Neuroethics (2022) 15:4
https://doi.org/10.1007/s12152-022-09484-0

ORIGINAL PAPER

On the Contribution of Neuroethics to the Ethics and Regulation of Artificial Intelligence

Michele Farisco · Kathinka Evers ·
Arleen Salles

Received: 16 March 2021 / Accepted: 10 December 2021

Abstract Contemporary ethical analysis of Artificial Intelligence (AI) is growing rapidly. One of its most recognizable outcomes is the publication of a number of ethics guidelines that, intended to guide governmental policy, address issues raised by AI design, development, and implementation and generally present a set of recommendations. Here we propose two things: first, regarding content, since some of the applied issues raised by AI are related to fundamental questions about topics like intelligence, consciousness, and the ontological and ethical status of humans, among others, the treatment of these issues would benefit from interfacing with neuroethics that has been addressing those same issues in the context of brain research. Second, the identification and management of some of the practical ethical challenges raised by AI would be enriched by embracing the methodological resources used in neuroethics. In particular, we focus on the methodological distinction between conceptual and action-oriented neuroethical approaches. We argue that the normative (often principles-oriented) discussion about AI will benefit from further integration of conceptual analysis, including analysis of some operative assumptions, their meaning in different contexts, and their mutual relevance in order to avoid misplaced or disproportionate concerns and achieve a more realistic and useful approach to identifying and managing the emerging ethical issues.

Keywords Ethics · Ethics of Artificial Intelligence · Artificial Intelligence Regulation · Neuroethics · Artificial Intelligence · Neuroscience

Introduction

AI research is growing rapidly raising various ethical issues related to safety, risks, trust, transparency, and accountability (among others), widely discussed in the literature [1]. One of the most recognizable outcomes of the ethical discussion is the publication of a number of guidelines intended to provide operational recommendations in response to the issues raised by AI design, development, and implementation [2–7]. Generally absent in the ethical discussion, however, is a consideration of whether the identification and management of many of those issues and the related regulation could benefit from the interfacing of AI ethics with other relevant fields such as neuroethics.

M. Farisco (✉) · K. Evers · A. Salles
Centre for Research Ethics & Bioethics, Uppsala University, Box 564, 751 22 Uppsala, Sweden
e-mail: michele.farisco@crb.uu.se

M. Farisco
Science and Society Unit, Biogem, Biology and Molecular Genetics Institute, Via Camporeale, Ariano Irpino, AV, Italy

A. Salles
Programa de Neuroetica, Centro de Investigaciones Filosoficas, Buenos Aires, Argentina

Published online: 03 February 2022

A summary of the historical development of AI as a discipline can help us to set the stage for exploring the potential connection between AI ethics/regulation and neuroethics.

At the beginning of AI research, an underlying conception of intelligence as logical reasoning about symbols prevailed. Accordingly, the first AI systems relied on the manipulation of symbols through logical rules. This approach has been baptized "Good Old-Fashioned AI" (GOFAI) [8], or simply "Symbolic AI". One of its main achievements is the development of so-called "expert systems." Introduced in the 70s, expert systems are computer systems emulating human decision-making and implementing if-then rules in order to infer new knowledge starting from pre-programmed facts and rules [9]. Notwithstanding important results achieved (including the famous 1997 IBM Deep Blue), symbolic AI has a number of shortcomings, including its limited ability to know facts outside its original datasets, its dependency on original programming by the researcher, and its rather inflexible architecture [10].

In the 1990s, AI research goals changed: rather than trying to emulate the computations of natural intelligence, the aim became building intelligent agents, i.e., "entities that sense their environment and act upon it" [11]. Within this new framework, AI is not necessarily or completely connected to algorithms (i.e., top-down instructions) for decision-making. Instead, it is an adaptive process of computation interacting with the environment for more efficient decision-making, that is, it displays a kind of bottom-up decision process implemented through reward signals. This strategy is at the core of contemporary Deep Learning Architectures, that basically rely on instructions that allow the system to learn from incoming data rather than on strong, top-down programming [12].

Importantly, although already inspired by biology from the beginning, AI research is increasingly looking at the brain as inspiration for more flexible and adaptable strategies that might result in more human-like systems.

The mutual relationship between neuroscience and AI has been critical for advancing both fields. On one hand, as outlined by Ullman, at the outset of AI research in the early 50s, the only known systems carrying out complex computations were biological nervous systems. Accordingly, AI researchers productively used knowledge about the brain as a source of inspiration [13] seeking to successfully emulate brain activities. To illustrate, research on neural networks and their translation onto computational systems took inspiration from how neurons in the brain function [14, 15]. On the other hand, one initial goal of creating correspondences between the functionalities of AI and the human brain was to promote a better understanding of the brain, the self, and the behaviour of biological organisms [16, 17]. Therefore, each field has arguably benefitted from its relation and mutual collaboration [13, 18, 19].

However, while neuroscience and AI as scientific fields have recognizable links and a productive collaboration, the same cannot be said of neuroethics and the AI ethics. Even when their interfacing has been recommended [20, 21] this is uncommon and often practically challenging because of the difference between their specific objects of interest and languages.

Here, we want to explore the advantages of the interfacing of neuroethics and AI ethics/regulation. We propose that their mutual engagement is desirable both for content-related and methodological reasons. Regarding content, AI ethics/regulation can benefit from the fact that underlying some specific ethical and societal issues raised by AI (e.g., the creation of potentially conscious AI, impact on autonomy and personal identity of AI-based brain implants, AI enabled or assisted monitoring of employment/academic performance; etc.), are notions such as intelligence and consciousness, and topics such as the ontological and ethical status of humans and machines [1] that neuroethics has been addressing since its beginnings. In turn, neuroethics as a field would benefit from interfacing with AI ethics/regulation insofar as it could be pushed to re-think how those concepts and notions are conceived.

Regarding methodology, recent emphasis within neuroethics on the importance of integrating conceptual and philosophical analysis into the scientific agenda from the beginning can enhance AI ethics/regulation´s methodology. We suggest that the identification and management of some of the practical ethical challenges raised by AI and AI-assisted technologies would be enriched by embracing some of the methodological resources of neuroethics.

An Excursion Into Neuroethics

Neuroethics is an interdisciplinary field that addresses scientific and philosophical, but also ethical, legal, social, and cultural questions raised by neuroscience and related technologies [22–25]. Its methodology can be conceptual, empirical, and normative (or a combination) depending on the perspective one wishes to emphasize [26]. Since the 2002's Dana Foundation Neuroethics Conference and onwards, this field has often been conceived in two ways: as a type of applied ethics aimed at providing a repertoire of ethical approaches to address the practical ethical and societal concerns raised by neuroscience research and its applications, e.g., privacy and the protection of neural data; or as an empirical, descriptive approach focusing on how neuroscientific findings can inform theoretical and practical issues, e.g., what is moral reasoning, how to understand choice ([27, 28]). More recently, a more basic research-oriented and conceptual approach, i.e. fundamental neuroethics [29, 30] has been gaining traction. Fundamental neuroethics takes as a starting point the view that conceptual analysis plays an important role not only in illuminating key operative notions (e.g. consciousness, self, and human identity), but also in examining issues such as what is the understanding of the same notions in different contexts (i.e., ethics and neuroscience) and their mutual relevance, how neuroscientific knowledge is constructed, what its underlying assumptions are and how they are justified, how results may be interpreted, and why or how empirical knowledge of the brain can be relevant to philosophical, social, and ethical concerns [26, 31, 32].

A fundamental neuroethics approach focuses on epistemic issues highlighting their impact on normative discussions. Importantly, it attempts to address conceptual gaps that may arise between neuroscientific and philosophical, including ethical, language. Without denying neuroscience's conceptual elements, the field is still conceptually limited regarding the potential normative biases and impacts of its methodologies, language, categories, and emerging results. Some of these limitations might be partly due to the fact that neuroscience is a relatively young science: it is reasonable to think that as it evolves so will its conceptual repertoire, including a form of conceptual self-assessment. However, other issues seem not to depend on the development of the field but appear intrinsic in nature. Consider, for example, neuroscience's strong focus on third-person perspectives. As it is in general presently pursued, neuroscience focuses primarily on objective perspectives that can be accounted for in third-person terms. However, this focus is epistemically insufficient to assess notions such as consciousness, experience, or self-awareness: since essentially subjective, they need to be approached from both third-person and first-person perspectives. This insight is not new. Neurophenomenology, as introduced by the neuroscientist Varela [33] and later developed by philosophers like Thompson [34], acknowledges the need to combine first and third person perspectives. Even further, the French neuroscientist Jean-Pierre Changeux wrote in 1983 that epistemologically, an adequate understanding of our subjective experiences must take “both introspective information and data gathered from anatomical observations and physical measurement” into account. (Changeux 82: 168). This "informed materialism" has been taken up and developed in fundamental neuroethics (Evers 83), but in neuroscience generally it is not commonly expressed or pursued.

What we here have called fundamental neuroethics explicitly aims at informing the ethical reflection on neuroscience and its applications. In this respect, fundamental neuroethics is characterized by three main features, related to content and to methodology [35]: it pursues *foundational analyses* within a *multidisciplinary research domain* using an *interdisciplinary methodology*. Topically, fundamental neuroethics pursues basic research and analyses foundational concepts and methods used in the neuroscientific investigations of notions like, for example, identity or consciousness [36]. These analyses necessarily involve both empirical scrutiny of the science in question and philosophical analyses of the concepts involved. This requires contributions from different disciplines, including the natural and social sciences, philosophy of science, philosophy of language, philosophy of mind, and moral philosophy and, accordingly, the combination of a variety of methods, e.g., empirical and conceptual methods depending on the different disciplines. Methodologically fundamental neuroethics is ipso facto interdisciplinary.

These features distinguish a fundamental neuroethics approach from normative or purely empirical neuroethical approaches. However, since all forms of neuroethics require some foundational analyses in order to be viable, it could also be said that all forms of neuroethics must somehow involve, or be developed on the basis of fundamental research in neuroethics (whether or not the label "fundamental neuroethics" is used).

Actual AI Regulation

Ethical reflection on AI is growing rapidly. There is a lively community of AI ethics scholars that address a variety of applied and conceptual issues. Both "academic" AI ethics and the AI field itself include explicit theoretical and foundational reflection [1, 37–42].

In addition to academic discussions, a number of practical guidelines and recommendations [43–45] aimed at supporting and improving policy making on AI and its applications [2, 4, 46–48] have been issued. Whether it is codes of ethics produced by professional bodies for their members and practitioners or by other regulatory bodies (IEEE Code of Ethics, 88; Simulationist Code of Ethics, 84) or governmental reports, statements, and declarations produced by ad hoc committees tasked with drafting policy documents (Barcelona Declaration for the proper development and usage of artificial intelligence in Europe, 88; High-Level Expert Group on AI, 86; OECD, 87), there are a number of such documents both in the public and private sectors.

Despite diverse backgrounds, these documents tend to have a common general objective, to focus on common themes, and to share a methodology. First, while targeting different non expert stakeholders (e.g., policymakers, general public, professional associations, etc.) [5], they aim to provide an adequate basis for achieving an ethically sound design, development, and application of AI and bio-inspired robotics [5–7, 49]. Second, these documents usually address potential limitations of human beings vis a vis AI, pointing to human beings' limited knowledge in the area of AI and robotics, their limited decision-making capacity regarding a technology they don't fully understand, and their limited power to control the development of the technology and its impacts. Finally, these documents tend to follow traditional practical ethics theorizing and methodology, usually taking inspiration from professional ethics codes, e.g. medical ethics. They are generally characterized by a top-down approach, often starting from a few classical fundamental principles rebaptized in this context as "human centred values" (e.g., human dignity, respect for autonomy, non-maleficence, beneficence, justice, and fairness) and complemented with other principles more tailored to technology in general and AI in particular (accountability, effectiveness, trust, transparency, and explicability, among others) which are taken to jointly provide an adequate basis for achieving an ethically sound design, development, and application of AI [5–7, 49].

While recognized as an important step forward, these documents present a few shortcomings. A common objection found in the literature is that current AI guidelines risk being ineffective because of their level of abstraction and the difficulty in translating them into action-oriented recommendations [7, 50, 51]. Accordingly, there have been recent attempts to focus on how actual organizations understand and address the ethical issues raised by AI [52] and to develop frameworks for actionability of the guidelines [53, 54]. Efforts to address this issue include providing preliminary landscape assessments (so as to bridge the distinction between what should be done and what can actually be done), calling for a richer engagement with diverse representative publics (so as to expand the scope of voices typically heard in the discussion of the issues), and for the creation of inclusive mechanisms for implementation.

These documents raise two additional concerns less discussed in the literature. One is that those guidelines often appear to conceive of AI ethics as a type of applied ethics that is methodologically principle-oriented. A second and related concern has to do with what appears to be a certain lack of theoretical and conceptual engagement with the issues, even when such reflection would be key to addressing the type of practical concerns that the documents attempt to address. More specifically, what seems lacking is not simply a conceptual analysis of the terms used, but primarily a thorough clarification of their different meaning in different contexts and of their mutual relevance. On this specific point the model offered by neuroethics can complement the conceptual work already pursued in AI ethics and informing AI regulation.

What Ethics for AI?

The Ethics Guidelines for Trustworthy AI developed by the High Level Experts Group on AI set up by the European Commission defines AI ethics as "a sub-field of applied ethics, focusing on the ethical issues raised by the development, deployment and use of AI" [2] (p.11). Analogously the IEEE Ethically Aligned Design document calls for "integrating applied ethics into education and research to address the issues of autonomous and intelligence systems" [3] (p. 59).

Concerning principles, the Guidelines from the High Level Expert Group say that "Trustworthy AI has three components, which should be met throughout the system's entire life cycle: (1) it should be lawful, complying with all applicable laws and regulations (2) it should be ethical, ensuring adherence to ethical principles and values and (3) it should be robust, both from a technical and social perspective since, even with good intentions, AI systems can cause unintentional harm" [2] (p. 2).

Analogously the Introduction of the IEEE document says: "As the use and impact of autonomous and intelligent systems (A/IS) become pervasive, we need to establish societal and policy guidelines in order for such systems to remain human-centric, serving humanity's values and ethical principles" [3] (p. 2).

These documents, which are illustrative among others that manifest a similar approach, suggest two things: 1. AI ethics is a type of applied ethics; 2. Identification and management of ethical issues raised by AI are importantly shaped by principle-oriented approaches (i.e., identification and application of principles within a top-down approach). Both points deserve further examination.

Recently, some authors have proposed that ethical analysis of AI be divided into three main content-oriented streams: specific issues related to the application of AI, e.g. Machine Learning; social and political issues arising in the digital society; and metaphysical questions about the nature of reality and humanity [55]. Whilst we consider this a theoretically insightful distinction, we should bear in mind that, in practice, the three levels of analysis should not be fully disconnected from each other, both topically and methodologically. We show below how a conceptually-informed ethical approach can help in implementing the intersection between the three streams, showing their respective relevance.

Relying heavily on an applied ethics methodology, current AI guidelines tend to focus on the first two applied streams more than the third, more theoretical one. This focus, we suggest, risks making the analysis incomplete, because the specific assessment of the ethical issues arising from AI application or from the digital society cannot abstract away from the conceptual (e.g., metaphysical) assumptions about relevant terms. Without denying the relevance of applied ethics, it is crucial to recognize the necessary contribution of conceptual analysis in understanding the ethical relevance of AI and its potential impact on human wellbeing. Preliminary conceptual clarification of terms, of their use in different contexts, and of their mutual relevance is also key in order to engage with AI developers so as to have the kind of dialogue that will ultimately result in an ethically sustainable AI.

To illustrate, in general, relevant guidelines and related policy-oriented documents tend to take the meaning of ambiguous notions such as intelligence, autonomy, consciousness, trustworthiness, and purposiveness for granted, and proceed to attribute them to or predicate them of AI [56]. However, how terms are conceived shapes and often biases the normative discussion, as illustrated in [57]. Indeed, careful conceptual analysis might reveal that some notions might be inappropriate or misleading when translated to AI from a different linguistic and semantic context [58]. In turn, AI should not be naively assumed as a large, uniform field lest we simplify drastically its complexity and neglect the variety of disciplines subsumed under the AI label with their own methods and approaches.

While it might be true that some recommendations of AI guidelines could be valid regardless of how some notions are understood, in the absence of a more thorough conceptual analysis the applicability and effectiveness of those guidelines will remain limited. Notably, some have argued that if AI regulation wants to go beyond a banning exercise to foster a proactive and positive ethical culture within AI, it needs to develop a vision of the good life and the good society, and to think how AI can contribute to them [1] but again, such a grand vision would require considerable conceptual clarity about the main notions, an assessment of their relevance to AI, and of the impact of AI upon them.

The prevalent AI ethics strategy as expressed in relevant guidelines is a type of "*ex post* AI ethics" (*ex post* meaning after the event, in distinction from *ex ante*, in anticipation to the event, both in terms of foundational reflection and timely analysis, cf. below), insofar as ethical reflection on AI appears mostly concerned with the recognition and management of actual or potential issues raised by AI and its impact on society often at the expense of more fundamental conceptual issues. While this distinction may bear some similarity with the distinction between re-active and pro-active ethics (with the former limited to finding out the right solutions for actual issues and the latter engaged in anticipating potential ethical challenges) it is not only a question of temporal precedence, but crucially one about of foundational reflection. Importantly, the two kinds of approaches are not so clearly distinguished in reality, the point being that the emphasis on one rather than the other eventually affect the ethical reflection and related recommendations.

The Conceptual Approach in the Ethical Analysis of AI: The Potential Contribution of Neuroethics

Recent literature highlights the importance of integrating ethical reflection from the beginning of the AI research and design process [59]. Ideally, conceptual analysis of ethically salient notions should be integral to the process of development of the technology itself thereby effectively shaping AI from the very beginning. This approach would be consistent with the ethics by design strategy, i.e. with the efforts to sustainably drive ethical behaviour in technology design, development, and use [60, 61]. Considering the concerns raised by AI, priority should be given to a reflection on foundational notions and interpretative categories (e.g., the values informing AI design and development) and the values considered essential to society, how to understand them, how they might be affected by AI development, and possibly how to align AI with them [62, 63].

Neuroethical reflection can play an important role in addressing some of the ethical and regulatory issues raised by AI especially considering that many of them touch on topics that have traditionally been object of neuroethical reflection. To illustrate this point, we can take the notions of intelligence and consciousness as illustrative cases.

Regarding intelligence, it is true that behavioural flexibility and innovation capacity, which a biological account of intelligence recognizes as critical [64], might also be expressed, at least partly, by AI. However, the needs and goals constitutive of these abilities are, at least to date, substantially different in biological organisms and in AI [65]. In the first case, they are the result of an emotional interaction with the world, i.e. the ability to evaluate external stimuli differentiating their respective salience for fulfilling specific goals. While some AI applications are able to recognize/label human emotions they do so in terms of information processes. This means that AI does not understand in the sense of empathizing with emotions as humans who can experience them do. At least at present, AI arguably lacks abilities that neuroethical reflection recognizes as ethically relevant and salient, such as what is generally called emotional and social intelligence [66, 67], notwithstanding some relevant conceptual and technical advancements in this direction [68], and a theory of mind. The conceptual and ethical reflection about intelligence provided by neuroethics can enrich the ethical discussion on the use of AI to replace humans in some specific contexts (for example, senior and child care) by fostering ethical reflection on how humans understand activities such as caring and what actions humans tend to value. Indeed, recognition of AI's lack of social and emotional intelligence might allow us to develop at least one criterion for assessing some AI uses, namely that regardless of the presence of actions commonly labelled as intelligent, the lack of some features (like emotional experience) typically taken to be morally relevant and valuable calls for caution when considering the role of AI in some specific human activities [65].

With regards to the notion of consciousness, some consider the hypothetical development of conscious AI one of the most pressing ethical issues [62, 69]. However, before engaging with the question of whether it is or is not (presently) attributable to AI, whether conscious AI could or should be developed [70–72], and which ethical consequences would follow, it is necessary to analyze the notion of consciousness conceptually [73, 74] . Importantly, the conceptual conceivability of conscious AI depends on the starting definition of consciousness. For instance, if conceived as a biological phenomenon for which the biological component plays a crucial role [75, 76], then simulating consciousness artificially would not be achievable, while if understood within a functionalist and computational framework its artificial implementation would be conceptually consistent even though practically not possible [73].

Concerning the technical feasibility of conscious AI, current deep learning neural networks (DLNNs) [12] show an impressive ability to recognize and classify complex input patterns, but they still make gross mistakes in such tasks. These mistakes reveal the inability of DLNNs to get the overall meaning or emotional significance of a scene, i.e. the human ability for generalization, conceptual learning, selective attention [77] and arguably consciousness [78].

Neuroethics has been extensively engaged in the conceptual analysis of consciousness and related ethical implications, also questioning its supposedly unique ethical relevance and implications for the ethics of non-conscious humans and non-human beings [79, 80]. To that extent, it can assist AI ethics in the assessment of potentially conscious AI and in informing relevant regulation.

From a practical point of view, several additional ethical issues arise at the intersection of neuroscience and AI. Consider, for example, AI-based neuro-enhancement technologies; the use of AI in BCI assistive and rehabilitative applications; the use of AI for monitoring and predicting individual choices; AI-based personality profiling technologies; among others. While these applications raise the well-known issues of informed consent, privacy, liability, transparency, and the possibility of dual-use and misuse, they do so in the context of some fundamental assumptions about the ontological and moral status of humans and ingrained beliefs about what human agency, autonomy, and intelligence are and mean. These are core issues that neuroethics has been addressing in the context of brain research, so it may arguably complement AI ethics in general and the development of regulation in particular.

In turn, AI ethics might offer neuroethics both the opportunity to delve deeper into some issues by pointing to aspects often unattended and possibly to approach old issues in a different way, e.g. how to understand human identity and humanness, agency and its possible attribution to non-human entities. This suggests that neuroethics and AI ethics would mutually benefit from interfacing when addressing some specific issues and from complementing their categories and methodologies.

In addition to these content-related reasons, neuroethics can help AI ethics and regulation also for methodological reasons. More specifically, fundamental neuroethics offers a potentially relevant methodological model. As mentioned above, fundamental neuroethics focuses on foundational issues through multidisciplinary research implementing an interdisciplinary methodology. As described by [35], the analysis performed by fundamental neuroethics has three main, interconnected foci: scientific descriptions (in terms of goals and purposes, methods, theoretical underpinnings, results), philosophical analyses (focused on meaning of key concepts and terminology, selection of methods, comparative assessment of theoretical underpinnings, possible scientific interpretations of results), and ethical and social considerations (e.g., about reliability of goals, adequate communication strategy, compliance of research conduct, ethical and social acceptability of the research). This kind of interdisciplinary reflection aimed at informing the ethical analysis is arguably highly relevant for making AI ethics/regulation more robust and effective.

It is true that a number of interdisciplinary theoretical approaches might provide the necessary contribution for overcoming the type of conceptual limitation relevant to the ethical analysis of AI. For instance, 4E cognition (embodied, embedded, enactive, and extended) relies on the premise that cognition is shaped by an interplay between brain, body, physical, and social environment [81]. Similarly, as mentioned above, neurophenomenology implements an interdisciplinary approach which elaborates a multidimensional conceptual analysis of notions like mind, intelligence, and consciousness [33]. The theoretical methodology developed by both these illustrations might be used to integrate first person, social, and emotional dimensions in the analysis of notions like intelligence and cognition by AI scholars. A fundamental neuroethics approach shares the interest in the general analysis of these dimensions, but importantly it is specifically concerned with informing ethical analysis. As illustrated in Fig. 1[1] using this approach

[1] For the sake of simplicity, this illustration of a conceptually-informed ethical analysis does not include the background normative dimensions that affect and possibly bias conceptual

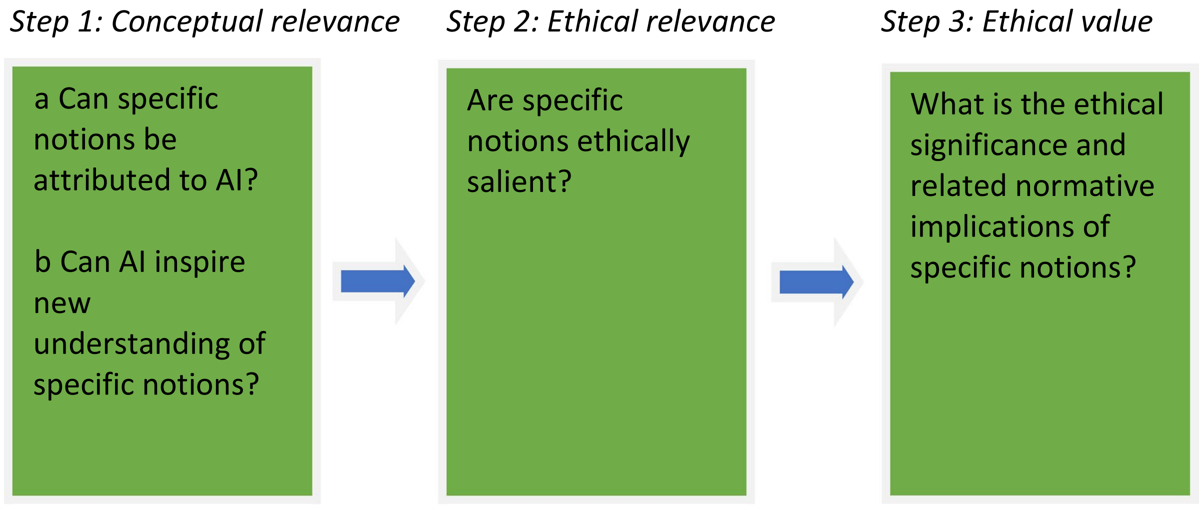
Step 1: Conceptual relevance
a Can specific notions be attributed to AI?
b Can AI inspire new understanding of specific notions?
Step 2: Ethical relevance
Are specific notions ethically salient?
Step 3: Ethical value
What is the ethical significance and related normative implications of specific notions?

Fig. 1 A conceptually-informed ethical analysis applied to AI

to analyse AI requires three steps: first, a determination of whether certain notions are attributable to AI and whether AI can provide a new understanding of those notions; second, if attributable to AI, whether they are ethically relevant (i.e., salient); and, finally, if so, whether they are ethically valuable (i.e., assessing their ethical significance and normative implications).

While holding that such conceptual component is key in the ethical analysis of AI, we are aware that questions might be raised regarding whether it can be practically implemented and how. For this, we can focus on the EU funded Human Brain Project (HBP). In the HBP philosophers and neuroethicists are collaborating with experts in AI, modelling, and bio-inspired robotics in order to identify crucial conceptual issues (e.g., how intelligence is understood in ethics and AI/robotics? What does “simulate brain activity” mean? How is "autonomous agent" defined by scientists? What is the technical meaning of “learning”? What is the role of biological plausibility? etc.) that, if not adequately assessed, might ultimately distort the ethical analysis of emerging results and related recommendations. This collaboration is pursued through different strategies/methods, including:

- Embedded research of ethicists/philosophers in scientific work packages
- Structured interviews with scientific researchers
- Co-authorship of interdisciplinary papers about emerging topics
- Co-authorship of opinion documents on the impact of neuroscience and AI on society
- Engagement activities with the public on the societal and ethical implications of scientific research, including their identification and strategies for assessing them.

This is, of course, work in progress, but still an example of an attempt to integrate conceptual, normative, and empirical analyses in order to shape ethical reflection on AI.

Towards a Collaboration of Neuroethics and AI Ethics

The above gives support to two main points. First, neuroethics and AI ethics would mutually benefit from collaboration (i.e., intersecting their language and categories) in order to identify, assess, and suggest effective strategies for managing ethical issues, particularly those arising at the intersection of

Footnote 1 (continued)
analysis. In fact, the meaning we give to terms is often the result of implicit or explicit evaluations.

neuroscience and AI. Neuroethics can offer relevant expertise in analyzing some issues and suggesting an effective methodology, while AI ethics might offer both the opportunity to recognize new issues and possibly to approach old issues in a different way, e.g. how to understand human identity and humanness, agency and its possible application to non-human entities, to mention a few. Second, even though subsuming AI ethics under the umbrella of applied ethics as actual guidelines seem to do covers important aspects of the ethical reflection on AI, the normative discussion of important applied issues (e.g., privacy, data protection, the impact of AI on job market, etc.) is incomplete and potentially misguided without advanced conceptual ethical reflection. This reflection should concern both the methodology and the content of AI ethics/regulation. Methodologically, the principle-oriented view of AI ethics usually taken for granted in relevant regulation is not free of controversy and in need of more explicit justification [50]. Particularly, it risks making AI ethics/regulation ineffective by being too vague and not sufficiently action-oriented. The theoretical and foundational reflection already present in academic AI ethics can be further enriched by an interfacing with neuroethics, because this discipline has been engaging with a number of relevant issues and offers a methodological model of interdisciplinary collaboration aimed at informing ethical deliberation. This is particularly relevant to AI regulation. With regard to contents, without a previous examination of key concepts and an analysis of fundamental (e.g., definitional) ethical issues, regulatory/normative reflection and related policy are blind. Thus, we propose that AI regulation would benefit from conceptual analysis as a starting point, an "*ex ante* AI ethics" that examines foundational issues whose clarification is necessary to achieve a balanced analysis of applied issues. Indeed, a preliminary conceptual analysis of foundational notions (e.g., autonomy or intelligence), including their meaning in different contexts and their mutual relevance, might reveal that some fears about risks or threats allegedly posed by AI (e.g., AI taking control of our society) are either misplaced or unrealistic, and that, even if realistic, such threats are less imminent than others which means that focusing on them might distract us from more pressing and current concerns.

Conclusion

AI systems are increasingly present in social contexts, from entertainment to work, from healthcare to education, among others. As use of AI technology becomes more common in a number of social domains, it becomes more ethically and socially impactful. Avoiding extreme attitudes (e.g. blind optimism regarding AI or disproportionate alarmism about its impact) requires scrutiny of concepts such as intelligence, action, interest, goal, consciousness, and autonomy, among others that are typically used when discussing AI. We have argued that a closer collaboration between neuroethics and AI ethics and then regulation is key for two reasons: 1. In assessing some of the ethical issues that increasingly arise at the intersection of neuroscience and AI, neuroethics and AI ethics can mutually enhance each other ultimately leading to more conceptually sound regulation; 2. the kind of interdisciplinary conceptual analysis illustrated by fundamental neuroethics can serve as a methodological model for AI ethics in general and for AI regulation in particular.

Specifically, conceptual reflection that goes beyond clarification of terms and focuses also on how concepts are elaborated and interpreted in different contexts and if/how they are relevant to each other should be an integral part of AI ethics/regulation. Accordingly, AI ethics can precede and more effectively inform the actual development of AI (*ex ante* AI ethics), complementing the practical analysis of AI consequences (*ex post* AI ethics) and making AI guidelines more effective, proactive, and action-inspiring. In this way, we may avoid the risk of being like the owl of Minerva that starts its flight only at twilight with limited possibilities to shape the state of affairs and understand them in time.

Authors' Contributions MF and AS conceived plan and structure of the paper. MF wrote the first draft and was responsible for general ideas. AS and KE contributed to revising and developing ideas. MF, AS, and KE were involved in revising the manuscript critically, gave final approval of the paper and agreed to be accountable for all aspects of the work.

Funding Open access funding provided by Uppsala University. This research was supported by the European Union's Horizon 2020 Framework Programme for Research and

Innovation under the Specific Grant Agreement No. (945539) (Human Brain Project SGA3).

Declarations

Conflicts of Interest/Competing Interests Not applicable.

References

1. Coeckelbergh, Mark., and AI ethics. 2020. The MIT press essential knowledge series. *Cambridge*. MA: The MIT Press.
2. HLEG. 2019. *Ethics Guidelines for Trustworthy AI*. European Commission: Brussels.
3. IEEE. 2019. *Ethically Aligned Design: A Vision for Prioritizing Human Well-being with Autonomous and Intelligent Systems, First Edition*.
4. EGE. 2018. *Statement on Artificial Intelligence, Robotics and ´Autonomous´ Systems*. European Commission: Brussels.
5. Jobin, A., M. Ienca, and E. Vayena. 2019. The global landscape of AI ethics guidelines. *Nature Machine Intelligence* 1: 389–399.
6. Ryan, M. and B.C. Stahl. 2020. Artificial intelligence ethics guidelines for developers and users: clarifying their content and normative implications. *Journal of Information, Communication and Ethics in Society.*
7. Hagendorff, T. 2020. The Ethics of AI Ethics: An Evaluation of Guidelines. *Minds & Machines* 30: 99–120.
8. Haugeland, J., and Artificial intelligence : the very idea. 1985. *Cambridge, MA*, 287. London: MIT Press.
9. Jackson, P. 1998. *Introduction to expert systems*. 3rd ed. International computer science series. Harlow: Addison-Wesley. xvii,542p.
10. Russell, S., and P. Norvig. 2010. Artificial Intelligence: International Version: A Modern Approach. *Englewood Cliffs*. NJ: Prentice Hall.
11. Russell, S. 2016. Rationality and Intelligence: A Brief Update. In *Fundamental Issues of Artificial Intelligence*, ed. V.C. Müller, 7–28. Cham: Springer International Publishing.
12. LeCun, Y., Y. Bengio, and G. Hinton. 2015. Deep learning. *Nature* 521 (7553): 436–44.
13. Ullman, S. 2019. Using neuroscience to develop artificial intelligence. *Science* 363 (6428): 692–693.
14. McCulloch, W., and W. Pitts. 1943. A logical calculus of ideas immanent in nervous activity. *Bull. Math. Biophys.* 5: 115–133.
15. Kleene, S.C. 1956. Representation of Events in Nerve Nets and Finite Automata. *Annals of Mathematics Studies* 34: 3–41.
16. Prescott, T. 2015. Me in the machine. *New Scientist* 225 (3013): 36–39.
17. Prescott, T. and D. Camilleri. 2018. *The Synthetic Psychology of the Self*, in *Cognitive Architectures*, M. Aldinhas Ferreira, J. Silva Sequeira, and R. Ventura, Editors. Springer: Cham, Switzerland.
18. George, D., M. Lazaro-Gredilla, and J.S. Guntupalli. 2020. From CAPTCHA to Commonsense: How Brain Can Teach Us About Artificial Intelligence. *Front Comput Neurosci* 14: 554097.
19. Hassabis, D., et al. 2017. Neuroscience-Inspired Artificial Intelligence. *Neuron* 95 (2): 245–258.
20. Ienca, M. 2019. *Neuroethics meets Artificial Intelligence*, in *The Neuroethics Blog*.
21. Ienca, M., and K. Ignatiadis. 2020. Artificial Intelligence in Clinical Neuroscience: Methodological and Ethical Challenges. *AJOB Neurosci* 11 (2): 77–87.
22. Illes, J., B.J. Sahakian, The Oxford, and handbook of neuroethics. Oxford library of psychology. 2011. *Oxford*, 935. New York: Oxford University Press. xxxix.
23. Johnson, L.S.M. and K.S. Rommelfanger. 2018. *The Routledge handbook of neuroethics*. Routledge handbooks in applied ethics. New York: Routledge, Taylor & Francis Group. xix, 509 pages.
24. Levy, N., and Neuroethics. 2007. *Cambridge, UK*, 346. New York: Cambridge University Press. xiii.
25. Marcus, S. C. A., and D. Foundation. 2002. *Neuroethics : mapping the field : conference proceedings, May 13-14, 2002, San Francisco, California*. New York: Dana Press. vii, 367 p.
26. Evers, K., A. Salles, and M. Farisco. 2017. *Theoretical framing of neuroethics: the need for a conceptual approach*, in *Debates about Neuroethics: perspectives on its development, focus and future*, E. Racine, Aspler, J., Editor. Springer International Publishing: Dordrecht. p. 89–107.
27. Marcus, S., and A. Charles. 2002. Dana Foundation., *Neuroethics : mapping the field : conference proceedings, May 13-14, 2002, San Francisco, California*. New York: Dana Press. vii, 367 p.
28. Roskies, A. 2002. Neuroethics for the new millenium. *Neuron* 35 (1): 21–3.
29. 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: p. S48–51.
30. Evers, K. 2009. *Neuroetique. Quand la matière s'éveille*. 2009, Paris: Odile Jacob.
31. Farisco, M., A. Salles, and K. Evers. 2018. Neuroethics: A Conceptual Approach. *Camb Q Healthc Ethics* 27 (4): 717–727.
32. Salles, A., K. Evers, and M. Farisco. 2019. The need for a conceptual expansion of neuroethics. *AJOB Neuroscience* 10 (3): 126–128.

33. Varela, F. 1996. Neurophenomenology: A methodological remedy for the hard problem. *Journal of Consciousness Studies* 3 (4): 330–349.
34. Thompson, E. 2007. *Mind in life : biology, phenomenology, and the sciences of mind*. Cambridge, Mass.: Belknap Press of Harvard University Press. xiv, 543 p., 8 p. of plates.
35. Evers, K. *Fundamental Neuroethics*, in *Neuroethics and cultural diversity*, M. Farisco, Editor. Forthcoming, ISTE-Wiley: London.
36. Evers, K., A. Salles, and M. Farisco. 2017. Theoretical Framing of Neuroethics: The Need for a Conceptual Approach. In *Debates About Neuroethics: Perspectives on Its Development, Focus, and Future*, ed. E. Racine and J. Aspler, 89–107. Cham: Springer International Publishing.
37. Floridi, L. 2013. *The ethics of information*. First edition. ed. Oxford: Oxford University Press. xix, 357 pages.
38. Taddeo, M.R. 2009. *Defining Trust and E-trust: Old Theories and New Problems.* International Journal of Technology and Human Interaction (IJTHI) Official Publication of the Information Resources Management Association 5(2): 23–35.
39. Taddeo, M.R. 2010. Modelling Trust in Artificial Agents, A first Step Towards the Analysis of E-Trust. *Minds & Machines* 20: 243–257.
40. Vakkuri, V. and P. Abrahamsson. 2018. *The Key Concepts of Articficial Intelligence*, in *2018 IEEE International Conference on Engineering, Technology and Innovation (ICE/ITMC)*. Stuttgart. p 1–6.
41. Hildt, E., K. Laas, and M. Sziron. 2020. *Editorial: Shaping Ethical Futures in Brain-Based and Artificial Intelligence Research.* Sci Eng Ethics.
42. Tolmeijer, S., et al. 2020. *Implementations in Machine Ethics: A Survey.* ACM Computing Surveys 53(6).
43. Tasioulas, J. 2018. *First Steps Towards an Ethics of Robots and Artificial Intelligence.* SSRN.
44. Boddington, P. 2017. *Towards a code of ethics for artificial intelligence*. Artificial Intelligence: foundations, theory, and algorithms. Cham, Switzerland: Springer. xix, 124 pages.
45. Turner, J. 2019. *Robot Rules. Regulating Artificial Intelligence*. London: Palgrave Macmillan.
46. Floridi, L., et al. 2018. *AI4People—An Ethical Framework for a Good AI Society: Opportunities, Risks, Principles, and Recommendations.* Minds & Machines.
47. AA.VV. 2018. *Should we fear artificial intelligence?*, in *In-depth Analysis*. European Union - STOA: Brussels.
48. Commission, E. 2019. *Building Trust in Human-Centric Artificial Intelligence*. Brussels: Eurpean Parlament.
49. Floridi, L. and J. Cowls. 2019. *A Unified Framework of Five Principles for AI in Society.* Harvard Data Science Review 1(1).
50. Mittelstadt, B. 2019. Principles alone cannot guarantee ethical AI. *Nat Mach Intell* 1: 501–507.
51. Rességuier, A., and R. Rodrigues. 2020. AI should not remain toothless! A call to bring back the teeth of ethics. *Big Data & Society* 7 (2): 1–5.
52. Stahl, B.C., et al. 2021. *Organizational responses to the ethical issues of artificial intelligence.* AI & Society.
53. Stix, C. 2021. *Actionable Principle for Artificial Intelligence Policy: Three Pathways.* Science and Engineering Ethics 27(15).
54. Morley, J., et al. 2020. *From What to how: An Initial Review of publicly Available AI Ethics Tools, Methods and Research to Translate Principles into Practices.* Science and Engineering Ethics 26: 2141–2168.
55. Stahl, B.C., et al. 2021. Artificial intelligence for human flourishing – Beyond principles for machine learning. *Journal of Business Research* 124: 374–388.
56. Ryan, M. 2020. In AI We Trust: Ethics, Artificial Intelligence, and Reliability. *Sci Eng Ethics* 26 (5): 2749–2767.
57. Salles, A., K. Evers, and M. Farisco. 2020. Anthropomorphism in AI. *AJOB Neurosci* 11 (2): 88–95.
58. Johnson, D.G., and M. Verdicchio. 2017. Reframing AI Discourse. *Minds & Machines* 27: 575–590.
59. McLennan S., et al. 2020. *An embedded ethics approach for AI development.* Nature Machine Intelligence 2: 488–490.
60. Forum, W.E. 2020. *Ethics by Design: An organizational approach to responsible use of technology*. Cologny/Geneva.
61. Stahl, B.C., et al. 2021. *From Responsible Research and Innovation to responsibility by design.* Journal of Responsible Innovation 1–24.
62. Tegmark, M. 2018. *Life 3.0 Being Human in the Age of Artificial Intelligence*. New York, NY: Alfred A. Knopf.
63. Havens, J.C. 2016. *Heartificial intelligence : embracing our humanity to maximize machines*. New York: Jeremy P. Tarcher/Penguin, an imprint of Penguin. xxxvi, 267 pages.
64. Roth, G. 2013. *The long evolution of brains and minds*. Dordrecht: Springer Science.
65. Farisco, M., K. Evers, and A. Salles. 2020. *Towards establishing criteria for the ethical analysis of AI.* Science and Engineering Ethics.
66. Gardner, H. 1985. *Frames of mind : the theory of multiple intelligences*. London: Heinemann. xii, 463 p.
67. Goleman, D., D. Goleman, and D. Goleman. 2004. *Emotional intelligence : why it can matter more than IQ ; Working with emotional intelligence*. London: Bloomsbury. xiv, 383 p.
68. Kirtay, M., et al. 2019. *Emotion as an emergent phenomenon of the neurocomputational energy regulation mechanism of a cognitive agent in a decision-making task.* Adaptive Behavior 0(0): 1059712319880649.
69. Bostrom, N. 2014. *Superintelligence : paths, dangers, strategies*. First edition. ed. xvi, 328 pages.
70. Dennett, D.C. 2019. *What can we do? We don't need artificial conscious agents. We need intelligent tools*, in *Possible Minds: Twenty-Five ways of Looking at AI*, J. Brockman, Editor. Imprint of Penguin Publishing Group: New York. p. 41–53.
71. Bentley, P.J., et al. 2018. *Should we fear artificial intelligence?, in In-depth Analysis*. Brussels: European Union - STOA.
72. Metzinger, T. 2021. *An Argument for a Global Moratorium onSynthetic Phenomenology.* Journal of Arti¯cial Intelligence and Consciousness 8(1): 1–24.
73. Dehaene, S., H. Lau, and S. Kouider. 2017. What is consciousness, and could machines have it? *Science* 358 (6362): 486–492.

74. Koch, C., and The feeling of life itself : why consciousness Is widespread but can't be computed. 2019. *Cambridge*. MA: MIT Press. pages cm.
75. Searle, J.R. 2007. Biological Naturalism. In *The Blackwell Companion to Consciousnss*, ed. M. Velmans and S. Schneider, 325–334. Malden MA, Oxford, Victoria: Blackwell Publishing Ltd.
76. Reber, A.S. 2019. *The First Minds : Caterpillars, 'Karyotes, and Consciousness*. New York: Oxford University Press. xxxii, 261 pages.
77. Lake, B.M., et al. 2017. *Building machines that learn and think like people.* Behav Brain Sci 40: e253.
78. Pennartz, C.M.A., M. Farisco, and K. Evers. 2019. Indicators and Criteria of Consciousness in Animals and Intelligent Machines: An Inside-Out Approach. *Front Syst Neurosci* 13: 25.
79. Levy, N. 2014. The Value of Consciousness. *J Conscious Stud* 21 (1–2): 127–138.
80. Farisco, M., and K. Evers. 2017. The ethical relevance of the unconscious. *Philos Ethics Humanit Med* 12 (1): 11.
81. Bruin, L.d., A. Newen, and S. Gallagher. 2018. *The Oxford Handbook of 4E Cognition*. Oxford handbooks. Oxford: Oxford University Press. xiii, 940 pages.
82. Changeux 1986: Changeux, J.-P. (1986). Neuronal man : the biology of mind. New York, Oxford University Press;
83. Evers, K. (2009). Neuroetique. Quand la matière s'éveille. Paris, Odile Jacob
84. Simulationist Code of Ethics (2015). https://scs.org/wpcontent/uploads/2015/12/Simulationist-Code-of-Ethics_English.pdf
85. Steels, L. and Lopez de Mantaras, R. (2018). The Barcelona Declaration for the Proper Development and Usage of Artificial Intelligence in Europe. AI Communications 31: 485 – 494.
86. HLEG (2019). Ethics Guidelines for Trustworthy AI. Brussels, European Commission
87. OECD. (2019). Recommendation of the Council on Artificial Intelligence. https://legalinstruments.oecd.org/en/instruments/OECD-LEG AL-0449.
88. IEEE Code of Ethics (2020). https://www.ieee.org/content/dam/ieeeorg/ieee/web/org/about/corporate/ieee-codeofethics.pdf

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.