Artificial Intelligence Review (2024) 57:133
https://doi.org/10.1007/s10462-024-10769-4

A method for the ethical analysis of brain-inspired AI

Michele Farisco[1,2] · G. Baldassarre[3] · E. Cartoni[3] · A. Leach[4] · M.A. Petrovici[5] · A. Rosemann[4,6] · A. Salles[1,7] · B. Stahl[4,8] · S. J. van Albada[9,10]

Accepted: 18 April 2024 / Published online: 3 May 2024

Abstract

Despite its successes, to date Artificial Intelligence (AI) is still characterized by a number of shortcomings with regards to different application domains and goals. These limitations are arguably both conceptual (e.g., related to the underlying theoretical models, such as symbolic vs.connectionist), and operational (e.g., related to robustness and ability to generalize). Biologically inspired AI, and more specifically brain-inspired AI, promises to provide further biological aspects beyond those that are already traditionally included in AI, making it possible to assess and possibly overcome some of its present shortcomings. This article examines some conceptual, technical, and ethical issues raised by the development and use of brain-inspired AI. Against this background, the paper asks whether there is anything ethically unique about brain-inspired AI. The aim of the paper is to introduce a method that has a heuristic nature and that can be applied to identify and address the ethical issues arising from brain-inspired AI (and from AI more generally). The conclusion resulting from the application of this method is that, compared to traditional AI, brain-inspired AI raises new foundational ethical issues and some new practical ethical issues, and exacerbates some of the issues raised by traditional AI.

Keywords Brain-inspired AI · NeuroAI · AI ethics · Philosophy of AI · Neuromorphic computing

1 Introduction

Because intelligence is a biological phenomenon, Artificial Intelligence (AI) is biologically inspired in its very essence. Nevertheless, AI may mimic biology to varying degrees.In some cases, AI research might be directly and extensively inspired by biology in general and by the brain in particular. In other cases it might attempt to go beyond biological instantiations of intelligence to implement mechanisms and types of intelligence not present in biological systems. Still, the mutual relationship between neuroscience and AI research has been important for advancing both approaches (Hassabis et al. 2017; Macpherson et al.

Extended author information available on the last page of the article

2021; Summerfield 2023) and has been recently indicated as crucial for the next-generation AI (Zador et al. 2023).

At the outset of AI research in the early 1950s, "the only known systems carrying out complex computations were biological nervous systems" (Ullman 2019). Accordingly, AI researchers productively used knowledge about the brain as a source of inspiration when seeking to create intelligent systems. This is true especially for AI paradigms alternative to the symbolic approach ("Good Old-Fashioned AI" or GOFAI), which was prevalent at the beginning of AI research and that aimed to reproduce the logical aspects of intelligence at a high functional level while neglecting the underlying brain mechanisms. In particular, research on artificial neural networks (ANNs) took inspiration from the mechanisms of brain functioning, such as the possibility of processing information based on multiple simple units similar to neurons, and their all-or-none signalling (Kleene 1956; McCulloch and Pitts 1943). At the same time, brain-inspired AI continued to inject new ways of thinking about how the brain works, in particular about neural computation (Saxe et al. 2020), suggesting new mechanisms for neural network models, alongside new research strategies in neuroscience.

This pattern has repeated multiple times, for example leading to the development of Deep Neural Networks (DNNs), which involve architectures inspired by the hierarchical structure of perceptual cortices, and currently yield state-of-the-art performance in several Machine Learning (ML) areas. While most principles underlying DNNs are grounded in early neural network models, at present we are witnessing a significant growth in their potential and applications, mainly owing to the increased availability of data ("Big Data") and the possibility of using specialised hardware such as Graphics Processing Units (GPUs). Many researchers believe that DNNs offer new possibilities for a mutual collaboration between neuroscience and AI insofar as they learn through sensory signals that are processed in ways that have some resemblance to sensory processing in the brain. This is true especially, but not exclusively, for unsupervised learning and reinforcement learning (Hassabis et al. 2017).

The increased collaboration between neuroscience and AI could result in further improvement of AI, possibly allowing it to overcome some of its current limitations. As mentioned above, the reverse is also possible, and AI can lead to a deeper understanding of the brain (Richards et al., 2019). Still, compared to current AI models, brains have advantages such as a greater capacity for generalisation and for learning from fewer examples (George et al. 2020), and a much lower power consumption (Attwell and Laughlin 2001).

Biological inspiration is not limited to human brains and cognitive reasoning: it may also come from organisms, processes, and phenomena occurring at different spatial and temporal scales (Floreano and Mattiussi 2008). As Floreano and Mattiussi claim, evolutionary, cellular, neural, developmental, immune, behavioral, and collective systems are some of the relevant sources of information.

Notwithstanding their great potential, biologically inspired AI in general and brain-inspired AI in particular are not immune to criticism. As the history of science and technology shows, the usefulness of deriving inspiration from biology cannot be taken for granted (AA.VV., 2012; Crick 1989), and it is theoretically possible for AI to develop along lines not consistent with brain architecture and functioning. Indeed, AI algorithms and techniques are often developed with the aim of solving particular tasks and only later are they compared with the brain (George et al. 2020; Gershman 2023; Lillicrap et al. 2020).

Calling for AI to emulate the brain thus risks being reductive and limiting depending on the goals pursued, as AI may fruitfully follow directions different from biology. Taking the brain as the privileged source of inspiration for further development of AI may present problems such as replicating the brain's limitations and biases in the development of AI. Conversely, assuming that AI could inform a better understanding of the brain raises issues as well. Some authors note that the relevance of DNNs for neuroscience is limited due to the characteristics of current approaches which lack sufficiently constrained learning rules, regularization principles, or architectural features (Hasson et al. 2020; Richards et al., 2019; Saxe et al. 2020). Current DNN strategies are thus limited in inspiring empirically testable hypotheses for brain research.

This article explores the conceptual, technical, and ethical issues arising during the development and use of biologically inspired AI with a focus on brain-inspired AI. The analysis results from the unique collaborative work of researchers from different disciplines in the EU-funded *Human Brain Project.* The research question underlying the ethical analysis is: is there anything ethically unique about brain-inspired AI? The paper starts with a discussion about biologically inspired and brain-inspired AI. Next, it introduces a method for the analysis of the practical ethical issues arising from AI in general and brain-inspired AI in particular. This is followed by an illustration of the method through two case studies (natural language processing and continual learning/context understanding). The article closes with an analysis of the fundamental ethical issues arising from brain-inspired AI, with two main *foci*: concepts and goals.

We propose the combination of two ethical approaches, fundamental and practical, in order to identify emerging ethical issues, prioritize their assessment, anticipate their impact on society, and maximize the benefits derived from brain-inspired AI.

The conclusion is that, compared to traditional AI, brain-inspired AI can raise qualitatively new foundational and practical ethical issues, and it can exacerbate some of the ethical issues raised by traditional AI.

2 Biologically inspired and brain-inspired AI: definition and philosophical reflections

2.1 Definition of biologically and brain-inspired AI

In its broadest sense, biological inspiration refers to the compatibility of AI with current knowledge in biology, particularly in neurobiology. Such a general description, while useful for the sake of introducing the concept, is not sufficiently constrained to be technically operationalizable. More specifically, an AI system is biologically inspired when its architecture and functioning include biological constraints that make specific parts of the system biologically realistic. Importantly, a biologically inspired AI system does not necessarily fully emulate or replicate the reference biological system, since different levels of biological realism are possible. Even if in theory biological inspiration can come from many different biological systems, the main trend today is to define biological realism of AI with specific reference to known biological principles of the brain, in particular mammalian and human brains. Of course, there is no such thing as *the* brain as brains vary substantially between both species and individuals of the same species. Furthermore, different organizational levels and regions of the same brain have different properties. The differences between brains have various causes. Roughly speaking, brain differences between species arise from evolutionary factors. Brain differences between individuals of the same species can be explained by the interplay of genetic and epigenetic factors (Changeux et al. 1973, 2021), further developmental factors (Bonduriansky and Day 2018), and factors reflecting the interaction with the environment like learning, nutrition, and disease. Moreover, the brain is a complex organ involving multi-level organization, including molecular, cellular, microcircuit, macrocircuit, and behavioural levels (Amunts et al. 2019). It is likely that not all these levels are equally relevant to the development of brain-inspired AI.

This has three implications in particular: (1) any general operational principle of the brain that we identify is the result of a statistical analysis or selection; (2) selection means choice, and this raises the issue of the criteria used to make the choice; (3) it is necessary to clarify what level of the brain is used as reference for brain-inspired AI. In short, biological inspiration conceived as a set of target features from the brain needs to be contextualized and specified. To illustrate: biological inspiration can be achieved by emulating a particular biological mechanism such as the spike-timing-dependent plasticity observed in biological synapses (Bi and Poo 1998, 2001; Gerstner et al. 1996; Schemmel et al. 2006; Song, Miller and Abbott 2000), or by using a biologically constrained model, such as the Hodgkin-Huxley model of neuronal action potentials (Hodgkin and Huxley 1952). It is also likely that more precise and technically relevant definitions of biological inspiration depend on the specific AI technique considered, and that there is a continuum of possible technologies with different levels of biological inspiration. These different levels are characterized by specific limitations and may raise different types of issues, including ethical issues.

As mentioned above, the biological brain, and more specifically the human brain, is usually uncritically assumed as the standard reference, either explicitly or implicitly. This tendency is evident, for instance, in these words by Jeff Hawkins: “From the moment I became interested in studying the brain, I felt that we would have to understand how it works before we could create intelligent machines. This seemed obvious to me, as the brain is the only thing that we know of that is intelligent” (Hawkins 2021). Yet, what seems clear to Hawkins and many others is actually not uncontested, both because it is not obvious that biological inspiration should be limited to the (human) brain (Floreano and Mattiussi 2008) and because it is not obvious that biological inspiration should be taken as a necessary requirement for improving AI. That it is not obvious that AI should be inspired by the brain is related to the fact that some functional features of the brain might reflect physiological constraints rather than optimal computational implementations. For instance, spikes themselves might not be necessary *in silico*, because copper makes for a better signal transmission substrate than cell membranes. Also, post-synaptic potentials (PSPs) do not have to decay gradually, because rectangular PSPs might lead to better functional performance of a network, but are simply more difficult to achieve with the electrophysiology of biological synapses.

2.2 Differences between current AI and the human brain

When compared to the human brain, current AI reveals a number of differences and limitations with regards to different domains and goals (Zador et al. 2023). These limitations are arguably of two main kinds: technical and conceptual. The technical limitations depend on the current technological stage of AI and are likely to be reduced and possibly overcome through further progress of knowledge and emerging technology. The conceptual limitations depend on the AI paradigms used, so overcoming them may require revised or new paradigms. To illustrate, present AI is still narrow, that is, it works for specific tasks in particular domains for which it is programmed and trained, and fails if environmental conditions are different from those in the training context (Marcus and Davis 2019). In this respect, its impressive success in specific applications is not yet translated into the capacity for solving broader and more general tasks. For some AI systems, this limitation might not be intrinsic but rather due to the time needed to adapt to new conditions and to learn how to behave in new environments, analogously to what happens with humans. Still, a consequence of such constraints is that even if omnipresent in our day-to-day lives, the applicability of AI in the real world and its related impact are still limited. Likewise, AI reliability has further development potential, for both technical and conceptual reasons. Indeed, some researchers argue for new approaches and paradigm changes in current AI (Marcus and Davis 2019), outlining that one of the main challenges for AI is building learning machines that are as *flexible and robust* as the biological brain and that have the same capacity for *generalization* (Sinz et al. 2019). Moreover, in contrast to AI, which at present can interact with a limited subset of variables within a contextualized environment, biological systems respond to a wide variety of stimuli over long periods of time, and their responses alter the environment and subsequent responses, giving rise to a kind of *systemic action-reaction cycle* (AA.VV., 2012). Other relevant abilities of the human brain that current AI systems are not able to fully replicate are "*multi-tasking*, *learning with minimal supervision*, […] all accomplished with high efficiency and *low energy cost*", as well as the ability to *communicate* via natural language (Poo 2018).

One possible strategy to improve the performance of current AI along the dimensions considered above is "to introduce structural and operational principles of the brain into the design of computing algorithms and devices" (Poo 2018). Relevant results have been obtained through neuromorphic systems, some of which show important advantages including increased computational power per unit of energy consumed (Cramer et al. 2022; Esser et al. 2016; Göltz et al. 2021a, b; Park, Lee and Jeon 2019) and robust learning (Buhler et al. 2017; Frenkel and Indiveri 2022; Renner et al. 2021; Wunderlich et al. 2019).

Another illustration comes from recent research that considers the biological feature that we are not born as clean slates, but already have a brain structure optimized by evolution for learning from our experiences. For instance, innate aspects of neural network structure may already gear networks toward effective task performance (Stöckl et al. 2022), and allow the generalization of learning within families of tasks (Bellec et al. 2020). Furthermore, cortical network models have been used to infer different possible learning rules (Zappacosta et al. 2018), including in spiking networks (Jordan et al. 2021). The spiking feature would add biological plausibility to most current AI systems, as would the inclusion of layer-specific connectivity as found in the cerebral cortex, including in particular layer-specific interactions between feedforward and feedback signals across cortical areas (Markov et al. 2014; Rao and Ballard 1999).

Yet the implementation of the strategies above is not simple, and the choice itself of taking biology, and specifically the brain, as a reference standard to calibrate and further develop AI systems needs a sound justification that goes beyond the acknowledgment that brains are the only thinking objects we are aware of. In fact, this does not *a priori* exclude the possibility of an AI regulated by different principles. Also, bio-inspired AI itself may outperform biology because it can profit from “better” (for example, faster) hardware (Billaudelle et al. 2021; Göltz et al. 2021a, b; Kungl et al. 2019b).

2.3 The risk of putting too much emphasis on emulating the brain

Focusing on general AI, some researchers have recently argued that biological inspiration and neuroscientific constraints should be regarded as strict requirements for AI until we understand the nature of intelligence (Hole and Ahmad 2021). In other words, since we do not yet understand the nature of intelligence in itself, we should take inspiration from the main example of intelligence we have in nature, that is, from the brain. This argument seems to put too much emphasis on the need to define or explain natural intelligence mechanisms before producing AI. This point was already criticized by Turing, who famously elaborated his imitation game to show that priority should be given to the operationalization of intelligence rather than to its theoretical definition (Dietrich et al. 2021; Turing 1950).

Philosophically, postulating that biological inspiration and resemblance to the human brain in particular should be paradigmatic for AI suggests an implicit endorsement of a form of anthropocentrism and anthropomorphism, which are both evidence of our intellectual self-centeredness and of our limitation in thinking beyond what we are (or what we think we are).

The debate around the biological plausibility of backpropagation is illustrative of the possibility of achieving results similar to biological intelligence by using principles that are not fully biologically compatible. Despite the fact that exact backpropagation is unlikely to happen in the brain (Crick 1989)(but see Haider et al. 2021; Lillicrap et al. 2016; Lillicrap et al. 2020; Millidge et al. 2022; Payeur et al. 2021; Pozzi et al. 2020; Sacramento et al. 2018; Song, Xu and Lafferty 2021 for a more informed discussion on this point), AI systems using backpropagation have achieved results comparable to and even better than humans in specific tasks (Pozzi et al. 2020).

From an operational point of view, the question of how to measure biological inspiration also arises. It seems that the quantification of biological inspiration depends on the domain and the brain level taken into consideration, so an AI system can be more or less biologically inspired depending on the particular level of its architecture to which we refer, and on the particular aspect of brain architecture and dynamics/physiology that inspired the AI system. To illustrate, it is possible to assess the biological realism of an ANN by checking if and how it abstracts from the behaviour of biological neurons. While this strategy might be sufficient to conclude that ANNs are biologically realistic from an operational point of view, it does not exclude that ANNs are not biologically realistic at another level, for instance at the computational level, since it is possible to implement different computations with the same underlying neuronal behaviour. Similarly, it is possible to produce the same network activity with different circuitry (Prinz et al. 2004), the latter point being very relevant for neuromorphic systems (Petrovici et al. 2014).

3 A method for the ethics of brain-inspired AI

For the sake of our analysis, ethics can be understood as the attempt to systematize and justify concepts of right and wrong, and to clarify the implications of these concepts for behaviour. We pursue this general goal through reflection on both fundamental/foundational and practical issues: the discrimination between right and wrong is justified by and relies on the identification of relevant values and principles. as well as on the definition of key notions like moral subject, moral reasoning, and moral action, among others.The definitional task is central in *fundamental* (or *foundational*) *ethics* that deals with the foundation, nature, and evolution of moral thought and judgment, and which can be distinguished from *practical ethics* that deals with concrete issues (e.g., ethical assessment of particular fields) (Evers 2007).

Accordingly, we can broadly distinguish two main kinds of ethical issues arising from AI, including brain-inspired AI: foundational (i.e., concerning both the justification of brain-inspired AI and its impact on how we think about fundamental moral notions) and practical (i.e., concerning the impact and implications of brain-inspired AI on our daily life). The first kind of issues involves theoretical analysis, while the second kind of issues involves mainly applied analysis, that is, the use of ethical theory to identify and address the practical issues related to the use of AI. Below, we begin by introducing a method for the analysis of the practical issues raised by AI and show how it can be applied by focusing on two case studies. We then present a foundational ethical analysis of key concepts and goals underlying brain-inspired AI.

3.1 Practical ethics of brain-inspired AI

Generally speaking, the practical issues raised by a technology can be intrinsic or extrinsic. Intrinsic issues are inseparable from how the technology itself works. Extrinsic issues are those emerging from a technology's deployment and use in different contexts, including those raised by the resulting relationship between humans and the relevant technology, such technology's impacts on society and the legitimacy of its intended goals.

We propose that the practical issues arising from biologically inspired AI, including brain-inspired AI, can be organized in terms of (at least) the following main levels:

- *Operational*, related to how AI works;
- *Instrumental*, related to how people use AI;
- *Relational*, related to how people see AI and to the resulting psychological and metaphysical human-AI relationship[1];
- *Societal*, related to the social and economic costs and consequences of the development and use of AI.

Importantly, the distinction between these levels is mainly for the sake of analysis: in practice the same factor may appear on different levels, for instance in terms of its "proximal" and "distal" effects. To illustrate, brain-inspired AI may have the proximal positive effect of needing less energy which may lead to the distal effect of cheaper systems and a conse-

[1] The way people perceive AI has an immediate impact on their own self-understanding and also on the way people conceive their nature in relation to the nature of AI.

quent more democratic access to it. More specifically, neuromorphic architectures can benefit from reduced power consumption compared to their more conventional counterparts, if they inherit relevant aspects of brain structure and dynamics. For example, the characteristic in-memory computing aspect of brain networks directly circumvents the notorious von-Neumann bottleneck (Indiveri and Liu 2015). Furthermore, spike latency codes may also provide benefits over conventional rate-based communication (Göltz et al. 2021a, b), but their scalability to large-scale applications still needs to be explored. And, specifically for analog neuromorphic substrates, the physical emulation of relevant dynamics as opposed to their simulation by an arithmetic logic unit can also yield benefits in terms of energy consumption and speed (Billaudelle et al. 2020).

Due to these advantages, large-scale brain-inspired AI substrates have the potential of being operable at a significantly reduced cost compared to conventional GPU clusters, thus democratizing the ownership and use of competitive AI hardware. Their reduced power consumption also implies a reduced carbon footprint, with evident benefits for the planet's climate. Furthermore, by being more affordable, they could place the relevant computational resources directly into the hands of the users, thus obviating the need for transmitting sensitive information to AI service providers in the cloud (Haider et al. 2021).

Also, the identified levels should not be seen as confined to biologically plausible technologies. The discussion of the ethical aspects of AI can be traced back to the beginning of digital and potentially autonomous systems (Wiener 1954). It has developed alongside technical progress (Dreyfus 1972; Whitby 1991) and gained prominence in recent years when the spectacular successes of machine learning became clear. There are various ways of approaching the ethics of AI (Coeckelbergh 2020; Dignum 2019; Stahl 2021) which include the distinction between foundational and practical and the four levels proposed here.

Operational issues of AI are those that have their basis in the very nature of AI. The prominent focus of established discussions of AI ethics are current machine learning technologies. These need large datasets for training and validating models. Where such datasets contain personal information, an operational issue would be the intrinsic threat of data protection violations (EDPS 2020; Kaplan and Haenlein 2019). Another operational issue is raised by the nature of neural networks used in much machine learning, because they are complex, opaque, and often not open to scrutiny, which leads to the recurrent issue of the need for explainable AI (Friedrich et al. 2022; Yeung 2018).

Instrumental concerns may be grounded on these operational questions and touch on how AI is used. One of the most prominent (and widely discussed) examples of this is the problem of biases and discrimination. Bias in machine learning can result from the replication of bias in social relationships that is reflected in the data and hence in machine learning models. Thoughtless use of these models may lead to the replication of bias in diverse contexts, for example in job selection and law enforcement (Birhane 2021; Team, 2018).

Societal issues typically arise less from the nature of AI and more from the way AI innovations influence and shape existing socio-economic structures. Relevant concerns refer to economic injustice and unfair distribution of risks and benefits (Walton and Nayak 2021; Zuboff 2019). Different issues arise when AI is inappropriately used to influence democratic processes or contribute to power concentration (Nemitz 2018; Parliament 2020). Further examples include AI's environmental impact (Nishant et al. 2020) and its potential to change the future of warfare (Brundage et al. 2018; Richards, Brockmann and Boulanini 2020).

Relational issues, which concern the ways in which AI changes the way we, as humans, see ourselves and whether it might alter our relationship with other humans, with technology, and with the world at large also have a prominent history. Addressing these issues has often led to speculative discussions about the potential of future AI, which are linked to concepts such as transhumanism, singularity, or superintelligence (Bostrom 2014). On a more immediate level, insofar as AI can serve as a metaphor for thinking about humans, changes to the capabilities of AI can change what characteristics we ascribe to ourselves and to each other.

We now focus on how these issues are raised in the context of brain-inspired AI.

Brain-inspired AI arguably promises new benefits and raises new risks for society (Doya et al. 2022). While there are some practical ethical issues raised by traditional AI that might be exacerbated by brain-inspired AI, others seem more specific to it. For example, intrinsic limitations and shortcomings of the brain might lead to specific practical ethical issues if somehow replicated by brain-inspired AI, as described in Table 1.

At the operational level, brain-inspired AI presents a number of potential benefits and risks. Among the possible benefits, brain-inspired AI has the potential to lead to *more environmentally friendly* systems because it may require less power than more traditional AI. This potential reduction derives from optimized functional architectures, including simplifications of the computing procedures and a minimization of memory access (Liu, Yu, Chai 2021). Furthermore, in many cases neuromorphic hardware shows increased computational power per unit of energy consumed (Cramer et al. 2022; Esser et al. 2016; Göltz et al. 2021a, b; Park, Lee and Jeon 2019). From a more general perspective, the brain notably uses much less energy than current computing systems (Attwell and Laughlin 2001), so that it is reasonable to take inspiration from it in order to optimize the energy consumption of AI. The reason for the high energy efficiency of the brain is still to be clarified, but it is likely that emulating some of its features (e.g., connectivity, dynamics, or algorithms) might help us to improve the energy efficiency of present hardware. Another possible strategy for saving energy through brain-inspired AI derives from analog neuromorphics specifically: by emulating relevant brain dynamics rather than simulating them (e.g., discharging a capacitor versus digitally calculating an exponential decay), one can gain energy efficiency and speed (Göltz et al. 2021b; Billaudelle et al. 2021).

Another possible benefit at the operational level is *more rapid optimization of systems* that are more strongly constrained than traditional AI (i.e., with a more restricted space of solutions for optimization).

On the other hand, brain-inspired AI has a limited capacity to emulate the brain because it necessarily refers to a specific level of brain organization (e.g., cellular level) or to specific phenomena and mechanisms, ignoring or streamlining its relation to other lower or higher levels (e.g., molecular or network level, respectively) or to other relevant phenomena and mechanisms. Moreover, brain activity is not independent from bodily influences (e.g., bodily activity seems to influence conscious perception (Park and Tallon-Baudry 2014) and confidence in subjective perception (Allen et al. 2016), and the molecular environment and metabolism regulate these brain-body couplings (Haydon and Carmignoto 2006; Jha and Morrison 2018; Petit and Magistretti 2016). This operational limitation raises the risk of *insufficient recognition of and flawed communication about the possibility of failure in fully emulating the brain*. That brain-inspired AI possibly fails to fully emulate the brain has ethical implications because there is a tendency, especially in lay people, to conceive the

Table 1 Practical ethical issues potentially arising from brain-inspired AI

Potential benefits	Potential risks
Operational Level	
- *More environmentally friendly systems*, because brain-inspired AI may require less power than more traditional AI - *More rapid optimization of systems* that are more strongly constrained than traditional AI (i.e., with a more restricted space of solutions for optimization)	- *Limited capacity of brain-inspired AI to emulate the brain*, leading to the risk of the insufficient recognition of this limitation and flawed communication about it - *Limited possibility for developers to optimize brain-inspired AI*, because it may be more strongly constrained than traditional AI and there is the risk of *a priori* excluding factors that are crucial for achieving the desired goals
Instrumental level	
Potential benefits	**Potential risks**
Applicability to certain domains for which traditional AI is less suited (e.g., because of energy- and learning-constraints)	*Risk of new kinds of brain-based crimes*, because brain-inspired AI may significantly improve the technology-mediated understanding of brain features and eventually lead to their exploitation (e.g., for brain hacking)
Relational level	
Potential benefits	**Potential risks**
Possibility to inspire more positive attitudes towards AI, because users' awareness of the 'brain-inspired' nature of the AI systems they use may trigger more favorable feelings towards AI's applications	- *Risk of hyped perception and misplaced trust in brain-inspired AI*, because of the risk of anthropomorphic attitudes and because of the background view of the brain as a paradigm of efficiency and effectiveness combined with inadequate recognition of and communication about the limited capacity of brain-inspired AI to emulate the brain - *Bewildering impact on ingrained beliefs about human identity*, particularly about human exceptionalism
Societal level	
Potential benefits	**Potential risks**
Possibility of new, cheaper, and more democratic products, including commercial applications, because brain-inspired AI may need less energy and may develop a broad domain of applications	*Risk of increased concentration of power in a few hands*, because of both the advanced technology required by brain-inspired AI and its applications domain extending that of more traditional AI

brain as paradigmatic (i.e., as a model of efficiency and effectiveness). It is possible that this view is projected onto brain-inspired AI if its above-mentioned limitation is not adequately recognized and communicated to the public. In other words, an ethical problem arising from the limited capacity of brain-inspired AI to emulate the brain is the lack of explicit and fair communication about this limitation. The tendency to assign human-like trustworthiness to Large Language Models (LLMs) like ChatGPT is an illustration of this risk.

The necessary selection of specific levels of brain organization as reference/target for brain-inspired AI can also *limit the possibility for humans to optimize the system if it fails to achieve its goals*: such failure might be caused by the fact that achieving the goal in question depends on factors (i.e., brain features) we are blind to because they were excluded by the preliminary selection (e.g., AI might be emulating one level without including its relation with other levels). In other words, if relevant aspects of brain organization or dynamics are excluded *a priori*, and they are crucial for optimizing a specific function, then there is limited possibility for developers to optimize the system (i.e., to improve its capacity to achieve the desired goal).

For example, it has been observed that it is useful for synapses to have access to multiple variables, beyond mere pre- and postsynaptic activity. One such type of variable is the activity of dendritic compartments, which are represented in morphologically complex neuron models, but not in point neuron models. Thus, choosing the simpler (point) model might impair the ability of a network to learn hierarchical representations (Haider et al. 2021; Sacramento et al. 2018). Furthermore, morphologically complex neurons can implement entire multilayer artificial neural networks (Beniaguev et al. 2021; Häusser and Mel 2003; Poirazi et al. 2003).

Another illustration of the point above is the inclusion of short-term plasticity or cortical oscillations into the dynamics of a modeled stochastic network. Without these, as the network learns to solve increasingly complex problems, it can become overly confident in one solution and is impaired in its ability to consider alternative solutions (Korcsak-Gorzo et al. 2022; Leng and Kakadiaris 2018).

A third example concerns the learning of time series, which relies on the presence of long enough transients in a network, often implemented by neuronal adaptation. If all specific time constraints are shorter than the transients in the signals to be learned (i.e., if all memory mechanisms are too short), learning becomes difficult or impossible (Bellec et al. 2020; Maass et al. 2002).

At the instrumental level, brain-inspired AI promises *applicability to certain domains for which traditional AI is less suited* because of the lack of necessary features, such as a sufficient capacity for online problem-solving and generalizability (i.e., the capacity to learn on the fly and to adapt to contextual variables that are different from the training data). While we are still limited in the ability to translate these capacities of the brain in AI systems, investing resources in trying to advance our knowledge of the brain's principles and mechanisms underlying these capacities, and in developing AI systems inspired by those principles and mechanisms appears a promising strategy.

On the other hand, the features of AI that could make its deployment more widespread and beneficial in some respects (like online problem solving) may have undesirable consequences if used for negative purposes, like fraudulent applications. In addition to the risks raised by traditional AI (e.g., security reduction, exploitation/violence, misuse/dual use, reduction or loss of human agency), *brain-inspired AI may raise additional risks of brain-based crimes*, because brain-inspired AI may significantly improve the technology-mediated understanding of brain features and eventually lead to their exploitation (e.g., for brain hacking, that is unauthorized access to and influence on subjective mental states/operations).

At the relational level, since it is modeled on the (human) brain, brain-inspired AI can inspire more empathetic feelings and this can lead to *more positive attitudes towards AI* (i.e., people are more willing to use AI) because of anthropomorphic tendencies. Considering that some cultures are still characterized by a form of technophobia (Weil and Rosen 1995), *awareness that AI is brain-inspired might diminish existing technophobic attitudes*.

On the other hand, the anthropomorphic feelings raised by brain-inspired AI may cause a *misplaced trust in its capabilities as well as hype about its application*, because of the background view of the brain as a paradigm of efficiency and effectiveness combined with inadequate recognition of and communication about the limited capacity of brain-inspired AI to emulate the brain. Furthermore, brain-inspired AI might be perceived as so close to human identity so as to have a *bewildering impact on ingrained beliefs about the relationships between humans and machines*. For instance, brain-inspired AI may be perceived as further evidence that human exceptionalism is illusory. This risk might again create either *disproportionate expectations* or *fears* about brain-inspired AI.

At the societal level, because of its potentially great energy efficiency and an impressive ability to handle vast amounts of data (Mehonic and Kenyon 2022), brain-inspired AI has the *potential for leading to new as well as to cheaper systems and commercial products*, which may eventually result in *more democratic and participative processes*. For instance, brain-inspired AI may lead to new clinical tools accessible to an increased number of stakeholders because of reduced costs compared to analogous traditional AI solutions. Yet the level of scientific knowledge and the kind of technology necessary to develop brain-inspired AI are so advanced compared to the technology available today that they raise the *risk of increased concentration of power in a few hands*, especially in the hands of tech companies and investors, analogously to what is happening with Large Language Models (LLM). The development of brain-inspired AI requires significant resources of different kinds, including economic and human resources. Adequate financial investments, both private and public, are necessary, and not all countries can afford such expenses, and some prefer to privilege other lines of research (Mehonic and Kenyon 2022). Furthermore, brain-inspired AI is cross-, multi-, and interdisciplinary, and it requires the collaboration of researchers from different fields. Not all public research institutions have the possibility to set up the necessary collaborative groups, and it is possible that only economically strong private companies can afford the related costs. Similarly, the potentially new application domains of brain-inspired AI or its improved performance in existing domains may amplify the uneven distribution of power between the rich and the poor that comes with powerful technology in general. For example, brain-inspired AI applications, especially when resulting from the work of private companies, may be subject to economic access gatekeeping, emphasizing issues of equity and justice concerning the possibility of taking advantage of the most recent technology. This unequal access may affect both private users at the domestic level and entire countries at the international level, exacerbating differences in a number of opportunities, including education, employment, scientific research, academic careers, and further technological advances. These risks are not exclusive to brain-inspired AI (as illustrated by the difference between free and paid versions of ChatGPT, which already raises the issue of equal access), but it is likely that brain-inspired AI will further emphasize them.

This analysis of the practical ethical issues arising from brain-inspired AI can now be illustrated by considering two possible fields of application that are benefitting from brain-inspired advancements: natural language processing and continual learning/context understanding.

3.1.1 Natural language processing

LLMs like GPT-3 and GPT-4 (Brown et al. 2020) have recently gained significant popularity due to impressive progress in their applicability, even by lay people in ordinary life contexts. For instance, discussions surrounding ChatGPT have exploded in the last few months. Since its release at the close of 2022, a growing number of people have harnessed its potential for a variety of purposes, including assistance in academic research (e.g., for summarizing or writing texts), in medicine (e.g., for assistance in diagnosis), in business (e.g., for content marketing), in education (e.g., for reviewing students essays), and in computer science (e.g., for coding). Notwithstanding these impressive results, ChatGPT is still brittle and fragile. This arises from the fact that the technology is still in its developmental stages and relies solely on one type of training data (i.e., text data). Compared to human intelligence, ChatGPT lacks embodiment, that is the sensorimotor abilities that humans use to explore the world through multisensory integration within a constitutive interaction between brain, body, and environment (Pennartz 2009). As a result of this interaction, human intelligence is multidimensional, and it is capable of online learning (i.e., of developing a realistic representation of the world that is adapted in real time). ChatGPT and LLMs in general presently lack this multidimensional and multisensory representation of the world, and therefore they are intrinsically exposed to limited and distorted knowledge.

Brain-inspired solutions might assist the improvement of LLMs on these aspects. To illustrate, potentially relevant results have been obtained incorporating biologically inspired neural dynamics into deep learning using a novel construct called spiking neural unit (SNU) (Woźniak et al. 2020). SNU has improved the energy efficiency of AI hardware, and it promises improvements in a number of tasks, including natural language processing. Also, a novel online learning algorithm for deep Spiking Neural Networks (SNNs), called online spatio-temporal learning (OSTL), has been introduced, with the potential for improving language modeling and speech recognition (Bohnstingl et al. 2022).

The application of the ethical framework described above to the case of natural language processing leads to the identification of the following ethically relevant potential benefits and risks (see Table 2).

At the *operational level*, an increased ability of AI systems to process natural language is likely to result in increased efficiency due to the use of heuristics based on the frequency of occurrence in natural language. However, there is the potential risk of greater difficulty in testing systems because there is a larger set of possible commands to check the functionality, leading to biased or inappropriate output going unnoticed.

At the *instrumental level*, there is the possibility of an easier use of AI (e.g., a richer and more flexible vocabulary to operate it) and more possibilities to exploit it (e.g., greater ease of use in varied contexts). In fact, AI systems will likely interact with human users in a more intuitive and direct way because more instructions/training data mediated by natural language will be possible. At the same time, it is also likely to increase the risk of unethical data processing and handling, for instance through more invasive AI systems, that is, systems characterized by an increased ability to identify and process sensitive data from humans.

At the *relational level*, an increased ability of AI systems to process natural language may either result in leading to more positive attitudes towards them (if they are perceived as closer to humans) or to hyped perception and misplaced trust in AI capacities (if anthropomorphic biases and unbalanced communication about the actual capacity of AI systems

Table 2 Potential ethically relevant benefits and risks arising from improved natural language processing by brain-inspired AI

Benefits	Risks
Operational level	
Increased efficiency due to the use of heuristics based on frequency of occurrence in natural language	Greater difficulty of testing systems because there is a larger set of possible commands, leading to biased or inappropriate output going unnoticed
Instrumental level	
Easier use of AI (e.g., a richer and more flexible vocabulary to operate it) and more possibilities to exploit it	Increased risk of unethical data processing (e.g., privacy infringement)
Relational level	
More positive attitudes towards AI systems	Risk of hyped perception and misplaced trust in AI systems and increased risk of "uncanny valley" effect.
Societal level	
More accessible AI systems (both easy to use and less expensive)	New forms of economic exploitations

prevail). Also, an increased risk of the so-called "uncanny valley" arises: a sense of uneasiness or eeriness experienced by humans when a technology is very similar to a human being but something appears "off" (Ciechanowski et al. 2019).

Finally, at the *societal level*, optimized natural language processing is likely to result in more accessible and "democratic" AI systems (i.e., more people will be able to use them because the interaction will be much easier and possibly less expensive for an increased commercialization), but at the same time it might trigger new forms of economic exploitation (e.g., through unequal capacity for user profiling) and a powerful impact on the job market (e.g., raising the risk of increasingly replacing humans in more "creative" activities like journalism or coding).

3.1.2 Continual learning and context understanding

Continual learning (also named open-ended learning) and context understanding present significant challenges to current AI, which is limited in its ability to learn new things on the fly and to adapt to new circumstances. Recent research has highlighted the possible contribution of neuroscience to improving the learning capacity of AI systems. For instance, the flexibility, adaptiveness to new circumstances, and fast learning capacity characteristic of the brain might derive in part from its capacity to autonomously learn on the basis of "intrinsic motivations" (e.g., curiosity, interest in novel stimuli or surprising events, and interest in learning new behaviours) (Barto 2004; Santucci et al. 2013). These are different from "extrinsic motivations" involving biological drives, such as hunger and pain, directed toward obtaining specific resources from the environment. Intrinsic motivations are maximally evident in children at play, and are, for example, related to novelty, surprise, and the success in accomplishing desired goals. The biological function of intrinsic motivations is to drive the learning of knowledge and skills that might be later used to find useful resources. The digital simulation of intrinsic motivations allows AI systems, such as DNNs and humanoid robots, to actively seek the knowledge they lack. This lets them adapt to new circumstances without the need for external guidance, eventually making AI more autonomous. Moreover, it speeds up learning processes as these are directed to the interactions that maximize knowledge and competence gain.

Directly related to the improvement of learning capability is the AI capacity for context understanding, that is, its ability to know relevant features of real-world contexts in order to appropriately act in them and to interact with other systems/agents within them. Time-continuous learning in substrates with time-continuous dynamics is the subject of a number of recent studies (Haider et al. 2021; Kungl et al. 2019a; Sacramento et al. 2018; Senn et al. 2023). This obviates the need for phases in learning, allowing AI agents to simply observe their surroundings in a time-continuous fashion, making it easier to embed them in complex, real-time contexts, which is one of the aspects current AI is missing.

Research has shown the advantages of attention for limiting the information stream to be processed, which may aid continual learning and context understanding. This is not only evidenced by the aforementioned LLMs relying on so-called transformers, which differentially weigh different parts of their input, but also by direct investigation of human learning (Niv et al. 2015). In this vein, a bio-inspired model that integrates bottom-up and top-down attention to scan the scene has been introduced (Ognibene and Baldassare 2015), showing the advantages of developing an integrated interplay between the two by continual reinforcement learning. This can lead an agent to autonomously find relevant stimuli in order to solve tasks, thus processing a smaller amount of information and speeding up learning. Another example is the recently proposed Attention-Gated Brain Propagation (BrainProp), a form of reinforcement learning that obviates the need for supervision to approximate error backpropagation (Pozzi et al. 2020). Attentional mechanisms are likely to become even more relevant as AI systems are developed that handle multimodal input streams, including not only text but also auditory and visual input.

The application of the ethical model presented above results in the following ethically relevant potential issues (see Table 3).

At the *operational level* the capacity for continual learning and the improvement of context understanding will likely result in more autonomous, flexible, and adaptable AI systems, which will be less prone to operative failures and more user-friendly. This will likely

Table 3 Potential ethically relevant benefits and risks arising from improved continual learning and context understanding by brain-inspired AI

Benefits	Risks
Operational level	
More autonomous, flexible, adaptable AI systems	Less transparent AI systems
Instrumental level	
Increased number of contexts in which to use AI systems	Less control on AI systems by the users
Relational level	
People more prone to see AI systems as reliable tools	Increased feeling of being insecure and/or less able than AI systems
Socio-economic level	
Possibility to maximize the advantage of using AI in different contexts	Risk of replacing human agents also in more creative activities

be counterbalanced by an increased risk of less transparent AI systems: since they will be more able to learn on the fly how to act, they will have evolving parameters and will eventually be more independent from top-down instructions and external monitoring, reducing the space for human understanding and supervision. This is an ethical issue because increased opacity of AI systems leads to an increased risk of unexpected operational failure with potential negative consequences, and fewer possibilities to prevent them.

At the *instrumental level*, the potential uses of AI (i.e., the contexts in which it can be used) will eventually be increased, but there is a risk that the capacity for controlling it by the users will be reduced.

At the *relational level*, an enhanced ability of AI for continual learning and for interacting with its surroundings might promote on the one hand a general perception of AI as more reliable and robust because it would be more able to adapt to changing external conditions. On the other hand, some people might feel more insecure, because of an increasing perception of AI systems as autonomous agents, and a self-perception as less capable than AI in an increasing number of activities.

At the *societal level*, optimized continual learning and context understanding will make it possible to maximize the advantage of using AI in many more contexts than traditional AI, but the risk of replacing human agents in different sectors, including those requiring more creativity and capacity for adaptation, will likely increase, eventually leading to concerns about a reduction or loss of human agency.

3.2 Fundamental ethics of brain-inspired AI

We have introduced four levels for the identification and the analysis of the practical ethical issues arising from brain-inspired AI, and have illustrated them with two case studies. Here we identify some fundamental/foundational ethical issues raised by brain-inspired AI. As mentioned above, these issues concern the justification of the attempt itself to build brain-inspired AI and its impact on how we think about fundamental moral notions. Therefore, we distinguish two main categories of fundamental ethical issues:

- **Those related to goals**, which refer to questions like: What is the driver of brain-inspired AI? What do we want to achieve by it?
- **Those related to concepts**, which include issues like epistemic risk, implicit assumptions about brain-inspired AI, and considerations about the historical, cultural, and societal contexts of brain-inspired AI.

3.2.1 Fundamental ethical issues emerging from brain-inspired AI in relation to goals

Different goals motivate the attempt to translate brain principles and features into AI, and each raises fundamental ethical issues (see Table 4 and 5).

As noted above, a first, main intention of the development of brain-inspired AI is *improving traditional AI*, by advancing in the following sectors (among others):

- More effective interaction of AI systems with the world.
- Optimized information processing.

Table 4 Illustrative goals of brain-inspired AI and related fundamental ethical issues

Goal of brain-inspired AI	Fundamental ethical issue
Improving traditional AI	What does improvement mean and for what? E.g., does it include awareness of the potential impact on different social groups or actors, including non-human? Does it include awareness of cultural diversity?
Making AI more autonomous	How might more autonomous AI systems impact human autonomy, including how people think about their autonomy? How is more autonomy better in this context? What does it mean?
Discovering operational principles in different sensorimotor and cognitive fields, which may be engineered and applied to AI systems	Does brain-inspired AI aim to replicate operational brain principles related to ethical reasoning? Can this introduce artificial moral agents?
Taking the brain as a model for advancing in the direction of Artificial General Intelligence (AGI)	How should AGI be conceptualized, particularly in relation to ethical dimensions? and why AGI?

Table 5 Fundamental ethical issues arising from brain-inspired AI

Fundamental issues	Related ethical questions
The brain as a model for AI	How can we be sure that applying brain principles in AI is the best possible option for advancing the development of AI and thus maximizing the benefit for our society?
	Are we maybe just seconding our anthropocentric bias which makes it hard if not impossible for us to think beyond the form of intelligence as we know it in nature?
	How unlikely is it that we will eventually end up replicating the shortcomings that make biological, and particularly human intelligence a source of risks and dangers for human society and the rest of the world?
Potential implications of brain-inspired AI for those concepts that are traditionally assumed as the basis for qualifying an agent as moral	If what characterizes and qualifies our biological intelligence can be realistically replicated (i.e., simulated) artificially, would it imply that there is nothing morally unique in our identity? If so, why would this be ethically problematic?
	Would our self-understanding as moral agents be impacted?
	If what characterizes us as moral agents (e.g., relevant cognitive and/or emotional abilities) can be artificially replicated through a brain-inspired AI, would this imply that such AIs qualify as moral agents? If so, would this be a problem? Why?
	Would a hypothetical artificial replication of our moral agency have consequences in how to discriminate between the good and the bad? How would this impact how we understand moral reasoning and decision-making?

- Improvement of AI systems' capability of on-line problem solving.
- Improvement of high assurance systems, like in manufacturing, information technology,navigation, etc.
- More effective real-world applications, like in banking, defense, education, finance, medicine, security, etc.
- More efficient robotics applications and better embodiment of AI.
- More flexible and autonomous AI.

A fundamental issue arising from the general goal of improving traditional AI concerns the *underlying understanding of improvement*. Related to this question is the issue of which improvements will lead to a better society. These are conceptual issues with direct ethical implications. For instance, to what extent does the desired improvement include considerations about potential societal and ethical impacts of brain-inspired AI on different social groups or actors, including non-human entities? Is the notion of improvement in this case only informed by technical considerations? Does it take into account cultural differences? In fact, culture, conceived as both socio-political and disciplinary identity, significantly impacts how we discriminate between what is good and what is bad (Farisco 2023). The impact of culture on our evaluations is expressed in the concept of *cultural model* as elaborated in cognitive anthropology. Cultural models are "mental representations shared by members of a culture" (Bennardo, De Munck 2014), which inform our knowledge and our evaluation of the world, including our ethical reasoning. For instance, people from more individualistic cultures may be prone to privilege values like privacy and confidentiality in the use of AI systems, while people from more collective cultures may manifest opposite inclinations, for instance giving priority to collective security rather than individualistic rights. This illustrates that the meaning of improvement is not self-evident or objective: It is necessary to acknowledge the controversy around the meaning of improvement and promote a thoughtful exchange of different opinions that are conducive to human well-being across the cultural landscape.

A second goal of brain-inspired AI is to *make AI more autonomous* through a better clarification and artificial replication of motivations, emotions, and values underlying human decision-making. A fundamental ethical question raised by this goal revolves around *possible impacts of more autonomous AI on humans*, including on how humans perceive their autonomy. For instance, when interacting with autonomous AI systems, people may not feel unique anymore, and may ultimately perceive these systems as a limiting factor for their own autonomy.

A third goal of brain-inspired AI is more knowledge-oriented: *discovering operational principles in different sensorimotor and cognitive fields, which may be engineered and applied to AI systems*. This way a positive epistemic feedback loop between brain-inspired AI and neuroscience may be realized: the first is inspired by the second and at the same time contributes to further advancing our knowledge of the brain. A fundamental ethical question concerns the *kind of brain principles that we aim to scale and apply to AI systems*: do they also include principles related to the capacity for ethical reasoning? If so, the possibility that brain-inspired AI systems may be equipped with moral reasoning cannot be logically excluded, even if the technical feasibility remains challenging. Also, informing AI with brain principles underlying capacities like empathy and sociability may help in developing more friendly and beneficial systems.

Finally, another motivation for brain-inspired AI is *advancing in the direction of Artificial General Intelligence (AGI)*, even if both its conceptual reliability and technical feasibility remain controversial (Summerfield 2023). A fundamental point to consider is that general intelligence is likely shaped by several factors, including bodily and environmental factors. This implies that limiting the focus to the brain risks being overly reductive and eventually ineffective for reaching AGI. At the ethical level, the question of *what kind of generality we refer to in the underlying concept of AGI and why we are seeking it* arises: does it also include ethically relevant features and if so, how are these dealt with? And what is the underlying motivation of seeking AGI? For instance, if informed by only cognitive features without the emotional and social dimensions of human intelligence, AGI may be eventually rather limited and even potentially dangerous: history shows that intelligence may in principle lead to great achievements as well as to great risks.

The aforementioned goals are not morally neutral in themselves. If achieved, they will necessarily lead to ethically relevant discussions regarding, for example, the notion of autonomy in general and of AI autonomy in particular, and about the necessity to align human and AI values. While this type of debate is not new within AI ethics, brain-inspired AI has the potential to bring it to the fore and exacerbate it insofar as it promotes advances in the direction of autonomous and ethically reasoning AI.

Importantly, all these goals, even if intrinsically bearing on ethical issues, are *a priori* neutral with regards to potential positive or negative societal consequences. In fact, the technical improvements potentially deriving from brain-inspired AI can equally lead to both good or bad applications. It is crucial to facilitate and to implement a multidisciplinary reflection and to set up a monitoring system in order to anticipate and identify relevant ethical issues in a timely manner. We propose combining the two ethical approaches introduced in this paper (i.e., fundamental and practical) as a promising strategy to identify emerging ethical issues, prioritize them, anticipate their impact on society, and eventually maximize the benefits deriving from brain-inspired AI.

3.2.2 Fundamental ethical issues emerging from brain-inspired AI in relation to concepts

This second group of fundamental ethical issues refers to theoretical considerations, which include epistemic risk, implicit assumptions about brain-inspired AI, and considerations about the historical, cultural, and societal contexts of brain-inspired AI. Epistemic risk means the possibility that the target model of brain-inspired AI (i.e., the brain) is not the best reference to optimize AI, either because we do not sufficiently know it or because of its own intrinsic limitations. This connects to implicit assumptions about brain-inspired AI, more specifically to the possibility of taking for granted that the brain is exemplary and that taking inspiration from it is the best possible strategy for optimizing AI. Finally, a number of questions arise from the interaction of brain-inspired AI with particular historical, cultural, and social contexts, and the resulting impact on ingrained foundational moral notions.

A first fundamental question concerns the justification of brain-inspired AI: why take the biological brain as a model for AI in the first place? This question is not only scientifically and technically relevant: it is also important from the point of view of public policy, Science and Technology Studies (STS), and ethics. From a public policy perspective a key question is whether the expected benefits of brain-inspired AI are substantial enough to prioritize its development rather than allocating resources to the further development of traditional AI. From an STS perspective, a point to address is what are the underlying politics and narratives used to frame brain-inspired AI as promising or cutting-edge. From an ethical perspective, the concern that emerges is whether biologically plausible and brain-inspired AI will actually lead to a maximization of benefits for society. After all, we cannot rule out the possibility that by developing brain-inspired AI we might be just following our already mentioned anthropocentric bias which makes it hard if not impossible for us to think beyond intelligence as we know it in nature. Is it possible that we tend to perceive brain-inspired AI as the optimal strategy to achieve societal benefits, because of our inability to see and appreciate alternatives to biological intelligence? We may be considering brain-inspired AI as more promising than traditional AI not for technical reasons but rather because of our cultural (i.e., anthropocentric) biases. Moreover, we must be mindful of the possibility that brain-inspired AI might end up replicating the shortcomings that make biological, and particularly human intelligence, a source of risks and harms for human societies at large. If so, this would have practical implications and an important theoretical dimension. These ethical questions touch upon our view of ourselves and our nature as intelligent agents, as well as our view of what is best for society and what role science and technology should play in it.

A second fundamental question concerns the possible implications of brain-inspired AI for those concepts that are traditionally assumed as the basis of morality, namely the conditions for agency in general and for moral agency in particular. Intentionality (i.e., wilful goal-oriented action) is usually assumed as a necessary and sufficient condition for agency (Davidson 1963, 1971). Yet this traditional view of agency has been complemented and sometimes criticized from different perspectives, highlighting, for instance, that the ability to reflect on and to care about the motivations of an action are peculiar to persons vs. non-persons (Frankfurt 1971), and that agency comprises different dimensions (i.e., authenticity, privacy, responsibility, trust) that are distinguished and reciprocally linked at the same time (Schönau et al. 2021). Attributing this plethora of concepts to AI is not uncontroversial, and brain-inspired AI might push us further in this direction, for instance if it replicates relevant functions of the human brain.

Regarding moral agency, brain-inspired AI might raise issues related to the traditional criteria for qualifying as a moral agent, and the sense and/or self-concept of being a moral agent. Moral agency is basically understood as the ability to discern right from wrong and to choose the relevant goals to pursue. While the issue of whether this capacity is mainly cognitive or emotional continues to be debated, there is wide consensus on the crucial role played by the brain and particularly by some specific cerebral areas and related functionalities (Verplaetse 2013).

A number of ethically relevant foundational questions may arise from the prospect of brain-inspired AI: if what characterizes and qualifies our biological intelligence can be realistically replicated (i.e., simulated) artificially, would this imply that there is nothing unique (both metaphysically and morally) to our natural identity? Is this morally problematic? If yes, for what reasons? The supposed uniqueness of humans has already been falsified by evolutionary biology, but brain-inspired AI poses a different threat, specifically to the supposedly unique identity of humans as moral agents. Specifically, if the features that characterize us as moral agents (e.g., relevant cognitive and/or emotional abilities) can be artificially replicated through brain-inspired AI, does this suggest that AIs can be moral agents? Would our self-understanding as moral agents be impacted? And would a hypothetical artificial replication of our moral agency have consequences for the way we discriminate between good and bad?

It is unlikely that implementing some brain-inspired principles in currently narrow AI will raise these kinds of issues, particularly because at present AI appears too constrained to the specific goals for which it has been pre-programmed, and its flexibility and robustness are too limited to enable the rise of artificial moral agency. These issues are more likely to be raised by the presence of a hypothetical human-like or human-level brain-inspired AI which at least at present appears far-fetched. Still, the theoretical possibility is, as such, ethically relevant, because it might be translated into technical feasibility, and ethics should not be taken to be just reactive but, importantly, proactive, anticipating possible future scenarios.

There is an additional fundamental issue potentially relevant to brain-inspired AI, one that arises from how brain-inspired AI is conceived and developed, and which will inevitably impact how it eventually operates. Since genetic, epigenetic, and environmental factors, including culture, socialization, and culturally/historically formed notions of identity shape the brain, the possibility of introducing biases in brain-inspired AI arises. This raises a number of ethically relevant issues: can brain-inspired AI inadvertently present a set of novel biases, that is, different from those raised by traditional AI? If so, how should these biases be evaluated from an ethical perspective? For instance, brain-inspired AI may enhance biases caused by the selection of training data (e.g., it may take only selected neuronal data as reference) if calibrated on particular sets of brain data (e.g., from adult brains rather than young brains, without accounting for important differences in terms of plasticity and dynamics).

The problem raised by biases in brain-inspired AI can also be turned around, by inquiring whether AI will be able to understand and apply the requisite manifestations of "diversity" and comprehend the "situatedness" of human cognition and perception in various types of AI-based predictions and applications. The point, which is ethically very salient, is that without awareness or understanding of the significance of the diversity of human knowledge, perception, and capacity of making sense of the world, many AI predictions and applications are themselves likely to be biased, of limited utility or eventually useless.

Moreover, implicit or explicit neuro-essentialist and neuro-reductionist cultural models (Vidal 2017) might impact how the general public perceives brain-inspired AI. In fact, if the brain is conceived as the core essence of human nature, and if human identity is eventually reduced to the brain, then imitating it means imitating what humans are, including ethically relevant features. As a consequence, brain-inspired AI might raise new fears, for instance about the supposed risk of creating artificial sentience leading to artificial forms of suffering (Metzinger 2021).

4 Conclusion

Brain-inspired AI is an attractive strategy for improving current AI, but it raises a number of technical and ethical issues. In this paper we summarized the main conceptual and technical aspects of brain-inspired AI, and introduced a method for its ethical analysis, distinguishing two main kinds of ethical issues: practical and fundamental/foundational.

We proposed that the practical issues arising from brain-inspired AI can be organized in terms of the following main levels:

- *Operational*, related to how AI works;
- *Instrumental*, related to how people use AI;
- *Relational*, related to how people see AI and to the resulting psychological and metaphysical human-AI relationship;
- *Societal*, related to the social and economic costs and consequences of the development and use of AI.

We described the fundamental/foundational ethical issues arising from brain-inspired AI as those issues that concern the justification of the attempt itself to build brain-inspired AI and its impact on how we think about fundamental moral notions. We identified two main categories of this second kind of issues: those related to goals and those related to concepts.

On the basis of the application of the method introduced in this paper we conclude that brain-inspired AI has the potential to raise new fundamental/foundational and practical ethical issues, as well as to exacerbate the practical ethical issues raised by traditional AI, which should be considered in relation to this promising approach.

Acknowledgements Special thanks to the participants of CRB scientific seminars, particularly to Kathinka Evers and Pär Segerdahl, and to two anonymous reviewers for extremely useful comments that improved the readability and clarity of the paper.

Author contributions M.F. wrote the manuscript and was responsible for general ideas. S.J.A. contributed to revising and developing ideas. All authors were involved in revising the manuscript critically and substantively for important intellectual content and gave final approval of the paper and agreed to be accountable for all aspects of the work. All authors read and approved the final manuscript.

Funding European Union's Horizon 2020 Framework Programme for Research and Innovation under the Specific Grant Agreement No. 945539 (Human Brain Project SGA3) (All); European Innovation Council Action No.101071178 (CAVAA) (M.F.); European Union Grant Agreements No. 604102, 720270, 785907 (M.A.P.); the Manfred Stärk Foundation (M.A.P.).
Open access funding provided by Uppsala University.

Declarations

Competing interests The authors declare no competing interests.

References

AA.VV. (2012) Is the brain a good model for machine intelligence? Nature, *482*(7386), 462–463. https://doi.org/10.1038/482462a

Allen M, Frank D, Schwarzkopf DS, Fardo F, Winston JS, Hauser TU, Rees G (2016) Unexpected arousal modulates the influence of sensory noise on confidence. eLife 5:e18103. https://doi.org/10.7554/eLife.18103

Amunts K, Knoll AC, Lippert T, Pennartz CMA, Ryvlin P, Destexhe A, Bjaalie JG (2019) The human brain project-synergy between neuroscience, computing, informatics, and brain-inspired technologies. PLoS Biol 17(7):e3000344. https://doi.org/10.1371/journal.pbio.3000344

Attwell D, Laughlin SB (2001) An energy budget for signaling in the grey matter of the brain. J Cereb Blood Flow Metab 21(10):1133–1145. https://doi.org/10.1097/00004647-200110000-00001

Barto AG (2004) Intrinsically Motivated Learning of Hierarchical Collections of Skills *Proceedings of the 3rd International Conference on Development and Learning* (pp. 112–119). Rome: Diamond Scientific Publishing

Bellec G, Scherr F, Subramoney A, Hajek E, Salaj D, Legenstein R, Maass W (2020) A solution to the learning dilemma for recurrent networks of spiking neurons. Nat Commun 11(1):3625. https://doi.org/10.1038/s41467-020-17236-y

Beniaguev D, Segev I, London M (2021) Single cortical neurons as deep artificial neural networks. Neuron 109(17):2727–2739e2723. https://doi.org/10.1016/j.neuron.2021.07.002

Bennardo G, De Munck VC (2014) Cultural models: genesis, methods, and experiences. Oxford University Press, New York

Bi, Poo MM (1998) Synaptic modifications in cultured hippocampal neurons: dependence on spike timing, synaptic strength, and postsynaptic cell type. J Neurosci 18(24):10464–10472. https://doi.org/10.1523/JNEUROSCI.18-24-10464.1998

Bi, Poo M (2001) Synaptic modification by correlated activity: Hebb's postulate revisited. Annu Rev Neurosci 24:139–166. https://doi.org/10.1146/annurev.neuro.24.1.139

Billaudelle S et al (2020) Versatile Emulation of Spiking Neural Networks on an Accelerated Neuromorphic Substrate. *2020 IEEE International Symposium on Circuits and Systems (ISCAS)*, 1–5. https://doi.org/10.1109/ISCAS45731.2020.9180741

Billaudelle S, Cramer B, Petrovici MA, Schreiber K, Kappel D, Schemmel J, Meier K (2021) Structural plasticity on an accelerated analog neuromorphic hardware system. Neural Netw 133:11–20. https://doi.org/10.1016/j.neunet.2020.09.024

Birhane A (2021) Algorithmic injustice: a relational ethics approach. Patterns 2(2):100205. https://doi.org/10.1016/j.patter.2021.100205

Bohnstingl T, Wozniak S, Pantazi A, Eleftheriou E (2022) Online spatio-temporal learning in deep neural networks. IEEE Trans Neural Netw Learn Syst. https://doi.org/10.1109/tnnls.2022.3153985

Bonduriansky R, Day T (2018) Extended heredity: a New understanding of inheritance and evolution. Princeton University Press, Princeton; Oxford

Bostrom N (2014) *Superintelligence: Paths, Dangers, Strategies* (First edition). Oxford University Press

Brown T, Mann B, Ryder N, Subbiah M, Kaplan JD, Dhariwal P, Amodei D (2020) Language Models are Few-Shot Learners. *Advances in Neural Information Processing Systems 33 (NeurIPS 2020)*

Brundage M, Avin S, Clark J, Toner H, Eckersley P, Garfinkel B, Amodei D (2018) *The Malicious Use of Artificial Intelligence: Forecasting, Prevention, and Mitigation*

Buhler FN, Brown P, Li J, Chen T, Zhang Z, Flynn MP (2017) 5–8 June 2017). *A 3.43TOPS/W 48.9pJ/pixel 50.1nJ/classification 512 analog neuron sparse coding neural network with on-chip learning and classification in 40nm CMOS* Paper presented at the 2017 Symposium on VLSI Circuits

Changeux JP, Courrège P, Danchin A (1973) A theory of the epigenesis of neuronal networks by selective stabilization of synapses. Proc Natl Acad Sci U S A 70(10):2974–2978

Changeux JP, Goulas A, Hilgetag CC (2021) A connectomic hypothesis for the hominization of the brain. Cereb Cortex 31(5):2425–2449. https://doi.org/10.1093/cercor/bhaa365

Ciechanowski L, Przegalinska A, Magnuski M, Gloor P (2019) In the shades of the uncanny valley: an experimental study of human–chatbot interaction. Future Generation Comput Syst 92:539–548. https://doi.org/10.1016/j.future.2018.01.055

Coeckelbergh M (2020) AI ethics. The MIT, Cambridge, MA

Cramer B, Billaudelle S, Kanya S, Leibfried A, Grubl A, Karasenko V, Zenke F (2022) Surrogate gradients for analog neuromorphic computing. Proc Natl Acad Sci U S A 119(4). https://doi.org/10.1073/pnas.2109194119

Crick F (1989) The recent excitement about neural networks. Nature 337(6203):129–132. https://doi.org/10.1038/337129a0

Davidson D (1963) Actions, reasons, and causes. J Philos 60(23):685

Davidson D (1971) Agency. In: Marras A, Binkley RW, Bronaugh RN (eds) Agent, Action, and reason. University of Toronto, Toronto, pp 1–37

Dietrich E, Fields C, Sullins JP, Van Heuveln B, Zebrowski R (2021) Great philosophical objections to Artificial Intelligence: the history and legacy of the AI wars. Bloomsbury Academic, London

Dignum V (2019) Responsible Artificial Intelligence: how to develop and use AI in a responsible way. Springer

Doya K, Ema A, Kitano H, Sakagami M, Russell S (2022) Social impact and governance of AI and neurotechnologies. Neural Netw 152:542–554. https://doi.org/10.1016/j.neunet.2022.05.012

Dreyfus HL (1972) What computers can't do; a critique of artificial reason, 1st edn. Harper & Row, New York

EDPS (2020) *Opinion on the European Commission's White Paper on Artificial Intelligence – A European approach to excellence and trust (Opinion 4/2020) (Opinion No. 4/2020)*

Esser SK, Merolla PA, Arthur JV, Cassidy AS, Appuswamy R, Andreopoulos A, Modha DS (2016) Convolutional networks for fast, energy-efficient neuromorphic computing. Proc Natl Acad Sci U S A 113(41):11441–11446. https://doi.org/10.1073/pnas.1604850113

Evers K (2007) Towards a philosophy for neuroethics. An informed materialist view of the brain might help to develop theoretical frameworks for applied neuroethics. *EMBO Rep, 8 Spec No*, S48-51. https://doi.org/10.1038/sj.embor.7401014

Farisco M (ed) (2023) Neuroethics and cultural diversity. ISTE-Wiley, London

Floreano D, Mattiussi C (2008) Bio-inspired Artificial Intelligence. MIT Press, Cambridge, MA

Frankfurt HG (1971) Freedom of the Will and the Concept of a person. J Philos 68(1):5–20

Frenkel C, Indiveri G (2022) 20–26 Feb. 2022). *ReckOn: A 28nm Sub-mm2 Task-Agnostic Spiking Recurrent Neural Network Processor Enabling On-Chip Learning over Second-Long Timescales* Paper presented at the 2022 IEEE International Solid- State Circuits Conference (ISSCC)

Friedrich AB, Mason J, Malone JR (2022) Rethinking explainability: toward a postphenomenology of black-box artificial intelligence in medicine. Ethics Inf Technol 24(1):8. https://doi.org/10.1007/s10676-022-09631-4

George D, Lazaro-Gredilla M, Guntupalli JS (2020) From CAPTCHA to Commonsense: how Brain can teach us about Artificial Intelligence. Front Comput Neurosci 14:554097. https://doi.org/10.3389/fncom.2020.554097

Gershman SJ (2023) What have we learned about artificial intelligence from studying the brain? Retrieved from https://gershmanlab.com/pubs/NeuroAI_critique.pdf website

Gerstner W, Kempter R, Van Hemmen JL, Wagner H (1996) A neuronal learning rule for sub-millisecond temporal coding. Nature 383(6595):76–78

Göltz J, Kriener L, Baumbach A et al (2021a) Fast and energy-efficient neuromorphic deep learning with first-spike times. Nat Mach Intell 3:823–835. https://doi.org/10.1038/s42256-021-00388-x

Göltz J, Kriener L, Sabado V, Petrovici MA (2021b) Fast and energy-efficient deep neuromorphic learning. ERCIM NEWS 125:17–18

Haider P, Ellenberger B, Kriener L, Jordan J, Senn W, Petrovici MA (2021) Latent equilibrium: a unified learning theory for arbitrarily fast computation with arbitrarily slow neurons. Adv Neural Inf Process Syst 34:17839–17851

Hassabis D, Kumaran D, Summerfield C, Botvinick M (2017) Neuroscience-inspired Artificial Intelligence. Neuron 95(2):245–258. https://doi.org/10.1016/j.neuron.2017.06.011

Hasson U, Nastase SA, Goldstein A (2020) Direct fit to Nature: an evolutionary perspective on Biological and Artificial neural networks. Neuron 105(3):416–434. https://doi.org/10.1016/j.neuron.2019.12.002

Häusser M, Mel B (2003) Dendrites: bug or feature? Curr Opin Neurobiol 13(3):372–383. https://doi.org/10.1016/S0959-4388(03)00075-8

Hawkins J (2021) *A Thousand Brains: A New Theory of Intelligence* (First edition). New York: Basic Books

Haydon PG, Carmignoto G (2006) Astrocyte control of synaptic transmission and neurovascular coupling. Physiol Rev 86(3):1009–1031. https://doi.org/10.1152/physrev.00049.2005

Hodgkin AL, Huxley AF (1952) A quantitative description of membrane current and its application to conduction and excitation in nerve. J Physiol 117(4):500–544. https://doi.org/10.1113/jphysiol.1952.sp004764

Hole KJ, Ahmad S (2021) A thousand brains: toward biologically constrained AI. SN Appl Sci 3(8):743. https://doi.org/10.1007/s42452-021-04715-0

Indiveri G, Liu S-C (2015) Memory and information Processing in Neuromorphic Systems. Proc IEEE 103(8):1379–1397. https://doi.org/10.1109/JPROC.2015.2444094

Jha MK, Morrison BM (2018) Glia-neuron energy metabolism in health and diseases: new insights into the role of nervous system metabolic transporters. Exp Neurol 309:23–31. https://doi.org/10.1016/j.expneurol.2018.07.009

Jordan J, Schmidt M, Senn W, Petrovici MA (2021) Evolving interpretable plasticity for spiking networks. eLife 10:e66273. https://doi.org/10.7554/eLife.66273

Kaplan A, Haenlein M (2019) Siri, Siri, in my hand: who's the fairest in the land? On the interpretations, illustrations, and implications of artificial intelligence. Bus Horiz 62:15–25

Kleene SC (1956) Representation of events in nerve nets and Finite Automata. Annals Math Stud 34:3–41

Korcsak-Gorzo A, Muller MG, Baumbach A, Leng L, Breitwieser OJ, van Albada SJ, Petrovici MA (2022) Cortical oscillations support sampling-based computations in spiking neural networks. PLoS Comput Biol 18(3):e1009753. https://doi.org/10.1371/journal.pcbi.1009753

Kungl AF, Dold D, Riedler O, Senn W, Petrovici MA (2019a) Deep reinforcement learning in a time-continuous model *Bernstein Conference*

Kungl AF, Schmitt S, Klahn J, Muller P, Baumbach A, Dold D, Petrovici MA (2019b) Accelerated physical emulation of bayesian inference in spiking neural networks. Front Neurosci 13:1201. https://doi.org/10.3389/fnins.2019.01201

Leng M, Kakadiaris IA (2018) 20–24 Aug. 2018). *Confidence-Driven Network for Point-to-Set Matching* Paper presented at the 2018 24th International Conference on Pattern Recognition (ICPR)

Lillicrap TP, Cownden D, Tweed DB, Akerman CJ (2016) Random synaptic feedback weights support error backpropagation for deep learning. Nat Commun 7:13276. https://doi.org/10.1038/ncomms13276

Lillicrap TP, Santoro A, Marris L, Akerman CJ, Hinton G (2020) Backpropagation and the brain. Nat Rev Neurosci 21(6):335–346. https://doi.org/10.1038/s41583-020-0277-3

Liu D, Yu H, Chai Y (2021) Low-power Computing with Neuromorphic Engineering. Adv Intell Syst 3:2000150. https://doi.org/10.1002/aisy.202000150

Maass W, Natschlager T, Markram H (2002) Real-time computing without stable states: a new framework for neural computation based on perturbations. Neural Comput 14(11):2531–2560. https://doi.org/10.1162/089976602760407955

Macpherson T, Churchland A, Sejnowski T, DiCarlo J, Kamitani Y, Takahashi H, Hikida T (2021) Natural and Artificial Intelligence: a brief introduction to the interplay between AI and neuroscience research. Neural Netw 144:603–613. https://doi.org/10.1016/j.neunet.2021.09.018

Marcus G, Davis E (2019) *Rebooting AI: building artificial intelligence we can trust* (First edition. ed.). New York: Pantheon Books

Markov NT, Vezoli J, Chameau P, Falchier A, Quilodran R, Huissoud C, Kennedy H (2014) Anatomy of hierarchy: feedforward and feedback pathways in macaque visual cortex. J Comp Neurol 522(1):225–259. https://doi.org/10.1002/cne.23458

McCulloch W, Pitts W (1943) A logical calculus of ideas immanent in nervous activity. Bull Math Biophys 5:115–133

Mehonic A, Kenyon AJ (2022) Brain-inspired computing needs a master plan. Nature 604(7905):255–260. https://doi.org/10.1038/s41586-021-04362-w

Metzinger T (2021) An argument for a global moratorium onSynthetic phenomenology. J Arti¯cial Intell Conscious 8(1):1–24

Millidge B, Tschantz A, Buckley CL (2022) Predictive coding approximates Backprop along Arbitrary Computation Graphs. Neural Comput 34(6):1329–1368. https://doi.org/10.1162/neco_a_01497

Nemitz P (2018) Constitutional democracy and technology in the age of artificial intelligence. Philosophical Trans Royal Soc A: Math Phys Eng Sci 376(2133):20180089. https://doi.org/10.1098/rsta.2018.0089

Nishant R, Kennedy M, Corbett J (2020) Artificial intelligence for sustainability: challenges, opportunities, and a research agenda. Int J Inf Manag 53:102104. https://doi.org/10.1016/j.ijinfomgt.2020.102104

Niv Y, Daniel R, Geana A, Gershman SJ, Leong YC, Radulescu A, Wilson RC (2015) Reinforcement learning in multidimensional environments relies on attention mechanisms. J Neurosci 35(21):8145–8157. https://doi.org/10.1523/JNEUROSCI.2978-14.2015

Ognibene D, Baldassare G (2015) Ecological active vision: four Bioinspired principles to integrate Bottom–Up and adaptive top–down attention tested with a simple camera-arm Robot. IEEE Trans Auton Ment Dev 7(1):3–25. https://doi.org/10.1109/TAMD.2014.2341351

Park, Tallon-Baudry C (2014) The neural subjective frame: from bodily signals to perceptual consciousness. Philosophical Trans Royal Soc B: Biol Sci 369(1641):20130208. https://doi.org/10.1098/rstb.2013.0208

Park, Lee J, Jeon D (2019) 17–21 Feb. 2019). *7.6 A 65nm 236.5nJ/Classification Neuromorphic Processor with 7.5% Energy Overhead On-Chip Learning Using Direct Spike-Only Feedback* Paper presented at the 2019 IEEE International Solid- State Circuits Conference - (ISSCC)

Parliament E (2020) The ethics of artificial intelligence: issues and initiatives (no. PE 634.452). EPRS | European Parliamentary Research Service, Bruxelles

Payeur A, Guerguiev J, Zenke F, Richards BA, Naud R (2021) Burst-dependent synaptic plasticity can coordinate learning in hierarchical circuits. Nat Neurosci 24(7):1010–1019. https://doi.org/10.1038/s41593-021-00857-x

Pennartz C (2009) Identification and integration of sensory modalities: neural basis and relation to consciousness. Conscious Cogn 18(3):718–739. https://doi.org/10.1016/j.concog.2009.03.003

Petit JM, Magistretti PJ (2016) Regulation of neuron-astrocyte metabolic coupling across the sleep-wake cycle. Neuroscience 323:135–156. https://doi.org/10.1016/j.neuroscience.2015.12.007

Petrovici MA, Vogginger B, Muller P, Breitwieser O, Lundqvist M, Muller L, Meier K (2014) Characterization and compensation of network-level anomalies in mixed-signal neuromorphic modeling platforms. PLoS ONE 9(10):e108590. https://doi.org/10.1371/journal.pone.0108590

Poirazi P, Brannon T, Mel BW (2003) Pyramidal neuron as two-layer neural network. Neuron 37(6):989–999. https://doi.org/10.1016/S0896-6273(03)00149-1

Poo M-m (2018) Towards brain-inspired artificial intelligence. Natl Sci Rev 5(6):785–785. https://doi.org/10.1093/nsr/nwy120

Pozzi I, Sander B, Roelfsema P (2020) Attention-Gated Brain Propagation: How the brain can implement reward-based error backpropagation. *Advances in Neural Information Processing Systems 33 (NeurIPS 2020)*

Prinz AA, Bucher D, Marder E (2004) Similar network activity from disparate circuit parameters. Nat Neurosci 7(12):1345–1352. https://doi.org/10.1038/nn1352

Rao RP, Ballard DH (1999) Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects. Nat Neurosci 2(1):79–87. https://doi.org/10.1038/4580

Renner A, Sheldon F, Zlotnik A, Tao L, Sornborger A (2021) *The Backpropagation Algorithm Implemented on Spiking Neuromorphic Hardware*. arXiv:2106.07030

Richards, Brockmann K, Boulanini V (2020) Responsible Artificial Intelligence Research and Innovation for International Peace and Security. Stockholm International Peace Research Institute, Stockholm

Richards, Lillicrap TP, Beaudoin P, Bengio Y, Bogacz R, Christensen A, Kording KP (2019) A deep learning framework for neuroscience. Nat Neurosci 22(11):1761–1770. https://doi.org/10.1038/s41593-019-0520-2

Sacramento J, Costa RP, Bengio Y, Senn W (2018) *Dendritic cortical microcircuits approximate the backpropagation algorithm*. Paper presented at the Proceedings of the 32nd International Conference on Neural Information Processing Systems, Montréal, Canada

Santucci VG, Baldassarre G, Mirolli M (2013) Which is the best intrinsic motivation signal for learning multiple skills? Front Neurorobot 7:22. https://doi.org/10.3389/fnbot.2013.00022

Saxe A, Nelli S, Summerfield C (2020) If deep learning is the answer, what is the question? Nat Rev Neurosci. https://doi.org/10.1038/s41583-020-00395-8

Schemmel J, Grubl A, Meier K, Mueller E (2006) *Implementing synaptic plasticity in a VLSI spiking neural network model* Paper presented at the The 2006 ieee international joint conference on neural network proceedings

Schönau A, Dasgupta I, Brown T, Versalovic E, Klein E, Goering S (2021) Mapping the dimensions of Agency. AJOB Neurosci 12(2–3):172–186. https://doi.org/10.1080/21507740.2021.1896599

Senn W, Dold D, Kungl AF, Ellenberger B, Jordan J, Bengio Y, Petrovici MA (2023) A neuronal least-Action Principle for Real-Time learning in cortical circuits. *bioRxiv*, 2023.2003.2025.534198. https://doi.org/10.1101/2023.03.25.534198

Sinz FH, Pitkow X, Reimer J, Bethge M, Tolias AS (2019) Engineering a less Artificial Intelligence. Neuron 103(6):967–979. https://doi.org/10.1016/j.neuron.2019.08.034

Song, Miller KD, Abbott LF (2000) Competitive hebbian learning through spike-timing-dependent synaptic plasticity. Nat Neurosci 3(9):919–926. https://doi.org/10.1038/78829

Song, Xu R, Lafferty J (2021) Convergence and alignment of Gradient Descent with Random Backpropagation weights. Adv Neural Inf Process Syst 34:19888–19898

Springer Nature Open Access eBooks https://doi.org/10.1007/978-3-030-69978-9 doi:10.1007/978-3-030-69978-9

Stahl BC (2021) *Artificial intelligence for a better future: an ecosystem perspective on the ethics of AI and emerging digital technologies SpringerBriefs in Research and Innovation Governance*, (pp. 1 online resource). Retrieved from Directory of Open Access Books https://directory.doabooks.org/handle/20.500.12854/67925

Stöckl C, Lang D, Maass W (2022) Structure induces computational function in networks with diverse types of spiking neurons. bioRxiv 202120052018444689. https://doi.org/10.1101/2021.05.18.444689

Summerfield C (2023) Natural General Intelligence: how understanding the brain can help us build AI. Oxford University Press, New York

Turing A (1950) Computing machinery and intelligence. Mind 59:433–460

Ullman S (2019) Using neuroscience to develop artificial intelligence. Science 363(6428):692–693. https://doi.org/10.1126/science.aau6595

Verplaetse J (2013) *The Moral Brain : Essays on the evolutionary and neuroscientific aspects of morality*

Vidal F (2017) *Being brains: making the cerebral subject* (First edition. ed.). New York: Fordham University Press

Walton N, Nayak BS (2021) Rethinking of marxist perspectives on big data, artificial intelligence (AI) and capitalist economic development. Technol Forecast Soc Chang 166:120576. https://doi.org/10.1016/j.techfore.2021.120576

Weil MM, Rosen LD (1995) A study of Technological Sophistication and Technophobia in University Students from 23 countries. Comput Hum Behav 11(1):95–133

Whitby B (1991) Ethical AI. Artif Intell Rev 5:201–204

Wiener N (1954) The human use of human beings; cybernetics and society. Houghton Mifflin, Boston

Woźniak S, Pantazi A, Bohnstingl T, Eleftheriou E (2020) Deep learning incorporating biologically inspired neural dynamics and in-memory computing. Nat Mach Intell 2(6):325–336. https://doi.org/10.1038/s42256-020-0187-0

Wunderlich T, Kungl AF, Muller E, Hartel A, Stradmann Y, Aamir SA, Petrovici MA (2019) Demonstrating advantages of Neuromorphic Computation: a pilot study. Front Neurosci 13:260. https://doi.org/10.3389/fnins.2019.00260

Yeung K (2018) Algorithmic regulation: a critical interrogation. Regul Gov 12(4):505–523. https://doi.org/10.1111/rego.12158

Zador A, Escola S, Richards B, Olveczky B, Bengio Y, Boahen K, Tsao D (2023) Catalyzing next-generation Artificial Intelligence through NeuroAI. Nat Commun 14(1):1597. https://doi.org/10.1038/s41467-023-37180-x

Zappacosta S, Mannella F, Mirolli M, Baldassarre G (2018) General differential hebbian learning: capturing temporal relations between events in neural networks and the brain. PLoS Comput Biol 14(8):e1006227. https://doi.org/10.1371/journal.pcbi.1006227

Zuboff S (2019) *The Age of Surveillance Capitalism: The Fight for a Human Future at the New Frontier of Power* (First edition). New York: PublicAffairs

Authors and Affiliations

Michele Farisco[1,2] · G. Baldassarre[3] · E. Cartoni[3] · A. Leach[4] · M.A. Petrovici[5] · A. Rosemann[4,6] · A. Salles[1,7] · B. Stahl[4,8] · S. J. van Albada[9,10]

✉ Michele Farisco
michele.farisco@crb.uu.se

[1] Centre for Research Ethics and Bioethics, Uppsala University, Uppsala, Sweden

[2] Biogem, Biology and Molecular Genetics Research Institute, Ariano Irpino (AV), Avellino, Italy

[3] Laboratory of Computational Embodied Neuroscience, Institute of Cognitive Sciences and Technologies, National Research Council of Italy, Rome, Italy

[4] Centre for Computing and Social Responsibility, De Montfort University, Leicester, UK

[5] Department of Physiology, University of Bern, Bern, Switzerland

[6] Centre for the Study of the Life Sciences, EGENIS, University of Exeter, Exeter, UK

[7] Institute of Neuroethics, Atlanta, GA, USA

[8] School of Computer Science, University of Nottingham, Nottingham, UK

[9] Institute of Neuroscience and Medicine (INM-6) Computational and Systems Neuroscience & Institute for Advanced Simulation (IAS-6) Theoretical Neuroscience & JARA-Institut Brain Structure-Function Relationships (INM-10), Jülich Research Centre, Jülich, Germany

[10] Institute of Zoology, University of Cologne, Cologne, Germany