Gen. **145**(10), 1351–1358 (2016). https://doi.org/10.1037/xge0000209.supp
67. Humphrey, N.: A History of the Mind: Evolution and the Birth of Consciousness. Springer, New York (1999)
68. Indiveri, G., et al.: Neuromorphic silicon neuron circuits. Front. Neurosci. **5**, 73 (2011)
69. Jackson, F.: Epiphenomenal qualia. In: Toribio, J., Clark, A. (eds.) Consciousness and Emotion in Cognitive Science: Conceptual and Empirical Issues, pp. 197–206. Routledge (1998)
70. Jankélévitch, V.: Vorlesung über Moralphilosophie: Mitschriften aus den Jahren 1962–1963 an der Freien Universität zu Brüssel. Österreich: Turia + Kant, Wien (2007)
71. Jamieson, D.: Morality's Progress: Essays on Humans, Other Animals, and the Rest of Nature. Oxford University Press, Oxford (2002)
72. Jones, A.K.P., Friston, K.J., Frackowiack, R.S.J.: Cerebral localisation of responses to pain in man using positron emission tomography. Science **255**, 215–216 (1992)
73. Kang, J.I., Huppé-Gourgues, F., Vaucher, E., Kang, J.I.: Boosting visual cortex function and plasticity with acetylcholine to enhance visual perception. Front. Syst. Neurosci. **8**(September), 1–14 (2014). https://doi.org/10.3389/fnsys.2014.00172
74. Kang, Y.N., Chou, N., Jang, J.W., Choe, H.K., Kim, S.: A 3D flexible neural interface based on a microfluidic interconnection cable capable of chemical delivery. Microsyst. Nanoeng. **7**(1), 66 (2021). https://doi.org/10.1038/s41378-021-00295-6
75. Keller, G.B., Mrsic-Flogel, T.D.: Predictive processing: a canonical cortical computation. Neuron **100**(2), 424–435 (2018)
76. Kirk, R.: In: Zalta E.N. (ed.) Zombies (Spring 2021 Edition). The Stanford Encyclopedia of Philosophy
77. Klein, C.: What pain asymbolia really shows. Mind **124**(494), 493–516 (2015)
78. Kneer, M.: Can a robot lie? Exploring the folk concept of lying as applied to artificial agents. Cognit. Sci. **45**(10), e13032 (2021). https://doi.org/10.1111/cogs.13032
79. Kolodny, O., Moyal, R., Edelman, S.: A possible evolutionary function of phenomenal conscious experience of pain. Neurosci. Conscious. **2021**(2), niab012 (2021)
80. Kulkarni, B., Bentley, D.E., Elliott, R., Youell, P., Watson, A., Derbyshire, S.W.G., Jones, A.K.P.: Attention to pain localization and unpleasantness discriminates the functions of the medial and lateral pain systems. Eur. J. Neurosci. **21**(11), 3133–3142 (2005)
81. Lamm, C., Majdandžić, J.: The role of shared neural activations, mirror neurons, and morality in empathy—a critical comment. Neurosci. Res. **90**, 15–24 (2015). https://doi.org/10.1016/j.neures.2014.10.008
82. Lamme, V.A.F.: Towards a true neural stance on consciousness. Trends Cognit. Sci. **10**(11), 494–501 (2006). https://doi.org/10.1016/j.tics.2006.09.001
83. Lamme, V.A.F.: How neuroscience will change our view on consciousness. Cognit. Neurosci. **1**(3), 204–220 (2010). https://doi.org/10.1080/17588921003731586
84. Larkum, M.E., Senn, W., Lüscher, H.R.: Top-down dendritic input increases the gain of layer 5 pyramidal neurons. Cereb. Cortex **14**(10), 1059–1070 (2004). https://doi.org/10.1093/cercor/bhh065
85. Lau, H., Rosenthal, D.: Empirical support for higher-order theories of conscious awareness. Trends Cognit. Sci. **15**, 365–373 (2011)
86. Lau, H.: Consciousness, metacognition, and perceptual reality monitoring. Preprint at arXiv https://doi.org/10.31234/osf.io/ckbyf (2020)
87. LeDoux, J.E.: Emotion, memory and the brain. Sci. Am. **270**(6), 50–57 (1994)
88. Ledoux, J.E., Brown, R.: A higher-order theory of emotional consciousness. Proc. Natl. Acad. Sci. U.S.A. **114**(10), E2016–E2025 (2017). https://doi.org/10.1073/pnas.1619316114
89. Mead, C.: Neuromorphic electronic systems. Proc. IEEE **78**(10), 1629–1636 (1990)

90. Mariello, M., Kim, K., Wu, K., Lacour, S.P., Leterrier, Y.: Recent advances in encapsulation of flexible bioelectronic implants: materials, technologies, and characterization methods. Adv. Mater. **34**(34), 2201129 (2022)
91. Mashour, G.A., Roelfsema, P., Changeux, J.P., Dehaene, S.: Conscious processing and the global neuronal workspace hypothesis. Neuron **105**, 776–798 (2020)
92. Mediano, P.A.M., Rosas, F.E., Bor, D., Seth, A.K., Barrett, A.B.: The strength of weak integrated information theory. Trends Cognit. Sci. **26**(8), 646–655 (2022). https://doi.org/10.1016/j.tics.2022.04.008
93. Metzinger, T.: Being No One: The Self-model Theory of Subjectivity. MIT Press, Cambridge (2004)
94. Metzinger, T.: Artificial suffering: an argument for a global moratorium on synthetic phenomenology. J. Artif. Intell. Conscious. **8**(01), 43–66 (2021)
95. Miller, L.F.: Granting automata human rights: challenge to a basis of full-rights privilege. Hum. Rights Rev. **16**(4), 369–391 (2015). https://doi.org/10.1007/s12142-015-0387-x
96. Morowitz, H.J.: Rediscovering the mind. Psychol. Today **14**(3), 12–18 (1980)
97. Müller, V.C.: Is it time for robot rights? Moral status in artificial entities. Ethics Inf. Technol. **23**(4), 579–587 (2021). https://doi.org/10.1007/s10676-021-09596-w
98. Nagel, T.: What is it like to be a bat? Philos. Rev. **83**(4), 435–450 (1974)
99. Nagasako, E.M., Oaklander, A.L., Dworkin, R.H.: Congenital insensitivity to pain: an update. Pain **101**(3), 213–219 (2003)
100. Newitz, A.: The curious case of the AI and the lawyer. New Scientist **255**(3396), 28 (2022)
101. Nickel, J.: In: Zalta, E.N. (ed.) Human Rights (Fall 2021 Edition). The Stanford Encyclopedia of Philosophy. https://plato.stanford.edu/archives/fall2021/entries/rights-human
102. Pal, D., Silverstein, B.H., Lee, H., Mashour, G.A.: Neural correlates of wakefulness, sleep, and general anesthesia: an experimental study in rat. Anesthesiology **125**(5), 929–942 (2016). https://doi.org/10.1097/ALN.0000000000001342
103. Papini, M.R., Fuchs, P.N., Torres, C.: Behavioral neuroscience of psychological pain. Neurosci. Biobehav. Rev. **48**, 53–69 (2015)
104. Parfit, D.: Reasons and Persons. Clarendon Press, Oxford (1984)
105. Pehle, C., Billaudelle, S., Cramer, B., Kaiser, J., Schreiber, K., Stradmann, Y., Weis, J., Leibfried, A., Müller, E., Schemmel, J.: The BrainScaleS-2 accelerated neuromorphic system with hybrid plasticity. Front. Neurosci. **16**, 795876 (2022)
106. Pennartz, C.M.: The Brain's Representational Power: On Consciousness and the Integration of Modalities. MIT Press (2015)
107. Pennartz, C.M.: Consciousness, representation, action: the importance of being goal-directed. Trends Cognit. Sci. **22**(2), 137–153 (2018)
108. Pennartz, C.M.A., Farisco, M., Evers, K.: Indicators and criteria of consciousness in animals and intelligent machines: an inside-out approach. Front. Syst. Neurosci. **13**, 25 (2019). https://doi.org/10.3389/fnsys.2019.00025
109. Pennartz, C.M.: What is neurorepresentationalism? From neural activity and predictive processing to multi-level representations and consciousness. Behav. Brain Res. **432**, 113969 (2022)
110. Pennartz, C.: The Consciousness Network: How the Brain Creates Our Reality. Taylor & Francis, London (2024)
111. Persaud, P., Varde, A.S., Wang, W.: Can robots get some human rights? A cross-disciplinary discussion. J. Robot. **2021**, 1–11 (2021). https://doi.org/10.1155/2021/5461703
112. Price, D.D.: Psychological and neural mechanisms of the affective dimension of pain. Science **288**(5472), 1769–1772 (2000)
113. Rainville, P., Duncan, G.H., Price, D.D., Carrier, B., Bushnell, M.C.: Pain affect encoded in human anterior cingulate but not somatosensory cortex. Science **277**(5328), 968–971 (1997). https://doi.org/10.1126/science.277.5328.968
114. Richardson, E., Alaluf, Y., Patashnik, O., Nitzan, Y., Azar, Y., Shapiro, S., Cohen-Or, D.: Encoding in style: a stylegan encoder for image-to-image translation. In: Proceedings of the IEEE/CVFConference on Computer Vision and Pattern Recognition, pp. 2287-2296 (2021).
115. Rodríguez, L.F., Ramos, F.: Development of computational models of emotions for autonomous agents: a review. Cognit. Comput. **6**(3), 351–375 (2014). https://doi.org/10.1007/s12559-013-9244-x
116. Rose, K: AI poses 'risk of extinction', industry leaders warn. The New York Times. https://www.nytimes.com/2023/05/30/technology/ai-threat-warning.html (2023)
117. Roy, K., Jaiswal, A., Panda, P.: Towards spike-based machine intelligence with neuro-morphic computing. Nature **575**(7784), 607–617 (2019)
118. Russell, S.: Provably beneficial. Artif. Intell. (2020). https://doi.org/10.1145/3490099.3519388
119. Schuman, C.D., Potok, T.E., Patton, R.M., Birdwell, J.D., Dean, M.E., Rose, G.S., Plank, J.S.: A survey of neuromorphic computing and neural networks in hardware. arXiv preprint at arXiv:1705.06963 (2017)
120. Seth, A.K., Bayne, T.: Theories of consciousness. Nat. Rev. Neurosci. **23**, 439–452 (2022). https://doi.org/10.1038/s41583-022-00587-4
121. Searle, J.R.: Minds, brains, and programs. Behav. Brain Sci. **3**, 417–457 (1980)
122. Senn, W., Dold, D., Kungl, A.F., Ellenberger, B., Bengio, Y., Sacramento, J., Jordan, J., Petrovici, M.A.: A neuronal least-action principle for real-time learning in cortical circuits. elife (2023). https://doi.org/10.1101/2023.03.25.534198
123. Smart, R.N.: Negative utilitarianism. Mind **67**(268), 542–543 (1958)
124. Simons, J.S., Garrison, J.R., Johnson, M.K.: Brain mechanisms of reality monitoring. Trends Cognit. Sci. **21**, 462–473 (2017)
125. Singer, M.G.: The paradox of extreme utilitarianism. Pac. Philos. Q. **64**, 242–248 (1983). https://doi.org/10.1111/j.1468-0114.1983.tb00197.x
126. Solms, M., Friston, K.J.: How and why consciousness arises: some considerations from physics and physiology. J. Conscious. Stud. **25**, 202–238 (2018)
127. Solms, M.: The hard problem of consciousness and the free energy principle. Front. Psychol. **9**(JAN), 1–16 (2019). https://doi.org/10.3389/fpsyg.2018.02714
128. Solms, M.: New project for a scientific psychology: general scheme. Neuropsychoanalysis **22**(1–2), 5–35 (2020)
129. Storm, J.F., Klink, P.C., Aru, J., Senn, W., Goebel, R., Pigorini, A., Avanzini, P., Vanduffel, W., Roelfsema, P.R., Massimini, M., Larkum, M., Pennartz, C.M.A.: An integrative, multiscale view on consciousness theories. Neuron **112**, 1532–1552 (2024)
130. Sunstein, C.R., Nussbaum, M.C. (eds.): Animal Rights: Current Debates and New Directions. Oxford University Press, Oxford (2004)
131. Takahashi, N., Ebner, C., Sigl-Glöckner, J., Moberg, S., Nierwetberg, S., Larkum, M.E.: Active dendritic currents gate descending cortical outputs in perception. Nat. Neurosci. **23**(10), 1277–1285 (2020). https://doi.org/10.1038/s41593-020-0677-8
132. Talbot, K., Madden, V.J., Jones, S.L., Moseley, G.L.: The sensory and affective components of pain: are they differentially modifiable dimensions or inseparable aspects of a unitary experience? A systematic review. Br. J. Anaesth. **123**(2), e263–e272 (2019). https://doi.org/10.1016/j.bja.2019.03.033
133. Tiku, N.: 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/ (2022)
134. Tononi, G., Edelman, G.M.: Consciousness and complexity. Science **282**(5395), 1846–1851 (1998). https://doi.org/10.1126/science.282.5395.1846
135. Tunyasuvunakool, K., Adler, J., Wu, Z., Green, T., Zielinski, M., Žídek, A., Hassabis, D.: Highly accurate protein structure prediction for the human proteome. Nature **596**(7873), 590–596 (2021)
136. United Nations Declaration on Human Cloning. https://en.wikipedia.org/wiki/United_Nations_Declaration_on_Human_Cloning (2005)
137. Urbanczik, R., Senn, W.: Learning by the dendritic prediction of somatic spiking. Neuron **81**(3), 521–528 (2014). https://doi.org/10.1016/j.neuron.2013.11.030
138. Van Gulick, R.: In: Zalta, E.N., Nodelman, U. (eds.) Consciousness (Winter 2022 Edition). The Stanford Encyclopedia of Philosophy. https://plato.stanford.edu/entries/consciousness/
139. Weaver, J.F.: My Client, the AI, Slate. https://slate.com/technology/2022/07/could-an-a-i-hire-a-lawyer.html (2022)
140. Williams, S.R., Fletcher, L.N.: A dendritic substrate for the cholinergic control of neocortical output neurons. Neuron **101**(3), 486-499.e4 (2019). https://doi.org/10.1016/j.neuron.2018.11.035
141. Whyte, C.J., Munn, B.R., Aru, J., Larkum, M., John, Y., Müller, E.J., Shine, J.M.: A biophysical model of visual rivalry links cellular mechanisms to signatures of conscious perception. BioRxiv (2023).
142. Yuk, H., Lu, B., Lin, S., Qu, K., Xu, J., Luo, J., Zhao, X.: 3D printing of conducting polymers. Nat. Commun. **11**(1), 1604 (2020)
143. Zahavi, D.: Self-awareness and Alterity: A Phenomenological Investigation. Northwestern University Press, Evanston (1999)
144. Zeng, T., Yang, Z., Liang, J., Lin, Y., Cheng, Y., Hu, X., Zhao, X., Wang, Z., Xu, H., Liu, Y.: Flexible and transparent memristive synapse based on polyvinylpyrrolidone/N-doped carbon quantum dot nanocomposites for neuromorphic computing. Nanoscale Adv. **3**(9), 2623–2631 (2021). https://doi.org/10.1039/d1na00152c