AI and Ethics
https://doi.org/10.1007/s43681-022-00201-4

ORIGINAL RESEARCH

Assessing the ethical and social concerns of artificial intelligence in neuroinformatics research: an empirical test of the European Union Assessment List for Trustworthy AI (ALTAI)

Bernd Carsten Stahl[1,2] · Tonii Leach[1]

Received: 14 May 2022 / Accepted: 21 July 2022

Abstract

Ethical and social concerns are a key obstacle to the adoption of artificial intelligence (AI) in the life sciences and beyond. The discussion of these issues has intensified in recent years and led to a number of approaches, tools and initiatives. Key amongst them is the idea of ex-ante impact assessments that aim to identify issues at the early stages of development. One prominent example of such ex-ante impact assessment is the European Union's (EU) Assessment list for Trustworthy AI (ALTAI). This article uses the findings of a large-scale application of the ALTAI to a large neuro-informatics project as an exemplar to demonstrate the effectiveness and limitations of the ALTAI in practice. The article shows that ex-ante impact assessment has the potential to help identify and address ethical and social issues. However, they need to be understood as part of a broader socio-technical ecosystem of AI. For ALTAI and related approaches to be useful in bio-medical research, they should be interpreted from a systems theory perspective which allows for their integration into the rich set of tools, legislation and approaches. The paper argues that ex-ante impact assessments have the best chance of being successful if seen applied in conjunction with other approaches in the context of the overall AI ecosystem.

Keywords ALTAI · Artificial intelligence · Ethics · Impact assessment

1 Introduction

The discussion of the ethics of artificial intelligence (AI) has moved beyond the conceptual stage and is now at the point where practical measures have been developed and are being trialled and tested. Many different approaches have been proposed, ranging from a plethora of guidelines [66] to more specific suggestions, such as labelling [12], standardisation [63] and certification [62], all the way to far-reaching regulation and regulation [47]. It stands to reason that these different approaches will overlap and maybe even converge to a significant degree, for example by legislation making use and enforcing of standards, certification being based on standards or guidelines being encouraged by regulation.

This large and quickly developing field of activity now calls for ways of ascertaining whether proposed measures have the desired effect. In general, the discourse concerning future governance structures of AI is based on the broad consensus that the benefits of AI need to be balanced against its risks. It is by no means certain, however, how this can best be achieved. It is difficult to predict the consequences of AI use and development for a number of reasons. This includes conceptual questions such as what types of technologies constitute AI and therefore should be covered by such measures. There are significant epistemological questions concerning the measures to be used to identify issues and whether or to which degree these can be quantified and compared. The temporal horizon of any measure of AI impact is difficult to determine, leading to a possible over-emphasis on short-term consequences and the neglect of potentially much more sensitive long-term outcomes.

However, despite the fact that there are likely to be some specific challenges related to AI, the problem of predicting possible impacts of social or technical developments is not

✉ Bernd Carsten Stahl
bstahl@dmu.ac.uk

Tonii Leach
antonia.leach@dmu.ac.uk

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

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

Published online: 19 September 2022

new. Conceptual and epistemological issues have had to be addressed to identify environmental, social or other impacts arising from other technologies, such as biotechnology or nanotechnology. The AI field can therefore build on existing experience when considering suitable governance arrangements. One well-established technology and innovation governance mechanism goes by the name "impact assessment". There are a number of approaches used to identify a broad range of impacts, for example in environmental impact assessment [57], social impact assessment [17, 18] or ethics impact assessment [32]. The field of technology assessment [52, 53] can be described as a framework for developing impact assessments independent of particular technologies or expected fields of impact. In addition, there are a number of more specific types of impact assessment with high relevance to AI, such as privacy impact assessment [35, 36, 64], data protection impact assessment [65] or ICT ethics impact assessment [109].

There are now more than 40 proposals for impact assessments that are suitable or specifically targeted at AI. Probably the most prominent one amongst these is the Assessment List for Trustworthy AI (ALTAI) that was proposed by the EU's High Level Expert Group on AI [4]. The ALTAI's prominence derives from the high-profile manner in which it was developed under the auspices of the European Commission, which means it is likely to figure prominently in any future AI-related regulation [47]. It has furthermore already been incorporated into the ethics self-assessment list of the research framework programme Horizon Europe. This paper therefore presents an empirical study of the application of an AI impact assessment based on ALTAI to a large-scale project working on the intersection of neuroscience and technology development. This project, the Human Brain Project is currently focussed on developing a distributed ICT research infrastructure for neuroscience. It hosts a number of activities that make use of current AI techniques, but it also holds the potential to generate new insights into the links between neuroscience and technology that can inspire the next generation of AI. The paper thus tries to answer the research question: To what extent does the application of an assessment of AI trustworthiness (ALTAI) to research activities allow for the identification and mitigation of social, ethical and technical benefits or problems of AI? The answer to this question informs our suggestion to interpret ex-ante assessments from the perspective of AI ecosystems and the conceptual and practical implications such a shift in perspective may have.

The answer to our research question is important in several respects. It constitutes an important contribution to knowledge in the ethics and governance of AI debate which continues to be held primarily on conceptual grounds. To the best of our knowledge, this is the first study of an application of an AI impact assessment outside of the context of development of the assessment. It thus provides valuable empirical insights into the strengths and weaknesses of the approach. These insights are of high interest to the scholarly debate surrounding ethical and social aspects of AI and they are of similar relevance to the community of practitioners. In the light of the rapidly growing use of AI in bio-medical research and practice, an understanding of the practice of undertaking an impact assessment in this field is sorely needed. The development of AI governance regimes must be driven by sound conceptual foundations, but these need to withstand empirical tests. This paper contributes to the body of empirical evidence that is required to support and justify such regimes.

The paper is organised as follows: It starts with a brief overview of the current debate on ethics and trustworthiness of AI which covers the concerns that an impact assessment would be expected to address, highlighting the role of the concept of 'trustworthiness' in this context. This provides the background for the description of the methodology of our empirical study. The findings section presents the insights gained from the study, and this is used to inform the discussion of strengths and weaknesses of utilising the ALTAI as an impact assessment tool for AI systems. The conclusion outlines the limitations of our study and points to next steps that will be required to ensure the empirical viability of AI impact assessments and AI governance approaches more broadly.

2 Ethics and trustworthiness of AI

An understanding of the current discussion of ethics and AI is helpful to understand the shape and implementation of the ALTAI. Before coming to this discussion, it is important to delineate this discourse by briefly characterising the concept of AI.

2.1 The concept of AI

There are a range of different technologies that can fall under the heading of AI in the broadest sense. Here we identify just a few of these technologies, as this has implications for the range of technologies that would fall within the remit of an AI impact assessment (AIIA). The range is furthermore important because different types of AI raise different concerns that an impact assessment should cover. However, it should be noted that this is not intended as a comprehensive overview, so much as an acknowledgement of the diversity of systems likely to be subjected to an AIIA.

An important starting point is the observation that there is no universally accepted definition of AI. It has been described a "term that can mean a lot of things" [100] that has been "seized upon by the media, marketing departments and commentators as shorthand, and to add narrative spice" [108], p. 286). A typical approximation of AI is that it consists of machines doing the "kind of things that only people used to be able to do" [49]. This includes the ability to solve problem and achieve goals, notably by understanding and learning from data, imitating human cognitive functions, such as vision and speech, and emulating human thinking and feeling [81]. A key problem with this type of definition is that it is subject to changing perceptions concerning what constitutes human cognitive functions.

In practice, the current AI discourse has been triggered by the rapid progress of a particular approach to AI, namely machine learning [98]. The idea of machine learning is not new, but the approach has proven immensely successful in recent years which was facilitated by the availability of large data sets, the supply of experts having the right skills and the provision of powerful computing capacity [19, 55]. Whilst machine learning in its current iterations (e.g. deep learning, supervised, unsupervised or reinforcement learning) takes centre stage, it should be clear that it is an example of what is commonly called narrow AI that is trained to perform specific tasks and cannot apply its models beyond its training environment [89]. Machine learning is the most prominent but not the only type of narrow AI which also includes approaches such as expert systems. In addition to narrow AI, there is the concept of general AI or artificial general intelligence (AGI), sometimes also called strong AI, which stands for technologies that have truly human cognitive capabilities [90]. Whilst AGI currently does not exist, it has long been discussed as a technical vision, but also as a potential ethical and social concern, as it might lead to AI with unpredictable capabilities, sometimes called superintelligence [24].

2.2 Ethical concerns

There is a large and fast growing body of literature on the ethics of AI [37, 42, 68, 94] that this paper cannot replicate. For the purpose of understanding the role of AI impact assessments, it is sufficient to highlight some of the key ethical and social issues that AI raises. The Assessment List for Trustworthy AI (ALTAI) (AI [4] identifies requirements as follows that are linked to these key ethical and social issues: human agency and oversight, technical robustness and safety; privacy and data governance; transparency; diversity, non-discrimination and fairness; societal and environmental wellbeing; and accountability. A brief (and non-comprehensive outline of some of the issues linked to these requirements is provided below to contextualise the landscape within which the ALTAI has been developed and, in the case of the HBP survey, deployed.

A number of the most prominent concerns about AI are caused by some of the technical features of machine learning (currently a key approach used in the field of AI). These include the requirements for large amounts of data for training and validating models and the opaque nature of the models themselves that are not easily assessed and verified. As a consequence there are worries about the reliability of machine learning systems which may have implications for safety [42] and raise concerns about security [15, 22, 29]. Machine learning system as IT systems are subject to established cybersecurity threats, but their unique features may also leave them open to novel threats [105].

Where such systems make use of personal data, this raises questions about privacy and data protection [44, 67, 102]. Privacy is a value and a human right which is also required to prevent other problematic uses of AI, notably for the purposes of surveillance.

The opaque nature of machine learning has given rise to the discussion of its impact on fairness [40]. It is now well-established that machine learning systems can include bias, for example by replicating biases in training data [2, 31] which can then lead to discrimination [72], for example on the basis gender or race. The opaque nature of machine learning makes it difficult to assess the scope of biases and resulting discrimination and it also serves as an obstacle to establishing accountability for discrimination.

In addition to the effects that AI can have on individuals, their rights and their chances to live a life according to their own design, there are numerous possible consequences of AI use that are related to broader social consequences. Many of these are related to economic questions. AI is widely believed to have beneficial economic results, create efficiencies, promote growth, and create wealth (AI [3]. However, such benefits will be achieved by engendering changes which may be problematic. Key amongst these is the worry about the justice of the distribution of economic benefits which, if unchecked, is likely to see the large tech companies reap the benefits whilst leaving smaller companies, individuals and developing countries as collateral damage [79]. Zuboff has coined the term "surveillance capitalism" to highlight this concern [112]. AI is recognised as a key component of the "datafication" of the economy [106] which fundamentally changes economic structures. This may lead to unemployment [22, 98], a hypothesis that is contested [108], but it will likely lead to changes in work [82] which can facilitate increased worker surveillance and control.

The broader consequences of AI use are not confined to the economy but can be found in other aspects of social life. The concentration of economic gains can exacerbate the concentration of political power [80]. The Cambridge Analytica/Facebook scandal has shown that AI can be used to influence the outcome of democratic elections, which is another pathway for AI to damage democratic structures [110]. AI technologies can facilitate automated surveillance, thereby posing a threat to freedom of expression and having a "chilling effect" on free speech [1]. AI has potentially significant (positive as well as negative) environmental consequences [81]. It is likely to shape the future of warfare and military structures with potentially deep ethical implications [29, 54, 83]. Overall, AI can empower humans, but it can also structure their spaces of actions in clandestine ways, thus reducing their perceived options, reducing human control and freedom [38, 102].

These are examples of frequently voiced concerns about current AI, mostly related to recent progress in machine learning. This broad range of concerns is extended even further if a broader concept of AI is used and AGI is considered as well. Whether ethical concerns linked to technologies that currently do not exist—and that may never come into being—is highly contentious. However, AGI is arguably a key target of AI research [59, 77], it is widely covered in science fiction and the media. The development of AGI raises a number of additional concerns such as the possibility of superintelligence which may not only be superior to human intelligence but may be indifferent or hostile to humans. The rise of AGI might change our view of human nature, promoting human enhancement, post- and transhumanism [37, 103]. It raises the spectre of machine consciousness [41] and resulting questions of machines as holder of rights and subject of responsibilities [20, 46, 91].

One important conclusion to be drawn from this overview of the AI ethics debate is that it does not lend itself to clear-cut and straightforward interventions that allow dealing with ethical and social issues in a straightforward manner. Elsewhere it has been proposed that it might be more fruitful to regard AI ethics using a systems theory perspective [94]. More specifically, one can interpret AI as an interlocking set of ecosystems whose ethical and social issues arise from the interaction of AI technology with humans and other components of the ecosystem [93]. We will return to these ideas during the discussion of our findings and use them to support our recommendations.

Having provided a brief and non-comprehensive overview of some of the key ethical concerns related to AI, a final step before discussing impact assessments is to look at the role of the term "trustworthiness" in the debate of ethics and AI.

2.3 Trustworthiness of AI

Trustworthiness is a central term in the debate on ethics and AI. It forms part of the title of the ALTAI system that we based our empirical investigation on. It is therefore important to understand the relationship of trustworthiness and ethics in the AI discourse in some more detail.

Trust can be defined as the "willingness of a party to be vulnerable to the actions of another party based on the expectation that the other will perform a particular action important to the trustor, irrespective of the ability to monitor or control that other party" ([73], p. 712). This highlights the social nature of trust which involves at least two actors and the psychological aspect of this relationship that requires one, sometimes called the trustor, to accept vulnerability by the other, sometimes called the trustee [69, 70]. Trust can be viewed in terms of its social function of facilitating collaboration in social systems [71]. However, to fulfil this social function, it has to rely on the perceptions of trustor and trustee which are closely related to their ethical positions.

This link between ethics and collaboration may explain why the concept of trust has been taken up in business ethics. Trust is described as a condition of successful collaboration that arises from a minimal threshold of ethical behaviour such as promise keeping that is required in business interactions [43]. This means that establishing trust is a condition of competitive success [25].

These insights from the general field of business have been taken up by the tech industry. In traditional business interaction, trust could normally develop between interacting parties through face-to-face interactions and familiarity. Where interactions are technologically mediated this source of trust disappears and other forms of trust are required for interaction to be successful [27]. This insight has inspired much research on the meaning of trust and possible ways of promoting it in intermediated relationships, such as electronic commerce or work in virtual organisations [16, 21, 58, 99, 107].

The nature of trust as a component of social interaction that depends at least on the two parties of trustor and trustee means that it is impossible to simply impose trust [51]. This explains the shift of terminology from trust to trustworthiness. By being trustworthy an organisation can highlight that it fulfils the criteria required to be trusted. The same, by implication, could work for technology. Spiegelhalter [92], for example, suggests that trustworthiness for AI systems requires transparency and explainability and proposes criteria to determine whether these requirements are fulfilled.

This short introduction to trustworthiness will suffice to explain the prominence of the term in current AI debates. Speaking of "trustworthy AI" raises fewer objections than speaking of "ethical AI" as the latter concept would have to contend with questions whether AI has agency and can be ethical in itself. However, at the same time, the use of trustworthy AI suggests that ethical issues are considered and taken care of. An AI system can only be trustworthy if it does not cause ethical problems when used, which implies that it does not discriminate unfairly, violate data protection, create unfair distribution of resources etc. The term "trustworthy AI" thus includes ethical consideration but goes beyond these, for example by including technical reliability.

This use of the term is not without problems [85]. It is based on an impoverished concept of trust, suggests an instrumental use of ethics and abridges broader discourses about the purpose and desirability of technology. We will return to some of these issues below. For the moment, however, it is sufficient to realise that the use of the term "trustworthy AI" covers ethical questions which are core to the AI HLEG's ALTAI list which was used in the empirical work informing this paper. This ALTAI list is probably the most prominent attempt to formalise the assessment of a broad range of social, ethical and technical issues in an impact assessment process. Such impact assessments are a prominent mechanism for assessing and addressing such issues, but so far there are few empirical data on whether and to what degree they are successful in achieving this aim. This paper therefore offers an account of an initial application of the ALTAI approach outside of the tests that it was subject to during its development process. The methodology used for this application is described in the next section.

3 Methodology

The paper describes the insights arising from the application of an AI impact assessment closely based on the ALTAI checklist in a large neuro-informatics project. The methodology section therefore provides some more detail of the ALTAI process before giving an overview of the project it was applied to and the details of the data collection process.

3.1 The ALTAI process

The ALTAI self-assessment list [4] forms one of the key outputs of the European Commission's approach to addressing the ethics of AI. This approach relied on the formation of a high-level expert group consisting of 52 experts covering various social, ethical and technical aspects of AI. The group was first convened in 2018 [26]. It developed this definition of AI:

> “Artificial intelligence (AI) refers to systems designed by humans that, given a complex goal, act in the physical or digital world by perceiving their environment, interpreting the collected structured or unstructured data, reasoning on the knowledge derived from this data and deciding the best action(s) to take (according to pre-defined parameters) to achieve the given goal. AI systems can also be designed to learn to adapt their behaviour by analysing how the environment is affected by their previous actions.
> As a scientific discipline, AI includes several approaches and techniques, such as machine learning (of which deep learning and reinforcement learning are specific examples), machine reasoning (which includes planning, scheduling, knowledge representation and reasoning, search, and optimization), and robotics (which includes control, perception, sensors and actuators, as well as the integration of all other techniques into cyber-physical systems)” ([8], p. 7).

The development of this definition was followed by its first major output, the ethics guidelines for trustworthy AI [6]. These guidelines established that trustworthy AI needs to be lawful, ethical and robust. In practice, the AI HLEG focussed on ethical AI, assuming that lawfulness would be dealt with by legal experts, whereas robustness is a technical capability that is subject to technical or scientific assessment. The guidelines determined four ethical principles (respect for human autonomy, prevention of harm, fairness, explicability) which were translated into 7 key requirements (human agency and oversight, technical robustness and safety, privacy and data governance, transparency, diversity, non-discrimination and fairness, societal and environmental wellbeing, and accountability. Based on the principles described in the guidelines, the AI HLEG developed policy and investment recommendations [7], sectorial considerations [5] and the ALTAI assessment list.

The work of the AI HLEG has not been without criticism [75, 104] but it has without doubt been highly influential. It represents the most visible effort of any major state or region to come to an inclusive consensus of how societies can deal with the opportunities and challenges of AI. It provides key input into planned regulation of AI, notably the EU AI Act [47] in which it is referenced, it has developed ‘cornerstone’ principles for trustworthy AI in the EU [97], and, furthermore, all EU projects must undertake an ethics review component on the basis of the work of the HLEG, constituting impact across a large research landscape.

The ALTAI assessment list plays a key role in the translation of the ideas produced by the AI HLEG into practice. It aims to provide assistance to AI developers to determine whether and to what degree the 7 requirements for trustworthy AI are realised. For this purpose, the 7 requirements are broken down into sub-sections (e.g. transparency is broken down into traceability, explainability and communication), each of which is assessed using a number of questions.

Whilst the ALTAI list is not the only example of an AI self-assessment or impact assessment [9, 61, 111] its visibility benefits from the central role it plays in EU AI policy. In addition, it has gained further force by virtue of the fact that it was incorporated into the ethics self-assessment of the Horizon Europe Research Framework Programme [48]. This document expresses the expectations of the European Commission when funding research and reflects the principles and categories of ethics review and assessment. In practice, all EU project proposals are evaluated against these criteria. AI has been integrated as an ethical issue in the ethics self-assessment. Applicants whose work includes the use or development of AI are "strongly encouraged" to "use the Assessment List for Trustworthy Artificial Intelligence (ALTAI) to develop procedures to detect, assess the level and address potential risks" ([48], p. 39). This use of the ALTAI list is the initial reason why we engaged with it, as we were asked to comment on the way ethics of AI is dealt with during an ethics review of the project we work for which we describe in the next section.

3.2 Ethics of AI in the human brain project

The Human Brain Project (HBP) is research project funded under "Future and Emerging Technologies Flagship" funding scheme that formed part of the European Commission's Horizon 2020 research framework programme. As a Flagship project, the HBP has a duration of 10 years (2013–2023), an EU budget of more than €400 m, comprises more than 100 partners with more than 500 individuals working on it. It brings together research from neuroscience and ICT/computer science with the aim of developing a distributed ICT research infrastructure for neuroscience [13]. Due to the fact that it includes many research and technology development activities across many disciplines, it has always been clear that the project raises significant ethical and social challenges [33, 84]. To ensure that these are addressed appropriately, the project from the outset included a programme of work on ethical and social issues [11], much of it famed in terms of responsible research and innovation [87, 88].

AI, with its fundamental link to the concept of intelligence (and the associated issues of value-leadenness and political ideology) [30], can be considered to fall within the remit of a number of associated disciplines, including (but not limited to) neuroscience, biology, and psychology [8]. It therefore stands to reason that a project like the HBP that brings together cutting-edge research in neuroscience and computer science would make use of current AI and at least have the potential to contribute to the next generation of AI technology, including at the level of fundamental or foundational research related to AI. It therefore did not come as a surprise that during an ethics review of the project in January 2021 the project was asked to assess the ethical aspects of its use and development of AI. However, the project itself is funded under the Horizon 2020 Research Framework Programme (the predecessor of Horizon Europe) which means that the ethics regime that governs the HBP had not previously covered AI. In addition, the new ethics guidance on AI only came into force in July 2021, so that at the time of the ethics review the way in which the ethics of AI is addressed in European projects was not yet fully known.

The result of the ethics review was a new requirement for the HBP to explain how AI used or developed in the project meets the criteria for trustworthiness. The report explaining this was to cover measures set in place to avoid potential bias, discrimination and stigmatisation; measures to ensure safety and prevention of harms (to humans, animals, environment); an explanation of how the respect of fundamental human rights and freedoms (e.g. human autonomy, privacy and data protection) will be ensured; measures to ensure fairness and explicability (paying particular attention to situations involving more vulnerable groups). In particular, the project was asked to explain how humans will maintain meaningful control over the most important aspects of decision-making process. This review of AI ethics was required to include an evaluation of the ethics risks related to the development/deployment of the AI systems/techniques and explain how the potential negative social impacts would be mitigated. Guidance on how to provide this information included references to the AI HLEG work, in particular the ALTAI guidelines. The responsibility for the creation of this report that was to be submitted as a deliverable to the European Commission was allocated to the authors of this paper due to their work as part of the Ethics and Society group of the HBP and ongoing work with researchers working on topics closely aligned to AI.

3.3 Survey design and delivery

A key challenge with applying an instrument such as the ALTAI self-assessment to the HBP is the size and complexity of the project. The HBP is divided into nine work packages comprising 114 tasks. It has a complex governance structure that is not conducive to central interventions. In particular with regard to AI, the Ethics and Society group had previously worked with scientists and technologists in the project to develop an Opinion on AI [10]. During this work, it became apparent that there was little consensus on which aspects of the HBP work should fall under the heading of AI and where in the project this work was done.

To overcome this fractured nature of AI work in the HBP, it was decided that a comprehensive overview was only to be achieved, if insights could be collected from all parts of the project. The suitable data collection tool was therefore deemed to be an online survey that would need to cover all parts of the project. Given the scope and breadth of AI relevant work undertaken in the HBP, it was deemed most appropriate to seek responses about each AI system being used/developed in the HBP from the researchers, scientists and technologists working on or with that system, as, regardless of discipline, these were the people most likely to be able to provide a comprehensive assessment of potential ethical issues in relation to the system. Recognising that the voluntary participation of the project's scientists and technologists was going to be of paramount importance for collecting the required information, the design of the online survey was undertaken in a collaborative manner. In practice, this meant that a number of researchers were consulted prior to the creation of the survey to assess possible issues and problems. An online, mixed methods survey was developed based on the ALTAI questions and implemented in the MS Office 365 Forms tool, and comprised a series of close-ended questions used as a screening tool to allow targeted questioning on ethical issues and concomitant mitigation factors that required an extended, text-based answer from only participants for whom the question was relevant. This survey was then shared with the AI experts consulted initially as well as the HBP's Ethics Advisory Board and members of the Ethics and Society team. After several iterations the survey was finalised. As a result of this process the online survey we used diverges somewhat from the ALTAI. It includes some questions and topics that ALTAI does not cover, for example the use of subliminal techniques, which were highlighted as a key concern of the EU AI Act [47] that rose to prominence after ALTAI was published. The full wording of the survey instrument is available in Appendix 1.

The distribution of the survey was endorsed by the HBP's Directorate, the main administrative and management governance body where it was determined that it would be distributed to all tasks of the project via the existing work package structure, relying on the work package managers to distribute the survey and encourage researchers to fill it in. The aim was to ensure that all tasks would have sight of the survey to ensure that tasks for which it was relevant would fill it in.

The survey was initially sent out to the WP managers on 15.10.2021. From 26.10.2021, missing tasks were chased, typically with the help of the relevant WP managers. The survey closed on 12.11.2021. The qualitative survey data were analysed in relation to the specifics of the requirement only, and the resulting deliverable was submitted to the EC on 31.12.2021. The data collection was originally undertaken as part of the contractual obligations of the HBP and was considered a reporting tool rather than research. However, the data collected were deemed to be of sufficient interest to warrant publication. We therefore applied for ethics permission from De Montfort University to collect consent from respondents for the use of the data as research data. This permission was granted on 21.03.2022. All respondents were then asked to consent to this use of the data. Only data for which relevant consent exists were included in the analysis presented in this paper.

The survey was responded to by 128 researchers from across the Human Brain Project. A further respondent answered by email, taking the total respondents to 129. Responses were received for 97 Tasks within the HBP, across all 9 Work Packages. There were 17 Tasks with no response. Two Tasks which had not begun at the time of survey distribution were disregarded for the purposes of the survey. This paper presents the analysis of 115 of these survey responses.

4 Findings

The findings of the survey presented here focus on insights that are of relevance to answering our research question, i.e. that give an answer to the question of the extent to which the application of ALTAI to research activities allows for the identification and mitigation of social, ethical, and technical benefits or problems of AI. We are thus not so much interested in the factual answers that HBP researchers give concerning their work, which is a specific project-related issue that is subject to ongoing ethics review. We are more interested in what the answers reveal about the nature, applicability, and usability of the self-assessment approach to trustworthy AI which is likely to be of interest to a broader set of stakeholders in the AI ethics community. The context of AI use in neuroscience should render the findings to be particularly pertinent to stakeholders in the bio-medical field.

4.1 Significant use of AI in the human brain project

The HBP is not an AI project per se. The first step in assessing the trustworthiness of AI was thus to ensure that a shared understanding of the terminology existed. Therefore, respondents were given the following definition of Artificial Intelligence:

> “Artificial intelligence (AI) systems are software (and possibly also hardware) systems designed by humans that, given a complex goal, act in the physical or digital dimension by perceiving their environment through data acquisition, interpreting the collected structured or unstructured data, reasoning on the knowledge, or processing the information, derived from this data and deciding the best action(s) to take to achieve the given goal. AI systems can either use symbolic rules or learn a numeric model, and they can also adapt their behaviour by analysing how the environment is affected by their previous actions” ([6], p. 24).

Respondents were asked whether, based on this definition, they developed or used AI within the HBP. Of the 115 responses, 75 stated that they do not work with AI. 40 respondents stated that they do work with AI. Respondents were asked to provide an overview of their work within the HBP. Figure 1 provides some insights into the responses to this question. Frequently cited areas of work included:

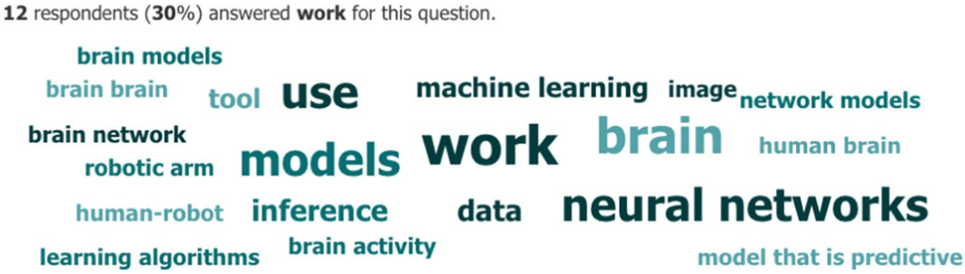

Fig. 1 Word cloud of key terms in responses about work undertaken in the HBP

- Machine learning
- Robotics software
- Neuromorphic computing platforms
- Neural networks

Respondents that stated they do not work with AI were screened out of the questionnaire at this stage.

Respondents were then asked about the significant use or development of AI. This form of words is based on the ethics requirement that triggered our investigation. By asking only about "significant" use or development, the ethics review request reflected the EU's risk-based approach to AI which is meant to reduce bureaucracy and ensure that ethics considerations do not needlessly hamper innovation. We followed the Hleg [4, 5] by defining a significant system as one that is designed to interact with humans in a way that may affect humans individually or society as a whole. 14 respondents indicated that they used or developed a significant AI system. The 26 respondents who indicated that they did not use or develop a significant AI system were screened out of the survey at this stage.

The 14 remaining respondents were asked to identify the type of AI system they used or developed. A combination of self-learning/autonomous, human-in-the-loop, human-on-the-loop and human-in-command systems were reported—2/6/3/3 responses respectively (see Fig. 2).

4.2 Key concerns and issues

Having established the significance and type of AI use in our sample, the following questions referred to frequently discussed ethical, social or legal concerns that could or did arise in the context of our respondents' work with AI. Given the mixed methods employed in this survey, for questions where an extended, open-ended text-based answer was required, we have adopted the standard qualitative research practice of providing illustrative examples of the responses to the questions in this section [14, 56, 74, 76]. This is particularly helpful in demonstrating the different approaches

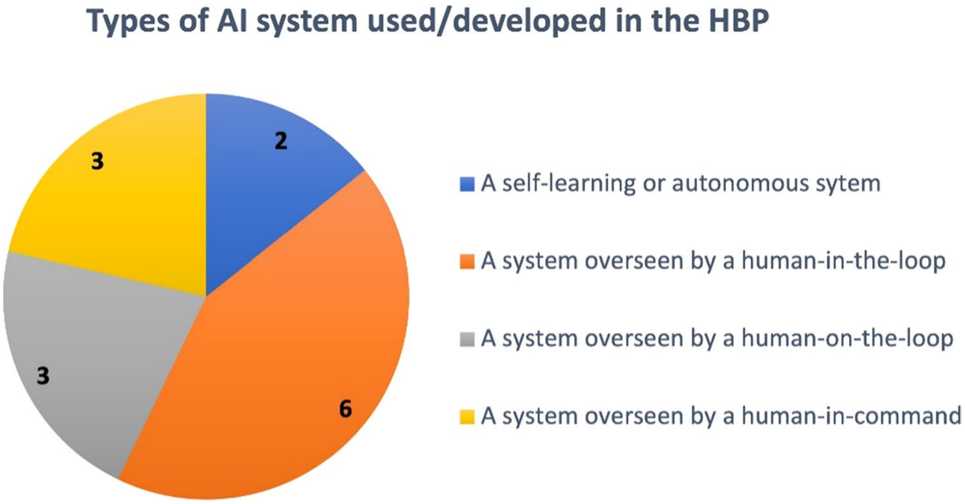

Fig. 2 Types of AI systems used or developed in the HBP: a self-learning or autonomous system; a system overseen by a Human-in-the-loop, which refers to the capability for human intervention in every decision cycle of the system; a system overseen by a Human-on-the-loop, which refers to the capability for human intervention during the design cycle of the system and monitoring the system's operation; or a system overseen by a Human-in-Command, which refers to the capability for a human to oversee the overall activity of the AI system (including its broader economic, societal, legal and ethical impact) and the ability to decide when and how to use the AI system in any particular situation

to, and responses elicited by, the ALTAI-based survey questions, allowing for a greater understanding of the ways in which the ALTAI tool is, or could be, (or, conversely, is not, or could not be) applicable in the HBP context. These responses are formatted in italics and indented paragraphs for easy identification.

First, the respondents were asked if the AI system could affect human agency or autonomy. They were provided with the following explanation of Human Agency and Oversight:

> the effect AI systems can have on human behaviour in the broadest sense. It deals with the effect of AI systems that are aimed at guiding, influencing or supporting humans in decision-making processes, for example, algorithmic decision support systems, risk analysis/prediction systems (recommender systems, predictive policing, financial risk analysis, etc.). It also deals with the effect on human perception and expectation when confronted with AI systems that 'act' like humans. Finally, it deals with the effect of AI systems on human affection, trust and (in)dependence [6].

Four respondents indicated that the AI system could affect human agency or autonomy, ten that it could not. The respondents who indicated that the AI system could affect human agency or autonomy were asked to explain measures put in place to ensure that human agency and autonomy, and human oversight, had been ensured. Responses included:

> ‘This very much specific to the particular application—as a platform provider we support users, and so far very few users are in the relevant territory, though the potential is there.’
> ‘A service level agreement is agreed when obtaining an EBRAINS account, and this is currently the only way for researchers outside [location] to access the [AI system].’
> ‘The clinicians will overview the results and make the final decision on how to use/interpret them.’

Each of these responses adopts a slightly different approach to mitigating risks to human agency or autonomy, which can be characterised as follows: first, that the system or platform is too far away/not application-specific enough for these issues to be usefully considered; and secondly, a user-centric approach which relies on the user to make an informed choice on using the system in the light of the potential for these issues, as explained in either the user agreement or on a decision-by-decision basis. In both cases, the active role in mitigating these issues appears to be deferred by the respondents.

The respondents were the asked if the AI system could have adversarial, critical or damaging effects (e.g. to human or societal safety) in case of risks or threats, such as design or technical faults, defects, outages, attacks, misuse, inappropriate or malicious use. Eight respondents indicated that the AI system could have these effects, six respondents indicated that it could not. Respondents who indicated that there was the possibility of adversarial, critical, or damaging effects to human or societal safety were asked to outline the measures put in place to ensure human or societal safety. Responses include:

> ‘Outputs of AI-based MIP analyses are controlled [sic] in various ways. The findings are weighted against other non-AI-based analyses, and eventually peer-reviewed and available for reproducibility. Risks and threats are in fact not different than those related to any type of inaccurate data analysis in medical research regardless of the use of AI.’
> ‘In the case of brain modelling software applied in a clinical setting, the diagnosis provided by the AI system is provided a simply another piece of information by the expert clinical team, which they can consider or not, when choosing the appropriate intervention for a patient.’

It is notable that responses to this question highlighted a consideration of ethical issues associated with human and societal safety that simultaneously acknowledges the risks posed by these systems, and reduces it as being no more of a risk than any other system, being not different ‘in fact’ or ‘simply another piece of information’, suggesting that the risks posed by AI to human and societal safety are considered, to an extent, as not AI specific enough to considered separately to other systems posing similar risks.

Respondents were asked if a low level of accuracy in the AI system could result in negative consequences, where accuracy is the predictive capability of the model and the ability to generalise well for unseen data. Eight respondents indicated that the AI system could have negative consequences given a low level of accuracy; six respondents indicated that there could not be negative consequences. Where respondents indicated that a low level of accuracy might have critical, adversarial, or damaging consequences, they were asked to outline the measures put in place to ensure a high level of accuracy. Responses included:

> ‘The model results are vetted by the modelling team in comparison to known good results.’
> ‘We regularly test changes to our code base to ensure accuracy.’

These responses demonstrate a reliance on rigorous testing to mitigate for low accuracy in an AI system, although a consideration of what an appropriate level of accuracy is not included here.

Respondents were asked if low reliability or reproducibility in the AI system could result in negative consequences. 7 respondents indicated that the AI system could have negative consequences given a low level of accuracy; the remaining 7 respondents indicated that there could not be negative consequences. Where respondents indicated that low reliability or reproducibility might have critical, adversarial, or damaging consequences, they were asked to outline the measures put in place to mitigate risks to human safety in the event of low reliability. Responses included:

> 'The risk would be out of bounds movement of the eventually involved robot arm. Safeguards will be put in place that will allow shut-down to prevent injury.'

This response demonstrates that, in the case of this particular system, although it is not at application stage yet, a consideration of future use cases has been made, and assumptions about mitigation processes outlined.

Respondents were asked about the possibility of malicious, inappropriate use or misuse of the AI system. Six respondents indicated that there was a possibility of this kind of use of the AI system; they were asked to outline the measures put in place to mitigate these risks. Responses included:

> 'Literally anything can be used for malicious purposes. Our work aims specifically at beneficial applications, but it is public-domain (as science should be), so it is impossible to exclude misuse.'
> 'As outlined in the previous answer, usage of the largest machines is limited to (a) those at the [location], or (b) those who have signed agreements / EBRAINS accounts.'

A number of different approaches to mitigating misuse came up in these responses. The first category considered the difficulties of predicting misuse—responses either argued that anything can be misused which makes prediction virtually impossible, or that the research being undertaken is too fundamental in nature to know what applications might emerge and therefore what scope for misuse might be possible. One mitigation strategy discussed was to limit who can access or use a system—however, this does not, in and of itself, prevent misuse so much as make accountability and traceability of misuse cases easier.

Respondents were asked whether the AI system could impact on a person's right to: privacy; physical, mental, or moral integrity; or data protection. Two respondents indicated that there was a possibility of this kind of impact; twelve respondents denied this; one respondent answered that the risks and mitigating measures are 'very much application specific'. The other respondent stated that their work is focussed on producing beneficial applications but that the nature of public domain science means that negative impacts on these specific rights cannot be excluded. Where personal data were used to train AI systems, respondents indicated that they complied with data protection principles, notably the GDPR [50].

Explainability is a key AI-related concern, and refers to 'the ability to explain both the technical processes of the AI system and the reasoning behind the decisions or predictions that the AI system makes' ([6], p.14). Respondents were therefore asked whether the AI system was designed to make decisions about human end-users. Four respondents indicated that the AI system was designed to do this; ten respondents indicated that the system was not designed to do this. The respondents that indicated that the AI system was designed to make decisions about human end-users were asked whether the AI system was likely to act as a 'black box'. Respondents were given the following definition of 'black box': 'AI systems whereby 'An explanation as to why a model has generated a particular output or decision (and what combination of input factors contributed to that) is not always possible' ([6], p. 14). Of the four respondents, 3 indicated that the AI system was not likely to act as a 'black box', one respondent indicated that the system was likely to act in this manner. The respondent who indicated the AI system could act as a 'black box' was asked to outline measures put in place to ensure the explainability of the AI system. The respondent answered that the ability to ensure explainability of the system would be 'application-specific'. Respondents furthermore pointed out that in none of the intended AI applications was there a danger that end-users might be misled about whether they interact with an AI.

Explainability is typically seen as important because it promises to avoid problems of bias and discrimination. We therefore asked respondents whether the AI system was being trained on data that may have issues of bias, discrimination, or that may suffer from a lack of diversity. Four respondents indicated that the AI system was being trained on data that may have these issues; ten respondents indicated that the data did not have these issues. The respondents that indicated that the AI system was being trained on data that may be lacking in diversity, or that may have issues of bias or discrimination were asked to outline measures in place to ensure that bias, discrimination, or a lack of diversity in the data was mitigated. Responses included the following:

> 'The biases are related to the geographical regions of origin of the datasets. Such biases are not raising issues of diversity or discrimination'
> 'The systems currently under development target clinical cohorts such as epilepsy, where it is not necessarily feasible to achieve fully diverse data sets.'

Responses to this question appeared to reflect ongoing debates about bias, diversity and discrimination in AI datasets, within the specific context of the HBP as a Euro-centric research project, as well as the disproportionate incidence rate of certain diseases (including epilepsy) across particular population demographics [60].

Respondents were asked whether the AI system might cause discriminatory or biassed outcomes. The three respondents indicated that the AI system might have this effect were asked to outline measures in place to mitigate this risk. Responses included the following:

> 'Fundamental problem of AI'
> 'As before: in principle, anything is possible. In our specific work, it is unlikely for users to be affected by the output of the AI in an ethically relevant way.'

These responses tend to focus on the limitations of this particular line of questioning in relation to the current research being undertaken: one respondent refers simply to the question of whether discrimination by an AI system in any given situation acts as a feature or a bug as a fundamental issue in this field, and another argues that a biassed or discriminatory outcome is entirely possible (although not, per their response, probable or likely), and that the scope of the system would inherently limit the impact of any potential outcomes in this direction.

Subsequently respondents were asked whether the AI system has been designed in an accessible way, for use by people with a range of abilities. Seven respondents indicated that the AI system had been designed to be accessible. The three respondents that indicated that the AI system had not been designed to be accessible to people with a range of abilities were asked to outline measures in place to ensure that the system is not inaccessible or likely to disadvantage specific groups of users through its design. Responses included the following:

> 'The systems currently under design or in use target as end-users, clinicians and provides them with standard document formats, such as PDF and HTML, yet the report results could be made available in alternative forms which would be more accessible.'
> 'The system is too far from an end-product to worry about these aspects/they can be addressed closer to market.'

These responses highlight that this question relies on the fundamental assumption that the system is at a completed stage of design to give an accurate response—where a system has yet to be completed/developed to application stage, it is, per se, impossible to say that it has been designed as fully accessible, therefore requiring respondents working in the pre-completion/pre-application stage to answer 'no' to this question, and therefore try to provide a justification for simply an incomplete system.

Respondents were asked whether the AI system is capable of using 'subliminal techniques'. Respondents were given the following definition of 'subliminal techniques': 'Subliminal techniques are practices that have 'a significant potential to manipulate persons through subliminal techniques beyond their consciousness or exploit vulnerabilities of specific vulnerable groups, such as children or persons with disabilities, to materially distort their behaviour in a manner that is likely to cause them or another person psychological or physical harm' ([47], sec. 5.2.2) Two respondents indicated that the AI system was capable of deploying subliminal techniques. They indicated these techniques would be application-specific, and that current conceivable applications would not affect users in an 'ethically relevant' way.

In answering a question about potential negative impacts on the environment, three respondents indicated that the AI system could have negative impacts on the environment, whereas eleven respondents indicated that the AI system could not have this impact. The respondents who indicated that the AI system could have a negative impact on the environment were asked to outline the potential negative impacts and describe the measures in place to mitigate these. Responses alluded to the fact that environmental impacts are likely to be application-specific, and also that these applications may be more or less beneficial or harmful to the environment depending on the direction of development.

A further concern about AI is its potential to impact human work and work arrangements. Six respondents indicated that the AI system could have this impact. When explaining how these issues are mitigated, responses included the following:

> 'System designed to support human worker in tasks that are dull (repetitive). Further, the system may be developed to alleviate mechanical load on human workers' musculoskeletal system (i.e. heavy payload carrying/manipulation). It is developed to improve work conditions of human workers, alleviate long-term musculoskeletal issues (and the associated socioeconomic burden).'
> 'heavy or dull tasks could be simplified'
> 'In the future co-working scenarios as we are trying to demonstrate may become deployable in manufacturing settings. This might affect human work patterns, but we would expect in a positive way.'

These answers again focus heavily on potential applications, and, in this particular case, with a bent towards robotic-integrated applications. It is also worth noting that this is the only question where responses routinely highlight positive outcomes in relation to an ethical issue.

Respondents were asked whether the AI system could have a negative impact on society at large or democracy. Three respondents indicated that the AI system could have such negative impacts; eleven respondents indicated that the AI system could not have negative impacts in this area. The respondents that indicated that the AI system could have negative impacts on democracy and society at large were asked to outline the potential negative impacts and describe the measures in place to mitigate these. Responses suggested that such wide-reaching impacts require a greater understanding of the applications likely to develop from this research than is currently available, but do not suggest that such negative impacts are a possibility.

Respondents were then asked whether the development process of the AI system, the sourcing of training data, and the processes, outcomes, and possible impacts of the AI system had been recorded. 13 respondents indicated that these records had been made.

The final question asked respondents whether there are ethical mechanisms in place to support the overall accountability and ethics practices in relation to the development of the AI system. Ten respondents indicated that these mechanisms were in place. The four respondents who indicated that ethical mechanisms were not in place were asked to outline how AI ethics had been integrated into the development of the AI system. Responses included the following:

> 'GDPR'
> 'Again, this is very much application-specific. These issues have yet to arise.'
> 'not at all. We are training a feed-forward neuronal network. No ethical concerns'
> 'We believe that the system based on synthetic data and at the current TRL is low ethical risk.'

Given the outward facing scope of the survey, it was to be expected that some respondents focussed here on external elements of integrating ethical mechanisms, rather than considering existing processes in respect of this within the HBP. In addition, it is clear that some respondents felt that many of the questions relating to ethical issues in this survey were not of particular relevance to the system they use/develop in the HBP at this time. As such, and given these factors, responses along the lines of 'no ethical concerns' are not unexpected. This acknowledgement that the respondents' views are not surprising does not imply that they represent the authors' position. The findings section was predominantly descriptive. The final section contains a critical reflection of our findings and the resulting recommendations.

5 Discussion and conclusion

This paper presents the findings from a real-life test of the ALTAI self-assessment. We engaged with a large number of scholars who can be described as working in the field of neuroinformatics to determine the extent to which the application of an assessment of AI trustworthiness (ALTAI) to research activities allow for the identification and mitigation of social, ethical and technical benefits or problems of AI. Before providing the answer to this research question, it is worth highlighting the limitations of our approach.

5.1 Limitations

We did not fully follow the ALTAI guidelines in our work. ALTAI states that the starting point for an AI impact assessment should be a fundamental rights impact assessment (FRIA), which we did not explicitly undertake. We would argue that our long-standing work on ethics and society in the HBP has highlighted the issues that a FRIA should, but we did not undertake a separate FRIA. Similarly, we did not link the HBP Data Protection Impact Assessments (DPIA) to the ALTAI. The HBP has done DPIAs but these were again not explicitly linked. In terms of delivery of the assessment, we chose the method of asking individuals via an online survey. ALTAI suggests undertaking the assessment in multidisciplinary teams which was not possible using this approach. We slightly altered the ALTAI questions in discussion with our respondents and other stakeholders which will hopefully have increased relevance and scope of the survey but will make it difficult to compare our responses with others using the original ALTAI wording. Our target audience furthermore consisted of scholars with a range of backgrounds across neuroscience and computer science, many of whom would not consider themselves to be AI researchers. And, finally, despite the large size of the HBP as a project, we identified only 14 individuals who considered themselves to be undertaking significant AI research. This small number, combined with the specific nature of the HBP means that we cannot claim any statistically significant insights that apply to AI research overall.

We nevertheless believe that the findings presented in this paper are important and allow us to draw conclusions of broader interest. The inclusion of ALTAI into the Horizon Europe ethics self-assessment guidance [48] means that requests for the application of ALTAI are likely to become the norm for any Horizon Europe projects involving AI. Many of these projects will be similar to the HBP in that they will aim at particular application areas from health to gaming and use AI as a means to achieve their scientific objectives. This means that the composition of ALTAI users will be similar to our respondents with various disciplinary backgrounds and variable expertise in AI. We furthermore believe that the application of ALTAI in such cases where it is required to address the ethics requirements of a project will be similar to our approach. The nature of ALTAI as a list of questions will tempt many projects to take a similar approach to the one we undertook, i.e. to ask the individual experts involved in AI development to go through these questions. This means that our approach can be expected to be typical and therefore our insights are likely to be relevant more broadly.

5.2 The value of assessing trustworthiness in AI

The responses we received to our questionnaire demonstrate that those scholars who self-identify as doing significant AI work show an awareness of some of the broader key areas of ethical concern focussed on in the ALTAI, and are aware of some strategies for seeking to mitigate such issues within their work—in particular, for example, testing processes to ensure the accuracy of a system's outputs were clearly described. This is not particularly surprising, given the high level of public attention to AI and its ethical consequences that is reflected in the media as well as research discourses. However, many responses also demonstrated a more limited grasp of how those ethical issues linked explicitly to the AI relevant research currently being undertaken in the HBP, and on occasion failed to identify appropriate mitigation mechanisms—for example, by regarding certain ethical concerns as inherently irrelevant to current research based on the stage of development, or by identifying the existence of biases in datasets used for research whilst simultaneously negating any consideration of the potential consequences of these biases.

Our research indicates, however, that the ALTAI approach as implemented here has fundamental limitations that one needs to be aware of. Some of these are of a conceptual nature. This starts with the inclusion of AI as an ethical issue into the Horizon Europe ethics self-assessment list. The inclusion of AI in the ethics self-assessment is understandable, given the high level of public debate which led to the formation of the HLEG and its ethics guidelines [6] and the ALTAI [4]. This inclusion of AI into the ethics self-assessment list is nevertheless highly contentious and arguably based on a fallacy of category. It should be obvious that AI is not an ethical issue but a technology, technique or family of techniques. The use or development of AI may raise all sorts of ethical issues as confirmed in this document, but it does not constitute one in and of itself.

This may appear like a peripheral observation, but it has manifest consequences for the way such issues can be dealt with. It has long been established that AI is a problematic concept and that no comprehensive definition exists [45]. Current proposals for definitions such as the one by the HLEG [8] that was used here or the one developed in the EU AI Act [47] tend to be too broad or too narrow. The responses to our survey supported the expectation that this could be problematic and showed that researchers were unsure whether their work is covered by the various definitions. This would not matter, if the issue were a consequence of technology use, e.g. bias and discrimination, but it does matter when the ethical issue is posited to be AI. By focussing on the concept of AI, the ALTAI approach therefore creates questions of applicability that are independent of the actual ethical and social consequences of the use of the technical system in question. This is a limitation that applies to the entire ethics of AI discourse which, by adopting the popular term AI, has to contend with the problem of defining it and thereby including or excluding specific technologies from the remit of ethical reflection. The responses we received to our survey demonstrate the validity of this concern which puts individual researchers into the difficult position of having to decide whether their work requires ethical scrutiny.

A further conceptual issue is the use of the term 'trustworthiness'. As indicated earlier, in practice this term is used to cover various ethical and social concerns. Ethical acceptability is arguably a contributing factor that allows building trust in technology. However, the focus on trustworthiness does not fully represent the range of ethical and social concerns. The reason for the prominence of the term may have something to do with the social dynamics of the AI HLEG. In practical terms during the application of the ALTAI as described in this article, the focus on trustworthiness was of limited importance. One could nevertheless argue that by focussing on trustworthiness one may lose from sight relevant concerns. In addition, trustworthiness may suggest a certain level of objectivity which can hide the need for ethical discourses.

Another limitation that was highlighted by the survey is related to the application of an ex-ante instrument such as ALTAI to research activities. The actual uses of a technology and their consequences are impossible to accurately predict at the research stage. This is an old problem of technology ethics and technology assessment [39]. However, the use of a tool such as ALTAI can be read as falsely suggesting that consequences can be accurately predicted. In the responses to our survey, we could see two possible consequences of this position. On the one hand, the attempt to predict consequences can lead to over-sensitivity and the attempt to consider all possible consequences. These are by definition infinite, which renders an ex-ante assessment infeasible. The other extreme is a desensitising to consequences, where all consequences are ignored due to their uncertainty. In practice, it is likely to be possible to predict some consequences with an accuracy that is sufficiently high to warrant actions based on the prediction. It is exceedingly difficult, however, to determine what that level of accuracy is and how it can be determined. The current version of the ALTAI survey provides little guidance on how such judgement calls can be made and justified. It is likely that a repository of experience will build up over time that will allow determining the required level of likelihood of possible consequences that need to be considered. At this point, there seems to be no mechanism, however, to collect experience of the application of ALTAI and build up a body of knowledge and good practice examples.

Such a body of knowledge would also be helpful for the steps following the ex-ante assessment. The bulk of ALTAI and related assessments focusses on the identification of possible consequences. They provide much less input into the way in which these issues can be addressed. The logic behind this approach seems to be that a reflection on various possible issues and the discussion of the questions contained in the assessment document will lead to the discovery of adequate solutions. This is likely to be the case frequently, but there is no guarantee that it will always be the case. In addition, the way in which the issues are addressed are rarely clear-cut and unambiguous. If, for example, a researcher is aware of the need for explainability and has included measures that are meant to support this feature, the question remains open whether the level of explainability is sufficient for the application and for the intended audiences. The assessment list itself provides no guidance on how such questions can be answered.

A final remark based on the uncertainty of consequences is in order concerning the term "significant" in the requirement. The European Commission is trying to reduce the administrative burden of ethics reviews by focussing only on significant ethical concerns. This is mirrored in the focus on high-risk AI in the EU AI Act. Whilst such a focus on the important cases is welcome, making a distinction of which technology or application will have significant consequences is impossible at the research stage. This adds to the overall problem of the uncertainty of future evaluations raised in the previous paragraph. Again, it is mostly left to the researcher to determine whether their work fulfils the definition of significant which can again lead to an over-sensitivity or an under-sensitivity with no obvious way of correcting either.

Our work thus shows that applying the ALTAI leads to conceptual and practical problems. The conclusion to be drawn from this short overview of the limitations of the approach is not to discard it altogether but to ensure that it is understood in the broader context of technology research, development and use. One way of achieving this is to see AI as conducive to human flourishing as a way to represent ethical and social concerns. Using such a lens, it helps to understand that AI is not a clear-cut and easily identifiable technology, but can better be conceptualised as a socio-technical innovation ecosystem or even as a set of interlinking ecosystems [96]. This means that interventions aimed at supporting human flourishing and addressing ethical issues need to be looked at using a systems perspective [94]. In such a view, an ex-ante assessment process like the one suggested by ALTAI and replicated here needs to be embedded in other processes that can shape the socio-technical innovation system in which AI is located.

The systems perspective can address several of the issues with the ALTAI application we have described. By focussing on AI as a system, more exactly as a socio-technical ecosystem that is embedded in and affects other socio-technical ecosystems, the conceptual delimitation of what counts as AI and what does not becomes less important, as the focus is on the activities and consequences of the existence of these ecosystems in which AI in the narrow technical sense is only one constituent part. The systems perspective moreover can help overcome the problematic assumption that there are linear and predictable causal chains in technology development which would allow the accurate prediction of consequences at the research and development stage of a technology. Instead, systems thinking highlights the interdependence of systems and their components and the need to reflect on feedback loops and other non-linear relationships.

Using such a systems perspective allows putting the ALTAI approach into a larger perspective. The ethical and regulatory ecosystem in which AI is used and developed includes a much broader array of interventions and ways of thinking about and dealing with ethical and social concerns than just ex-ante assessments. The discussion of AI ethics includes national and international legislation and regulation, such as the proposed EU AI Act [47]. It is important to realise, however that it is not just the big headline activities that govern AI, but that there are numerous other activities, rules, tools, etc. that have an influence. In legislative terms, for example, there are laws and regulations around data protection, competition, intellectual property, liability and many others that can influence AI development and use. Principles of human rights have a bearing on AI as do other national and international policies, ranging from taxation and procurement to environmental sustainability and defence. In many cases, these are already integrated into organisational routines through various processes and requirements. Organisations tend to have risk assessment processes in place that can include AI. More broadly, organisational governance structures including data governance, auditing but increasingly also ethics reviews can be applied to AI. Corporate social responsibility and stakeholder engagement work can be developed to include a sensitivity to technology. In addition, there are numerous other approaches and guidance mechanisms aimed at ensuring ethical acceptability of AI. These include high-level ethical frameworks [66, 86], standardisation initiatives [63] including certification schemes [34, 62], professional body support [28], and various development methodologies [23, 101] and ethics tools [78].

Keeping this complexity of the AI ecosystem in mind, it is easy to see that an ex-ante impact assessment such as the ALTAI has a potentially important role to play. It can serve as a sensitising mechanism that encourages researchers, developers and users to consider issues and deliberate how these may be addressed. Such a reflective exercise can be triggered by other aspects of the ecosystem, for example, it can be mandated by regulation. It can also function as input into subsequent steps, for example when the outcomes of ALTAI are used as starting points for the choice of an appropriate development methodology. This perspective takes the pressure off the ex-ante assessment to be the definitive solution and thereby renders the conceptual and practical concerns less significant, as they may be picked up elsewhere in the larger system. This leads to the question of how ALTAI and similar approaches can be integrated into the larger AI Ecosystem.

5.3 The integration of ex-ante assessments into the AI ecosystem

The responses to our online survey have highlighted some gaps in the consideration of ethical issues related to AI research in the HBP. Simply highlighting the existence of such gaps does not resolve them, and the nature of the survey means that further research will need to be undertaken to gain a more comprehensive understanding of such. We will therefore work with the respondents to ensure that ethical issues related to AI are better understood and that existing and novel ways of dealing with and thinking about them are applied appropriately, both in relation to research in the HBP, and (hopefully) beyond. This will be a specific task for the project team, and will be dealt with in the context of the ethics and society work of the HBP as described elsewhere [95]. However, it is a task that finds its equivalent on the broader societal scene where ALTAI and similar need to be integrated into the AI ecosystem.

Due to the complexity of ecosystems overall and the highly international and interdisciplinary nature of AI, it is difficult to provide clear guidelines or recommendations on how this is to be achieved. However, the observations and insights discussed earlier can provide some pointers to activities that are likely to strengthen ethical and broader social considerations in the AI ecosystem.

The first of these suggestions applies to all AI ecosystem participants. It refers to the concept of AI. The term has now become entrenched in the debate and will be impossible to change by decree. It is therefore important to ensure that it explicitly covers all those digital technologies that are capable of giving rise to concerns and goes beyond narrow AI in the sense of current machine learning or symbolic AI approaches. It should thus cover all technologies that can collect or process digital data and use them to act on the world, so that technologies like the Internet of Things, edge computing, quantum computing etc. are explicitly included.

A second suggestion will be to explicitly integrate ex-ante assessments into the broader AI ethics landscape. It is not enough to speculate that such assessments can be useful in various ways, for example as triggers and input into ethics by design development methodologies or risk management structures. These links need to be more clearly defined and pathways through the AI ecosystem need to be charted. This is a task that at some level requires legislation and regulation on the national or international scale. However, many other stakeholders, such as professional bodies, AI-driven organisations can play a role in linking different aspects of the AI ecosystem.

In order for these links between different aspects and approaches to ethics of AI to come together, it will be important to go beyond pathways and procedures. A crucial component will be the collection and sharing of experience and good practice. This calls for creating and maintaining organisations that serve as official gatekeepers of the ethics of AI. In practice, it is likely that these will be the national AI champions or regulators who have a role as gatekeepers of AI more broadly. Such organisations where they already exist will have good access to experiences of using different tools such as ex-ante assessments and will be able to collect data on their strengths and weaknesses. Such gatekeeper organisations could be public and regulatory bodies such as current data protection authorities that exist in many countries. They could also be academic institutions, such as the Alan Turing Institute in the UK which brings together leading AI research from across the country and provides input into policy development.

The collection of experience and good practice will need to go beyond the ex-ante assessment and should cover examples of addressing specific issues. Examples of current approaches are the work done on explainable AI or identification of bias in data and algorithms. These are areas that currently attract significant research interest. It is not yet established which approaches are most suitable to address which types of technologies or applications. For the ex-ante assessment to be fruitful, they need to be linked to suitable mechanisms for addressing the concerns that the assessment reveals.

These suggestions cannot claim to be comprehensive. They provide a starting point for thinking about ethical and social questions of AI in terms of broader AI ecosystems and address the questions how such ecosystems can be shaped to promote beneficial uses and outcomes and avoid undesirable ones. Due to the lack of predictability of interventions in systems, it is important that the AI ecosystems are endowed with an ability to learn and modify. This is a key feature of technical machine learning systems and this ability to learn and adapt must be a characteristic of the broader socio-technical ecosystems into which technical AI is embedded.

In this view of AI, it is plausible to assume that ex-ante approaches such as the ALTAI can play an important role in identifying and paving the way for addressing ethical and social concerns. Taking the systems nature of AI seriously offers ways to shape these AI ecosystems to general advantage. A failure to take this position and the resulting assumption that ethical and social issues can be comprehensively addressed via ex-ante assessments is likely to have the opposite effect. We believe that our research has demonstrated and provided the evidence to support this view. As a result, we hope that we can contribute to the further development of ALTAI and similar ex-ante approaches and to ensure that they are embedded into the broader societal and political discussion and can find suitable roles in the AI ecosystem to ensure that benefits of AI are retained whilst its problems are addressed early.

Appendix 1: Survey instrument

The Ethics review report of the M9 Ethics Review of the HBP SGA3 (Ares (2021)2194932-30/03/2021) defined a new ethics requirement:

NEW REQUIREMENT in relation to the significant use or development of AI in HBP and the development and now availability of related AI ethics guidance in Europe.

In case Artificial Intelligence will be used:

The applicant/beneficiary must explain how the developed/used AI meets the criteria for trustworthiness. The report must be submitted as a deliverable and must cover in particular the following:

- Measures set in place to avoid potential bias, discrimination and stigmatisation;
- Measures set in place to ensure safety and prevention of harms (to humans, animals, and environment);
- Detailed explanation on how the respect of fundamental human rights and freedoms (e.g. human autonomy, privacy and data protection) will be ensured;
- Measures to ensure fairness and explicability (paying particular attention to situations involving more vulnerable groups).

The applicant/beneficiary must explain how humans will maintain meaningful control over the most important aspects of decision-making process.

The applicant/beneficiary must evaluate the ethics risks related to the development/deployment of the AI systems/techniques and explain how the potential negative social impacts will be mitigated. (p.10).

This questionnaire forms the basis for this new AI Deliverable, which sits within the remit of Task 3.8. The findings of this questionnaire will only be used for the completion of this deliverable. However, you will be asked if you would like to be contacted regarding any follow up research related to this field of research.

If you have any questions or concerns about this questionnaire, please contact Tonii Leach (antonia.leach@dmu.ac.uk).

Please find the applicable De Montfort University, UK, Data Protection Policy and Privacy Notice here: https://www.dmu.ac.uk/policies/data-protection/data-protection.aspx

Section 1: Background Information

Please provide the following information about yourself and your work in the HBP.

1. Name:
2. Email address:
3. Job title:
4. Please select all work packages you are affiliated with:

WP1
WP2
WP3
WP4
WP5
WP6
WP7
WP8
WP9

5. Please provide the task number of all tasks you are affiliated with:

Section 2: AI use in the HBP

This section will ask about the use or development of AI (Artificial Intelligence) within the HBP.

For the purposes of this deliverable, the following definition of AI is used:

> "Artificial intelligence (AI) systems are software (and possibly also hardware) systems designed by humans that, given a complex goal, act in the physical or digital dimension by perceiving their environment through data acquisition, interpreting the collected structured or unstructured data, reasoning on the knowledge, or processing the information, derived from this data and deciding the best action(s) to take to achieve the given goal. AI systems can either use symbolic rules or learn a numeric model, and they can also adapt their behaviour by analysing how the environment is affected by their previous actions" ([6, 7], p. 24).

High Level Expert Group on Artificial Intelligence (2019) The Assessment List for Trustworthy Artificial Intelligence.

6. Given the definition of AI provided, do you, within your role in the HBP, develop or use AI?

Yes

No

7. Please provide a brief overview of your work using or developing AI in the HBP.

You are welcome to provide weblinks where these would be helpful.

Section 3

Thank you.

You will now be asked a few, short questions about the AI you develop/use within your role in the HBP.

If you use/develop more than one AI system, please think about the AI system you use/develop most often when answering this questionnaire. We will ask you to complete this questionnaire for each additional AI system you use/develop in your role with the HBP. We will be in contact by email to provide you with information in completing the additional questionnaires.

These questions should not take more than 15 min to complete.

Section 4: Significance of the AI system

This question aims to understand the significance of the AI system.

8. Is the AI system designed to interact with, guide or take decisions for human end-users that affect humans or society?

Yes

No

Section 5: Trustworthiness of AI

The following questions relate to trustworthiness of the AI system you use/develop in your role with the HBP, and the seven requirements of trustworthy AI as outlined in the The Assessment List For Trustworthy Artificial Intelligence (ALTAI) [6, 7].

Section 6: Human Agency and Oversight

Human Agency and Oversight refers to the effect AI systems can have on human behaviour in the broadest sense. It deals with the effect of AI systems that are aimed at guiding, influencing or supporting humans in decision-making processes, for example, algorithmic decision support systems, risk analysis/prediction systems (recommender systems, predictive policing, financial risk analysis, etc.). It also deals with the effect on human perception and expectation when confronted with AI systems that 'act' like humans. Finally, it deals with the effect of AI systems on human affection, trust and (in)dependence.

9. What type of AI system do you use/develop?

- Human-in-the-loop refers to the capability for human intervention in every decision cycle of the system.
- Human-on-the-loop refers to the capability for human intervention during the design cycle of the system and monitoring the system's operation.
- Human-in-command refers to the capability to oversee the overall activity of the AI system (including its broader economic, societal, legal and ethical impact) and the ability to decide when and how to use the AI system in any particular situation.

A self-learning or autonomous system

A system overseen by a Human-in-the-Loop

A system overseen by a Human-on-the-Loop

A system overseen by a Human-in-Command

10. Could the AI system affect human agency or autonomy?

Yes

No

11. Please explain measures put in place to ensure that human agency and autonomy, and human oversight have been ensured.

Section 7: Technical Robustness and Safety

Technical Robustness and Safety refers to dependability (the ability to deliver services that can justifiably be trusted) and resilience (robustness when facing changes, for example to the use context, such as the application domain or life cycle phase).

12. Could the AI system have adversarial, critical or damaging effects (e.g. to human or societal safety) in case of risks or threats, such as design or technical faults, defects, outages, attacks, misuse, inappropriate or malicious use?

Yes

No

13. Please outline the measures put in place to ensure that human and societal safety has been ensured.

14. Could a low level of accuracy of the AI system result in critical, adversarial or damaging consequences?

Accuracy here refers to the predictive capability of the AI model and the ability to generalise well for unseen data.

Yes

No

15. Please outline the measures put in place to ensure that a high level of accuracy has been ensured.

16. Could the AI system cause critical, adversarial, or damaging consequences (e.g. pertaining to human safety) in case of low reliability and/or reproducibility?

Yes

No

17. Please outline the measures put in place to mitigate risks to human safety in the event of low reliability.

18. Is there a risk of the possible malicious use, misuse or inappropriate.

use of the AI system?

Yes

No

19. Please outline the measures put in place to mitigate risks related to the malicious use, misuse or inappropriate use of the AI system.

Section 8: Privacy and Data Governance

Privacy is a fundamental right that is likely to be particularly affected by AI systems. Prevention of harm to privacy also necessitates adequate data governance that covers the quality and integrity of the data used, its relevance in the light of the domain in which the AI systems will be deployed, its access protocols and the capability to process data in a manner that protects privacy.

20. Could the AI system impact on the right to privacy, the right to physical, mental and/or moral integrity and the right to data protection?

Yes

No

21. Please outline the measures put in place to ensure the right to privacy and the right to data protection.

22. What type of data is your AI system being trained on?

Human data

Non-human primate data

Other animal data

Synthetic data

Anonymous statistics derived from data.

23. Does the human data that the AI system is being trained on include personal data (including any special categories of personal data)?

Yes

No

24. Please outline the measures put in place relating to good data governance (including, where appropriate, GDPR (General Data Protection Regulation) and DPIAs (Data Protection Impact Assessments).

Section 9: Transparency

A crucial component of achieving Trustworthy AI is transparency which encompasses three elements: (1) traceability, (2) explainability and (3) open communication about the limitations of the AI system.

This question refers to 'blackbox' systems. These are defined as:

AI systems whereby 'An explanation as to why a model has generated a particular output or decision (and what combination of input factors contributed to that) is not always possible' ([6, 7], p.14).

This question also refers to 'end-users'. These are defined as:

> 'An end-user is the person that ultimately uses or is intended to ultimately use the AI system. This could either be a consumer or a professional within a public or private organisation. The end-user stands in contrast to users who support or maintain the product, such as system administrators, database administrators, information technology experts, software professionals and computer technicians.'

25. Is the AI system designed to make decisions about human end-users?

Yes

No

26. Is the AI system likely to act as a 'blackbox'?

Yes

No

27. Please outline measures put in place to ensure the explainability of the AI system.

28. Is the AI system designed to interact with human end-users?

Yes

No

29. Please outline how you ensure that human end-users are aware that they are interacting with an AI system.

Section 10: Diversity, Non-discrimination and Fairness

Inclusion and diversity should be ensured throughout the entire AI system's life cycle. AI systems (both for training and operation) may suffer from the inclusion of inadvertent historic bias, incompleteness, and bad governance models. The continuation of such biases could lead to unintended (in)direct prejudice and discrimination against certain groups or people, potentially exacerbating prejudice and marginalisation.

30. Is the AI system being trained on data that may have issues of bias (historic or otherwise), discrimination, or a lack of diversity?

Yes

No

31. Please outline measures put in place to mitigate bias, discrimination or a lack of diversity in the data.

32. Could the AI system act in a way that might cause discriminatory/biassed outcomes for humans/end-users?

Yes

No

33. Please outline measures put in place to mitigate biassed/discriminatory outcomes from the AI system.

34. Does the AI system interact with humans?

Yes

No

35. Is the AI system accessible or designed for use by people with a range of abilities?

Yes

No

36. Please outline measures taken to ensure that the design of the AI system is not inaccessible or does not disadvantage particular groups of users.

37. Is the AI system capable of utilising 'subliminal techniques'?

Subliminal techniques are practices that have 'a significant potential to manipulate persons through subliminal techniques beyond their consciousness or exploit vulnerabilities of specific vulnerable groups such as children or persons with disabilities to materially distort their behaviour in a manner that is likely to cause them or another person psychological or physical harm.'

Yes

No

38. Please outline measures taken to prevent prohibited use subliminal techniques.

Section 11: Societal and Environmental Well-being

The broader society, other sentient beings and the environment should be considered as stakeholders throughout the AI system's life cycle. The effects of AI systems must therefore be carefully monitored and considered. Sustainability and ecological responsibility of AI systems should be encouraged, and research should be fostered into AI solutions addressing areas of global concern, for instance the Sustainable Development Goals. Overall, AI should be used to benefit all human beings, including future generations. AI systems should serve to maintain and foster democratic processes and respect the plurality of values and life choices of individuals. AI systems must not undermine democratic processes, human deliberation or democratic voting systems or pose a systemic threat to society at large.

39. Are there potential negative impacts of the AI system on the environment?

Yes

No

40. Please outline the potential negative impacts of the AI system on the environment and measures in place to mitigate these.

41. Could the AI system impact human work and work arrangements?

Yes

No

42. Please outline the potential impacts on human work and work arrangements and measures in place to mitigate these where they are negative.

43. Could the AI system have a negative impact on society at large or democracy?

Yes

No

44. Please outline the potential negative impacts on society at large or democracy and measures in place to mitigate these.

Section 12: Accountability

The principle of accountability necessitates that mechanisms be put in place to ensure responsibility for the development, deployment and/or use of AI systems.

45. Is the development process of the AI system, the sourcing of training data, and the processes, outcomes, and possible impacts of the AI system recorded?

Yes

No

46. Please outline how the ability of a third party to audit the AI system has been ensured.

47. Are there mechanisms in place (such as Ethics Review Boards, places to register ethical concerns, ethics support systems) to discuss the overall accountability and ethics practices, including potential unclear grey areas in relation to the AI system?

Yes

No

48. Please outline how AI ethics have been integrated into the development of the AI system.

49. Do you have any other concerns about the use or development of AI in the HBP?

Section 13: Other AI Systems

50. Do you use or develop any other AI systems in your role with the HBP?

Yes

No

51. How many other AI systems do you use/develop?

1

2

3

4

5 or more

Section 14

Thank you.

We will ask you to complete this questionnaire for each AI system you use/develop in your role with the HBP.

We will be in contact by email to provide you with information in completing the additional questionnaires.

Section 15: Other AI Research in the HBP

This deliverable requires us to consider the ethical implications of all significant uses of AI in the HBP—however, given the large number of researchers, tasks and work packages in the HBP, identifying each case of AI is difficult.

As such, we would greatly appreciate your help in identifying other researchers who might be using/developing AI in the HBP.

If you are aware of any other researchers, or specific tasks, within that HBP that use or develop AI, we would greatly appreciate you sharing this information so that we can contact them regarding this deliverable.

Please provide name[s] and/or email addresses for individual researchers, or the Task number or title for research projects.

52.Details of other researchers/tasks using or developing AI in the HBP:

Section 16: Follow up Research

The researchers involved in this deliverable (members of Task 3.8) would like to undertake some follow up research with developers/users of AI in the HBP on the topic of ethical and trustworthy AI.

53. Please provide any additional information or feedback in relation to this questionnaire that you wish to be considered in any further research.

54. Would you be happy to be contacted in the future (within SGA 3) about any follow up research?

Yes

No

Section 17: Thank you

Thank you for completing this questionnaire. Your responses have been recorded.

If you have any questions or concerns about this questionnaire, please contact Tonii Leach (antonia.leach@dmu.ac.uk).

You may now close this page.

Acknowledgements The authors would like to acknowledge the contribution of the many HBP project members and members of the Ethics Advisory Board who reviewed earlier versions of the survey, gave feedback on formulation, provided responses, and allowed us to use the data for this paper.

Author contributions Both authors contributed to the paper in equal parts. TL led on the survey design, data collection and analysis as well as the production of the deliverable. BS led on the drafting of the journal paper.

Funding The paper arises from research undertaken by the EU-funded Human Brain Project SGA3, GA no 945539.

Availability of data and materials The survey instrument is included as an annex to the paper. Individual responses cannot be shared, to ensure confidentiality of respondents.

Declarations

Conflict of interest There are no conflicts of interest related to the paper.

References

1. Access Now.: Human Rights in the Age of Artificial Intelligence. Access Now (2018).
2. Access Now Policy Team: The Toronto Declaration: Protecting the Right to Equality and Non-discrimination in Machine Learning Systems. Access No, Toronto (2018)

3. AI Council.: AI Roadmap. Office for Artificial Intelligence, Department for Business, Energy & Industrial Strategy, and Department for Digital, Culture, Media & Sport, London (2021).
4. Hleg, A.I.: Assessment List for Trustworthy AI (ALTAI). European Commission, Brussels (2020)
5. Hleg, A.I.: Sectorial considerations for Trustworthy AI—taking AI's context specificity into account. European Commission, Brussels (2020)
6. Hleg, A.I.: Ethics Guidelines for Trustworthy AI. European Commission—Directorate-General for Communication, Brussels (2019)
7. Hleg, A.I.: Policy and investment recommendations for trustworthy Artificial Intelligence. European Commission-Directorate-General for Communication, Brussels (2019)
8. Hleg, A.I.: A definition of AI: main capabilities and scientific disciplines. European Commission, Brussels (2018)
9. AI Now Institute.: Algorithmic impact assessments: a practical framework for public agency accountability (2018).
10. Aicardi, C., Bitsch, L., Datta Burton, S., Evers, K., Farisco, M., Mahfoud, T., Rose, N., Rosemann, A., Salles, A., Stahl, B., Ulnicane, I.: Opinion on trust and transparency in artificial intelligence—ethics & society. Hum. Brain Project (2021). https://doi.org/10.5281/zenodo.4588648
11. Aicardi, C., Reinsborough, M., Rose, N.: The integrated ethics and society programme of the Human Brain Project: reflecting on an ongoing experience. J. Responsib. Innov. (2017). https://doi.org/10.1080/23299460.2017.1331101
12. AIEI Group.: From Principles to Practice—An Interdisciplinary framework to operationalise AI ethics. VDE/Bertelsmann Stiftung (2020).
13. Amunts, K., Ebell, C., Muller, J., Telefont, M., Knoll, A., Lippert, T.: The Human Brain Project: creating a European research infrastructure to decode the human brain. Neuron **92**, 574–581 (2016). https://doi.org/10.1016/j.neuron.2016.10.046
14. Aronson, J.: A pragmatic view of thematic analysis. Qual. Rep. **2**, 1–3 (1995)
15. Babuta, A., Oswald, M., Janjeva, A.: Artificial Intelligence and UK National Security—Policy Considerations (Occasional Paper). Royal United Services Institute for Defence and Security Studies (2020).
16. Beatty, P., Reay, I., Dick, S., Miller, J.: Consumer trust in e-commerce web sites: a meta-study. ACM Comput. Surv. **43**, 141–1446 (2011). https://doi.org/10.1145/1922649.1922651
17. Becker, H.A.: Social impact assessment. Eur. J. Oper. Res. **128**, 311–321 (2001). https://doi.org/10.1016/S0377-2217(00)00074-6
18. Becker, H.A., Vanclay, F.: The International Handbook of Social Impact Assessment: Conceptual and Methodological Advances. Edward Elgar Publishing, Cheltenham (2003)
19. Bengio, Y., Lecun, Y., Hinton, G.: Deep learning for AI. Commun. ACM **64**, 58–65 (2021). https://doi.org/10.1145/3448250
20. Benjamins, R.: A choices framework for the responsible use of AI. AI Ethics (2020). https://doi.org/10.1007/s43681-020-00012-5
21. Bhattacherjee, A.: Individual trust in online firms: Scale development and initial test. J. Manag. Inf. Syst. **19**, 211–241 (2002)
22. Boden, M.A.: Artificial Intelligence: A Very Short Introduction, Reprint edition. ed. OUP Oxford, Oxford (2018).
23. Borenstein, J., Grodzinsky, F.S., Howard, A., Miller, K.W., Wolf, M.J.: AI ethics: a long history and a recent burst of attention. Computer **54**, 96–102 (2021). https://doi.org/10.1109/MC.2020.3034950
24. Bostrom, N.: Superintelligence: Paths, Dangers, Strategies, Reprint edition. ed. OUP Oxford, Oxford (2016).
25. Bowie, N.E.: Business Ethics: A Kantian Perspective. Blackwell Publishers, New York (1999)
26. Brattberg, E., Rugova, V., Csernatoni, R.: Europe and AI: leading, lagging behind, or carving its own way? Carnegie Endowment for International Peace (2020).
27. Brenkert, G.G.: Trust, business and business ethics: an introduction. Bus. Ethics Q. **8**, 195–203 (1998)
28. Brinkman, B., Flick, C., Gotterbarn, D., Miller, K., Vazansky, K., Wolf, M.J.: Listening to professional voices: Draft 2 of the ACM code of ethics and professional conduct. Commun. ACM **60**, 105–111 (2017). https://doi.org/10.1145/3072528
29. Brundage, M., Avin, S., Clark, J., Toner, H., Eckersley, P., Garfinkel, B., Dafoe, A., Scharre, P., Zeitzoff, T., Filar, B., Anderson, H., Roff, H., Allen, G.C., Steinhardt, J., Flynn, C., hÉigeartaigh, S.Ó., Beard, S., Belfield, H., Farquhar, S., Lyle, C., Crootof, R., Evans, O., Page, M., Bryson, J., Yampolskiy, R., Amodei, D.: The Malicious Use of Artificial Intelligence: Forecasting, Prevention, and Mitigation (2018). arXiv:1802.07228 [cs].
30. Cave, S.: The problem with intelligence: its value-laden history and the future of AI. In: Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society, pp. 29–35. Association for Computing Machinery, New York (2020).
31. CDEI.: Interim Report: Review into Bias in Algorithmic Decision-making. Centre for Data Ethics and Innovation (2019).
32. CEN-CENELEC.: Ethics assessment for research and innovation—Part 2: Ethical impact assessment framework (CEN Workshop Agreement No. CWA 17145-2:2017 (E)). CEN-CENELEC, Brussels (2017).
33. Christen, M., Biller-Andorno, N., Bringedal, B., Grimes, K., Savulescu, J., Walter, H.: Ethical challenges of simulation-driven big neuroscience. AJOB Neurosci. **7**, 5–17 (2016). https://doi.org/10.1080/21507740.2015.1135831
34. Cihon, P., Kleinaltenkamp, M.J., Schuett, J., Baum, S.D.: AI Certification: advancing ethical practice by reducing information asymmetries. IEEE Trans. Technol. Soc. **2**, 200–209 (2021). https://doi.org/10.1109/TTS.2021.3077595
35. Clarke, R.: Privacy impact assessment: its origins and development. Comput. Law Secur. Rev. **25**, 123–135 (2009). https://doi.org/10.1016/j.clsr.2009.02.002
36. CNIL.: Privacy Impact Assessment (PIA) Good Practice. CNIL (2015).
37. Coeckelbergh, M.: AI Ethics. The MIT Press, Cambridge (2020)
38. Coeckelbergh, M.: Technology, narrative and performance in the social theatre. In: Kreps, D. (ed.) Understanding Digital Events: Bergson, Whitehead, and the Experience of the Digital, pp. 13–27. Routledge, New York (2019)
39. Collingridge, D.: The Social Control of Technology. Palgrave Macmillan, London (1981)
40. de Laat, P.B.: Companies Committed to Responsible AI: From Principles towards Implementation and Regulation? Philos. Technol. (2021). https://doi.org/10.1007/s13347-021-00474-3
41. Dehaene, S., Lau, H., Kouider, S.: What is consciousness, and could machines have it? Science **358**, 486–492 (2017)
42. Dignum, V.: Responsible Artificial Intelligence: How to Develop and Use AI in a Responsible Way, 1st edn. Springer, Berlin (2019)
43. Donaldson, T., Dunfee, T.W.: Ties that Bind: a Social Contracts Approach to Business Ethics. Harvard Business Press, Harvard (1999)
44. EDPS.: EDPS 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). EDPS (2020).
45. Elsevier: Artificial Intelligence: How knowledge is created, transferred, and used—Trends in China, Europe, and the United States. Elsevier, Amsterdam (2018)

46. Etzioni, A., Etzioni, O.: Incorporating ethics into artificial intelligence. J Ethics **21**, 403–418 (2017). https://doi.org/10.1007/s10892-017-9252-2
47. European Commission.: Proposal for a Regulation on a European approach for Artificial Intelligence (No. COM (2021) 206 final). European Commission, Brussels (2021).
48. European Commission.: EU Grants: How to complete your ethics self-assessment—V2.0. Brussels (2021).
49. FRA.: Getting the future right—Artificial intelligence and fundamental rights. European Union Agency for Fundamental Rights, Luxembourg (2020).
50. GDPR.: REGULATION (EU) 2016/679 OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation). Official Journal of the European Union L119/1 (2016).
51. Grabner-Kraeuter, S.: The role of consumers' trust in online-shopping. J. Bus. Ethics **39**, 43–50 (2002)
52. Grunwald, A.: Technology Assessment in Practice and Theory, 1st edn. Routledge, Abingdon (2018)
53. Grunwald, A.: Technology assessment or ethics of technology? Ethical Perspect. **6**, 170–182 (1999)
54. Guterres, A.: The Highest Aspiration—A Call to Action for Human Rights. United Nations (2020).
55. Hall, W., Pesenti, J.: Growing the artificial intelligence industry in the UK. Department for Digital, Culture, Media & Sport and Department for Business, Energy & Industrial Strategy, London (2017).
56. Hammer, D., Berland, L.K.: Confusing claims for data: a critique of common practices for presenting qualitative research on learning. J. Learn. Sci. **23**, 37–46 (2014). https://doi.org/10.1080/10508406.2013.802652
57. Hartley, N., Wood, C.: Public participation in environmental impact assessment—implementing the Aarhus convention. Environ. Impact Assess. Rev. **25**, 319–340 (2005). https://doi.org/10.1016/j.eiar.2004.12.002
58. Hoffman, D.L., Novak, T.P., Peralta, M.: Building consumer trust online. Commun. ACM **42**, 80–85 (1999)
59. Hole, K.J., Ahmad, S.: A thousand brains: toward biologically constrained AI. SN Appl. Sci. **3**, 743 (2021). https://doi.org/10.1007/s42452-021-04715-0
60. Huber, R., Weber, P.: Is there a relationship between socioeconomic factors and prevalence, adherence and outcome in childhood epilepsy? A systematic scoping review. Eur. J. Paediatr. Neurol. **38**, 1–6 (2022). https://doi.org/10.1016/j.ejpn.2022.01.021
61. IEEE.: IEEE 7010-2020—IEEE Recommended Practice for Assessing the Impact of Autonomous and Intelligent Systems on Human Well-Being (Standard). IEEE (2020).
62. IEEE.: IEEE SA—The Ethics Certification Program for Autonomous and Intelligent Systems (ECPAIS) [WWW Document] (2019). https://standards.ieee.org/industry-connections/ecpais.html. Accessed 4 Oct 2020).
63. IEEE Computer Society.: IEEE Standard Model Process for Addressing Ethical Concerns during System Design—7000-2021 (Standard), 7000-2021 (2021).
64. Information Commissioner's Office.: Privacy Impact Assessment Handbook, v. 2.0 (2009).
65. Ivanova, Y.: The Data Protection Impact Assessment as a Tool to Enforce Non-discriminatory AI. Lecture Notes in Computer Science (2020).
66. Jobin, A., Ienca, M., Vayena, E.: The global landscape of AI ethics guidelines. Nat. Mach. Intell. **1**, 389–399 (2019). https://doi.org/10.1038/s42256-019-0088-2
67. Kaplan, A., Haenlein, M.: 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 (2019)
68. Kazim, E., Koshiyama, A.S.: A high-level overview of AI ethics. PATTER (2021). https://doi.org/10.1016/j.patter.2021.100314
69. Koehn, D.: The nature of and conditions for online trust. J. Bus. Ethics **43**, 3–19 (2003)
70. Koehn, D.: Trust and business: barriers and bridges. Bus. Prof. Ethics J. **16**, 7–28 (1997)
71. Lane, C., Bachmann, R.: The social constitution of trust: supplier relations in Britain and Germany. Organ. Stud. **17**, 365–395 (1996)
72. Latonero, M.: Governing artificial intelligence: upholding human rights & dignity. Data & Society (2018).
73. Mayer, R.C., Davis, J.H., Schoorman, F.D.: An integrative model of organizational trust. Acad. Manag. Rev. **20**, 709–734 (1995)
74. Mayring, P.: Qualitative content analysis. Companion Qual. Res. **1**, 159–176 (2004)
75. Metzinger, T.: Ethics washing made in Europe. Der Tagesspiegel (2019).
76. Miles, M.B., Huberman, A.M.: Qualitative Data Analysis: An Expanded Sourcebook. SAGE, Thousand Oaks (1994)
77. Montes, G.A., Goertzel, B.: Distributed, decentralized, and democratized artificial intelligence. Technol. Forecast. Soc. Change **141**, 354–358 (2019). https://doi.org/10.1016/j.techfore.2018.11.010
78. Morley, J., Floridi, L., Kinsey, L., Elhalal, A.: From what to how: an initial review of publicly available AI ethics tools, methods and research to translate principles into practices. In: Floridi, L. (ed.) Ethics, Governance, and Policies in Artificial Intelligence, Philosophical Studies Series, pp. 153–183. Springer International Publishing, Cham (2021). https://doi.org/10.1007/978-3-030-81907-1_10
79. Muller, C.: The Impact of Artificial Intelligence on Human Rights, Democracy and the Rule of Law (No. CAHAI(2020)06-fin). Council of Europe, Ad Hoc Committee on Artificial Intelligence (CAHAI), Strasbourg (2020).
80. Nemitz, P.: Constitutional democracy and technology in the age of artificial intelligence. Philos. Trans. R. Soc. A **376**, 20180089 (2018). https://doi.org/10.1098/rsta.2018.0089
81. Nishant, R., Kennedy, M., Corbett, J.: Artificial intelligence for sustainability: challenges, opportunities, and a research agenda. Int. J. Inf. Manag. **53**, 102104 (2020). https://doi.org/10.1016/j.ijinfomgt.2020.102104
82. Rai, A., Constantinides, P., Sarker, S.: Next-generation digital platforms: toward human–AI hybrids. MIS Q. **43**, iii–x (2019)
83. Richards, L., Brockmann, K., Boulanini, V.: Responsible Artificial Intelligence Research and Innovation for International Peace and Security. Stockholm International Peace Research Institute, Stockholm (2020).
84. Rose, N.: The human brain project: social and ethical challenges. Neuron **82**, 1212–1215 (2014). https://doi.org/10.1016/j.neuron.2014.06.001
85. Ryan, M.: In AI we trust: ethics, artificial intelligence, and reliability. Sci. Eng. Ethics **26**, 2749–2767 (2020). https://doi.org/10.1007/s11948-020-00228-y
86. Ryan, M., Stahl, B.C.: Artificial intelligence ethics guidelines for developers and users: clarifying their content and normative implications. J. Inf. Commun. Ethics Soc. (2020). https://doi.org/10.1108/JICES-12-2019-0138
87. Salles, A., Bjaalie, J.G., Evers, K., Farisco, M., Fothergill, B.T., Guerrero, M., Maslen, H., Muller, J., Prescott, T., Stahl, B.C., Walter, H., Zilles, K., Amunts, K.: The human brain project: responsible brain research for the benefit of society. Neuron **101**, 380–384 (2019). https://doi.org/10.1016/j.neuron.2019.01.005

88. Salles, A., Evers, K., Farisco, M.: Neuroethics and philosophy in responsible research and innovation: the case of the human brain project. Neuroethics **12**, 201–211 (2019). https://doi.org/10.1007/s12152-018-9372-9
89. Schölkopf, B., Locatello, F., Bauer, S., Ke, N.R., Kalchbrenner, N., Goyal, A., Bengio, Y.: Toward causal representation learning. Proc. IEEE **109**, 612–634 (2021). https://doi.org/10.1109/JPROC.2021.3058954
90. Shneiderman, B.: Design lessons from AI's two grand goals: human emulation and useful applications. IEEE Trans. Technol. Soc. **1**, 73–82 (2020). https://doi.org/10.1109/TTS.2020.2992669
91. Smith, N., Vickers, D.: Statistically responsible artificial intelligences. Ethics Inf Technol (2021). https://doi.org/10.1007/s10676-021-09591-1
92. Spiegelhalter, D.: Should we trust algorithms? Harv. Data Sci. Rev. (2020). https://doi.org/10.1162/99608f92.cb91a35a
93. Stahl, B.C.: Responsible innovation ecosystems: Ethical implications of the application of the ecosystem concept to artificial intelligence. Int. J. Inf. Manag. **62**, 102441 (2022). https://doi.org/10.1016/j.ijinfomgt.2021.102441
94. Stahl, B.C.: Artificial Intelligence for a Better Future: An Ecosystem Perspective on the Ethics of AI and Emerging Digital Technologies, SpringerBriefs in Research and Innovation Governance. Springer International Publishing, Berlin (2021)
95. Stahl, B.C., Akintoye, S., Fothergill, B.T., Guerrero, M., Knight, W., Ulnicane, I.: Beyond research ethics: dialogues in neuro-ICT research. Front. Hum. Neurosci. (2019). https://doi.org/10.3389/fnhum.2019.00105
96. Stahl, B.C., Andreou, A., Brey, P., Hatzakis, T., Kirichenko, A., Macnish, K., Laulhé Shaelou, S., Patel, A., Ryan, M., Wright, D.: Artificial intelligence for human flourishing—beyond principles for machine learning. J. Bus. Res. **124**, 374–388 (2021). https://doi.org/10.1016/j.jbusres.2020.11.030
97. Stix, C.: The ghost of AI governance past, present and future: AI governance in the European Union (2021). https://doi.org/10.48550/arXiv.2107.14099
98. Stone, P., Brooks, R., Brynjolfsson, E., Calo, R., Etzioni, O., Hager, G., Hirschberg, J., Kalyanakrishnan, S., Kamar, E., Kraus, S.: Artificial Intelligence and Life in 2030. One hundred year study on artificial intelligence: Report of the 2015–2016 Study Panel. Stanford University, Stanford, CA (2016). http://ai100.stanford.edu/2016-report. Accessed 6 Sept 2016.
99. Thaw, Y.Y., Mahmood, A.K., Dominic, P.D.D.: A Study on the Factors That Influence the Consumers Trust on Ecommerce Adoption (2009). arXiv:0909.1145 [cs].
100. UK Government.: National AI Strategy (2021).
101. Umbrello, S., van de Poel, I.: Mapping value sensitive design onto AI for social good principles. AI Ethics **1**, 283–296 (2021). https://doi.org/10.1007/s43681-021-00038-3
102. UNESCO.: First draft of the recommendation on the Ethics of Artificial Intelligence (No. SHS / BIO / AHEG-AI / 2020/4 REV.2). UNESCO, Paris (2020).
103. Vallor, S.: Technology and the Virtues: A Philosophical Guide to a Future Worth Wanting. Oxford University Press, Oxford (2016)
104. Veale, M.: A critical take on the policy recommendations of the EU high-level expert group on artificial intelligence. Eur. J. Risk Regul. (2020). https://doi.org/10.1017/err.2019.65
105. Veale, M., Binns, R., Edwards, L.: Algorithms that remember: model inversion attacks and data protection law. Philos. Trans. R. Soc. A Math. Phys. Eng. Sci. **376**, 20180083 (2018)
106. Walton, N., Nayak, B.S.: Rethinking of Marxist perspectives on big data, artificial intelligence (AI) and capitalist economic development. Technol. Forecast. Soc. Change **166**, 120576 (2021). https://doi.org/10.1016/j.techfore.2021.120576
107. Welty, B., Becerra-Fernandez, I.: Managing trust and commitment in collaborative supply chain relationships. Commun. ACM **44**, 67–73 (2001)
108. Willcocks, L.: Robo-Apocalypse cancelled? Reframing the automation and future of work debate. J. Inf. Technol. **35**, 286–302 (2020). https://doi.org/10.1177/0268396220925830
109. Wright, D.: A framework for the ethical impact assessment of information technology. Ethics Inf. Technol. **13**, 199–226 (2011). https://doi.org/10.1007/s10676-010-9242-6
110. Yeung, K.: Algorithmic regulation: a critical interrogation. Regul. Gov. **12**, 505–523 (2018)
111. Zicari, R.V., Brodersen, J., Brusseau, J., Düdder, B., Eichhorn, T., Ivanov, T., Kararigas, G., Kringen, P., McCullough, M., Möslein, F., Mushtaq, N., Roig, G., Stürtz, N., Tolle, K., Tithi, J.J., van Halem, I., Westerlund, M.: Z-Inspection®: a process to assess trustworthy AI. IEEE Trans. Technol. Soc. **2**, 83–97 (2021). https://doi.org/10.1109/TTS.2021.3066209
112. Zuboff, P.S.: The Age of Surveillance Capitalism: The Fight for a Human Future at the New Frontier of Power, 01 edn. Profile Books, London (2019)