DOI:
10.5281/zenodo.4588647

Document Version
Publisher's PDF, also known as Version of record

Human Brain Project

Ethics & Society

Trust and Transparency in Artificial Intelligence

Co-funded by
the European Union

Authors:

Christine Aicardi, Foresight Lab, King's College London
Lise Bitsch, Danish Board of Technology Foundation
Saheli Datta Burton, Foresight Lab, King's College London
Kathinka Evers, Centre for Research Ethics & Bioethics, Uppsala University
Michele Farisco, Centre for Research Ethics & Bioethics, Uppsala University
Tara Mahfoud, Foresight Lab, King's College London
Nikolas Rose, Foresight Lab, King's College London
Achim Rosemann, Centre for Computing and Social Responsibility, De Montfort University
Arleen Salles, Centre for Research Ethics & Bioethics, Uppsala University
Bernd Stahl, Centre for Computing and Social Responsibility, De Montfort University
Inga Ulnicane, Centre for Computing and Social Responsibility, De Montfort University

Contributors:

We gratefully acknowledge input, conversations and support from the Ethics Advisory Board (especially Josep Domingo-Ferrer, Michaela Mayrhofer, Kristin Bergtora Sandvik and Sven Nyholm), SP12 colleagues, the Ethics Rapporteurs and the HBP editorial team.

How to cite this document:
Ethics & Society, C. Aicardi, L. Bitsch, S. Datta Burton, K. Evers, M. Farisco, T. Mahfoud, N. Rose, A. Rosemann, A. Salles, B. C. Stahl & I. Ulnicane (2021) Trust and Transparency in Artificial Intelligence. Opinion of the Human Brain Project's Ethics & Society Subproject. DOI: 10.5281/zenodo.4588648

DOI: 10.5281/zenodo.4588648

Layout: Søren B. Jepsen, Danish Board of Technology Foundation
Cover Illustration: Søren B. Jepsen, Danish Board of Technology Foundation
First print 2021©

This research has received funding from the European Union's Horizon 2020 Framework Programme for Research and Innovation under the Specific Grant Agreement No. 785907 (Human Brain Project SGA2)

This research has received funding from the European Union's Horizon 2020 Framework Programme for Research and Innovation under the Specific Grant Agreement No. 945539 (Human Brain Project SGA3)

Trust and Transparency in Artificial Intelligence

Index

List of acronyms / abbreviations

AI	Artificial Intelligence
AI HLEG	The EU's High Level Expert Group on AI
EC	European Commission
EPSRC	Engineering and Physical Sciences Research Council
G20	Group of 20
HBP	Human Brain Project
HLEG	High Level Expert Group
OECD	Organisation for Economic Cooperation and Development
SGA	Specific Grant Agreement
SGA3	The third phase of the HBP (2020-2023)
SP	Subproject
USNSTC	United States National Science and Technology Council
WEF	World Economic Forum

Executive Summary

The Ethics and Society Subproject of the Human Brain Project has developed this Opinion to provide key insights into the current discussion on the social, ethical and regulatory aspects of artificial intelligence. The EU and numerous other bodies are promoting and implementing a wide range of policies aimed to ensure that AI is beneficial - that it serves society. The HBP as a leading project bringing together neuroscience and ICT is in an excellent position to contribute to and to benefit from these discussions. This Opinion therefore highlights some key aspects of the discussion, shows its relevance to the HBP and develops a list of six recommendations.

Part 1 introduces the aims and rationale of the Opinion. Part 2 defines AI and identifies the areas in which AI technology is developed and used in the HBP. Part 3 discusses the concepts of trust and trustworthiness and why these are central to discussions on the ethics and governance of AI. Part 4 explores the role of transparency and why it is seen as an essential element in the development of safe and trustworthy AI. Part 5 examines broader societal perspectives on AI, in particular possible implications for politics and democracy. Part 6 discusses the relevance of the previous sections for the HBP. The Opinion ends with a Conclusion and an overview of the six Recommendations.

Recommendations for the Human Brain Project

Recommendation One: Provide an overview of AI-related activities in the HBP

In order to create a basis for further ethical, social and RRI reflection, we recommend that the HBP undertakes a comprehensive overview of ongoing and emerging AI activities within the project.[1]

- The overview builds on previous HBP work on AI
- The overview will be a basis for developing an HBP approach to AI which positions itself in relation to the EU policies on AI.
- Such an overview needs to pay attention to contextual factors that may be playing a role in AI research (such as political agendas, commercialization, personal interests, and aspects such as gender, class, age, and race, among others).

Recommendation Two: Involve clinicians and other users and beneficiaries

We recommend that the HBP identifies those who are envisaged as users and beneficiaries of AI based technologies (e.g., clinicians, patients, citizens, public services, interest organisations, etc.) and involves them in the formulation of research problems and in the initial design of research projects. This would contribute to raising awareness of the needs and preferences of different groups of users.

- For clinical applications, this is a crucial step towards gaining the trust that is necessary for clinical translation. Researchers and engineers in the HBP need to understand and acknowledge the tacit, experiential knowledge involved in clinical reasoning, which has made it difficult in the past to incorporate algorithmic tools into the clinic.

1 RRI stands for Responsible Research and Innovation which according to the EU Horizon 2020 programme 'is an approach that anticipates and assesses potential implications and societal expectations with regard to research and innovation, with the aim to foster the design of inclusive and sustainable research and innovation'. Available at https://ec.europa.eu/programmes/horizon2020/en/h2020-section/responsible-research-innovation (Last accessed on 21 December 2020)

Recommendation Three: Include Ethics and RRI in the HBP's AI education programme

We recommend that the Human Brain Project develops an educational programme on AI for PhD students, early and mid-career researchers, as well as corporate partners, which specifically addresses AI Ethics and Responsible AI.

- Education could help researchers understand the possible societal and behavioural contexts of implementation for AI based applications and systems.
- A better understanding of these implications can help to anticipate undesirable effects at an early stage and lead to better technologies and solutions.
- Education should include training with scholars from the humanities and social sciences to address societal and ethical issues of AI-based and AI-developing research from a more interdisciplinary perspective.

Recommendation Four: Focus on the ethical and societal implications of commercialization

We recommend that the HBP and its partner projects undertake further work on the ethical, social and RRI dimensions of the translation of research into commercial AI and robotics products and services.

- This line of work should include a concern with the societal, political, economic and environmental consequences that are caused by disruptions of existing systems of production, social organisation, administration and political control.

Recommendation Five: Examine the effects of the international transfer of AI technologies from the EU to other world regions

We recommend that the HBP and its corporate partners give careful consideration to the possible implications of the international transfer of AI and robotics applications developed in the EU to other world regions.

- This should involve a concern with the ways in which individual products and services interact with and transform different social, digital and natural environments at both micro and macro levels, including the lives and well-being of different groups of technology users and citizens, in all their diversity.
- At the same time, decisions to enable or constrain international transfer should consider the international competitiveness of AI technologies developed in the EU.

Recommendation Six: Develop new methods to integrate RRI in the HBP's strategy to facilitate commercial exploitation of project findings and inventions

We recommend the development of new methods and ethical criteria to integrate RRI evaluation in the HBP's strategy to enable the exploitation and commercial use of emerging AI and (neuro)robotics applications.

- Such methodologies must be designed for researchers in the HBP and the private sector, including for firms and organisations involved in the development and distribution of these technologies
- They must be usable in public-private sector partnerships and consider the well-being and situation of key stakeholders, in particular those whose lives will be influenced by emerging AI and robotics products (e.g., employees, citizens, consumers, patients, etc.)
- They must be context-specific and consider RRI policies enshrined by H2020 and other relevant policies, including the EU's forthcoming rules on the trade of dual use items.[2]

2 https://www.consilium.europa.eu/en/press/press-releases/2020/11/09/new-rules-on-trade-of-dual-use-items-agreed/#

1. Introduction

Neuroscience and computer science are two key disciplinary domains contributing to, and benefiting from, ongoing developments in the highly interdisciplinary and varied field of artificial intelligence (AI). The Human Brain Project (HBP) undertakes interdisciplinary research across these disciplinary domains. For this reason, the HBP can learn from, but also contribute to, the broader discourse on AI. This includes the discussion of ethics and AI. The HBP's Ethics and Society Subproject has, therefore, developed this Opinion, to explore the relevance of the discussion of social and ethical issues in AI for the HBP.
The aim of this Opinion is to deepen and enrich the understanding of some of the key ethical and social concerns raised by AI in general and to suggest options to the HBP and other key actors for increased awareness and responsiveness to such issues. In particular, this Opinion focuses on trust, trustworthiness and transparency.

Trust and trustworthiness are central to discussions on the ethics of AI, as demonstrated in a 2019 review of the global corpus of principles and guidelines on ethical AI (Jobin et al. 2019). Transparency and open science are also core values of responsibility in scientific research and innovation. The recent European Commission (EC) White Paper On Artificial Intelligence - A European approach to excellence and trust also places questions of trust at its heart (European Commission 2020). These issues have become prominent in areas including ethical, legal, and regulatory debates over the use of algorithms in healthcare, with corporations and other organisations involved in the development and distribution of these technologies and the demand for explainability and accountability of seemingly inscrutable systems getting stronger (Price 2017; Vayena et al. 2018). As we discuss below, transparency is increasingly believed to be a mandatory step towards the safety and trustworthiness of AI and machine learning systems, and it is seen as integral to ethically-aligned design principles. Transparency is also consistently identified in ethics guidelines as a key requisite to building and achieving trust and trustworthiness (Jobin et al. 2019).

For this reason, the Ethics and Society Subproject of the HBP undertook social sciences and humanities research, organised a series of consultations, webinars, and workshops with citizens, experts, policy-makers, scientists and engineers and other stakeholders, to identify societal and ethical aspects of HBP research as it related to AI. These form the basis of this Opinion and its recommendations. They suggest that concerns about trust and transparency are central to the work of the HBP.

The present Opinion is intended first and foremost for an audience of researchers and developers in the HBP. It is intended as a starting point for more exhaustive work around trust and trustworthiness and their key requisites during the next HBP funding period (SGA3, 2020-2023). Its aim is to deepen our understanding of societal and ethical issues in relation to AI, and to point out ways in which the use or development of AI-based technologies and techniques in the HBP can be steered towards societal benefit.

2. Defining AI in the HBP

There is a multiplicity of definitions of AI, as an umbrella term for machine learning, autonomous systems, intelligent data mining and smart information systems. According to the European Economic and Social Committee's Opinion:
There is no single accepted and rigid definition of AI. AI is a catch-all term for diverse sets of techniques as well as research agendas, and thus for a large number of sub-fields such as: cognitive computing (algorithms that reasons and understand at a higher (more human) level), machine learning (algorithms that can teach themselves tasks), augmented intelligence (cooperation between human and machine) and AI robotics (AI embedded in robots).

Since the HBP is an EC funded project, we paid special attention to definitions of AI coming from the EU policy sphere. The EU's High-level expert group on AI (AI HLEG) has developed Ethics guidelines for trustworthy AI that define AI or AI systems as follows (p. 36):[3]

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 numerical model, and they can also 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:[4]

- machine learning (of which, deep learning and reinforcement learning are specific examples),
- machine reasoning (which includes planning, scheduling, knowledge representation and reasoning, search, and optimisation), and
- [AI-controlled] robotics (which includes control, perception, sensors and actuators, as well as the integration of all other techniques into cyber-physical systems).

Those who are creating AI systems may have very different levels of expertise. Some may be using black-boxed AI tools, i.e., off-the-shelf commercial packages or open source libraries, where all they need to know is what input is required and what output can be expected. Some may be developing their own AI tools based on existing, well-documented algorithms. Some may be doing advanced research in AI itself, working on the development of new kinds of algorithms.

Within the HBP, there have been several attempts to define the HBP contribution to AI, yet no agreement has so far been reached. The process of identifying AI-related work in the HBP is ongoing. However, at this point we can identify the following areas where HBP work touches upon or is relevant to AI, according to the definition by the HLEG quoted above:

3 The AI HLEG has developed a more detailed AI definition in its document 'A definition of AI: Main capabilities and scientific disciplines'.

4 We should note, in passing, that this definition ascribes to computer software capacities, notably perception and reasoning, that are conventionally restricted to humans – a move that some might consider controversial.

- **AI in Neuroscience:** The HBP develops and uses deep learning neural networks for data analysis and interpretation. The HBP also anticipates that new neuroscientific insights gained from HBP research on recurrent networks may inspire new processing architectures for artificial recurrent networks. In addition, the HBP will provide open access neuroscience data resources, which are expected to contribute training and testing data for AI-driven methods in neuroscience.
- T**ools for the Clinical Translation of Neuroscience:** In the Medical Informatics Platform of the HBP, machine learning methods are being used and developed for the diagnosis of neurological disorders such as epilepsy, Alzheimer's and Parkinson's.
- **Brain Modelling, Simulation and Emulation:** The HBP is developing a suite of brain models of different levels of brain activity (i.e., from synapses, to neurons, to populations, to brain areas, and whole-brain), with varying degrees of detail and biological plausibility. The HBP, therefore, brings together research into biological intelligence and consciousness with research into artificial intelligence, which interact with one another through the Project.
- **Brain-Inspired Hardware:** The HBP is developing analogue and digital neuromorphic computing hardware which provides the substrate for running simulations of the biological neural networks models developed in the Project. The hardware design of these new computing platforms is based on detailed knowledge derived from neuroscience and is therefore a more biologically realistic approach to machine learning in AI than the ones currently in use, which are based on more abstract neural systems.
- **Neurorobotics:** The HBP Neurorobotics Platform aims to better understand the relationship between animal and human brains, bodies, and the environments they are embedded in. This is done by embedding HBP-developed brain models into simulated and physical robotic bodies and environments (Aicardi et al 2020).

The EBRAINS infrastructure, which will be the main output of the SGA3, combines most of the above tools and services and therefore can be used for AI-related research. By collecting and making available neuroscience data, it also provides the data sets that are required for training and testing certain AI techniques.

3. Trust and Trustworthiness

Calls for trustworthy AI have gained a prominent place in current AI ethics debates and in the documents of national and international organisations (Ulnicane et al 2020; Ulnicane et al 2021). These include: the EU Ethics Guidelines for Trustworthy AI (European Commission 2019), the US National Artificial Intelligence Research and Development Strategic Plan (USNSTC 2019), the World Economic Forum's White Paper Digital Transformation of Industries (WEF 2016), the OECD Principles on AI (OECD 2019), IBM's Principles for Trust and Transparency (IBM 2018), and the G20 Statement on Trade and the Digital Economy (2019), amongst many others.
Trust is typically described as a three-place relationship where Agent A (the trustor) places her or himself in a relationship of dependency on Agent B (the trustee) with regard to a particular range of Action C (Mayer et al. 1995). The trustee can be a person, a social group, an institution, or in the case of AI, an algorithm, a set of data or an operating system. Trust inevitably involves vulnerability and risk, because it is not certain whether the trustee will behave as expected, and therefore can harm the trustor's rights, interests or wellbeing (Schoorman et al. 2007).

Most of the existing literature on trust has examined the relationships between humans, human groups as well as individuals, organisations and social institutions. Trust towards non-human entities and complex technology systems such as AI has only been examined more recently (Lankton et al. 2015). A helpful way to approach trust in AI is through the notion of "epistemic trust" as distinct from moral trust. The term refers to people's willingness to accept that the knowledge and information provided by, for example, scientific or diagnostic devices, social media or intelligent systems such as AI is accurate and reliable and can be used as a basis for learning and decision-making (Koenig and Harris 2007; Wiltholt 2013).

This requires evidence. In order to trust AI or any other technology system, it has to prove (or to be proven) that it is trustworthy. As the philosopher Onora O'Neill has said, it is the trustworthiness of an organisation, or a new technology, that matters in the first place; not the level of trust that people have. Trust is the response that follows when trustworthiness has been demonstrated (O'Neill 2013). However, judgements on trustworthiness depend on clear criteria, competence, reliable methods and honesty (O'Neill 2013; Spiegelhalter 2020).

In EU policy documents, references to trust and trustworthiness include calls for trustworthy research, trustworthy AI developers and organisations, trustworthy design principles and algorithms, and the responsible deployment of AI applications. They also underline the importance of the trust of citizens and consumers. Yet, although many AI policy reports and documents speak of trust and trustworthiness, they rarely define and operationalise these terms. The 2019 EU Ethics Guidelines for Trustworthy AI (c) is an exception but it tends to understand trust in moral (not merely epistemic) terms. It uses the following definition of trust: *(1) a set of specific beliefs dealing with benevolence, competence, integrity, and predictability (trusting beliefs); (2) the willingness of one party to depend on another in a risky situation (trusting intention); or (3) the combination of these elements.*

In addition to the somewhat controversial attribution of trustworthiness (rather than reliability) to AI systems (beyond researchers and designers) a problem with this definition is that terms such as benevolence, competence and integrity are vague, generalised categories, which offer little practical guidance on how to implement these "virtues" in the context of actual research, deployment and commercialisation practices (Naughton 2020). Nevertheless, the EU Guidelines state that the development of trustworthy AI requires adherence to the following three components that should be met throughout an AI system's entire life cycle:

(1) It should be lawful, ensuring compliance with all applicable laws and regulations; (2) it should be ethical, demonstrating respect for, and ensure adherence to, ethical principles and values; and (3) it should be robust, both from a technical and social perspective, since, even with good intentions, AI systems can cause unintentional harm. Trustworthy AI concerns not only the trustworthiness of the AI system itself, but also comprises the trustworthiness of all processes and actors that are part of the system's life cycle (European Commission 2019).

According to these guidelines, trustworthy AI is one that respects human dignity and rights, serves and protects our interests, satisfies people's needs, does not infringe on their freedoms, and fosters democratic processes (European Commission 2019). The document acknowledges that trust in AI depends not only on the technology itself, but also on the trustworthiness of the engineers, corporations, advertisers, regulators, and the social, political and technical processes through which AI systems are developed and used.

While this is important, the guidelines ignore other questions. For example, how can trustworthiness be measured? Which criteria, methods and forms of expertise shall be used? Who are the arbiters of trust? Does it make sense to seek trustworthiness (other than just reliability) in algorithms? And who are the trusting parties: governments, scientific organisations, NGOs, technology users, citizens at large – or all together? And how can the involvement of laypeople, citizens and consumers in decisions surrounding AI be improved? These questions remain to be answered.

Moreover, whether an AI system is socially or technically "robust" can only be determined on a case-by-case basis, which requires careful assessment and ongoing monitoring. A possible way forward, at least in the EU, would be the development of a multi-phase evaluation structure for new algorithms and AI applications that is implemented by a regulatory body, similar to the assessment of pharmaceuticals (Spiegelhalter 2020).

3.1 Trust: concerns for citizens and society

Another central question is who will keep the oversight on how, by whom and for which purposes digital data are collected and processed? Current systems of data collection and use are primarily designed as “one-way mirrors”. Citizens and consumers can always be identified, but they never know which kind of information is gathered, how these data are used and for which purposes, and which types of organisations, companies or governments units are involved in these processes. Such practices are fundamentally dangerous, because they compromise anonymity, informed consent and security. It also creates distrust and suspicion.

A central challenge of AI based systems is that they depend on data collection. Data are everywhere and seemingly disparate data can be combined to paint clear pictures of groups and individuals. It is possible that whoever collects, processes and owns the largest amounts of data, is likely to hold power in the future. Such “data inequality” is likely to intensify over the years, when data collection will be even more widespread, and the use and range of data will be more extensive and revolutionised by technological advances (Bitsch, Kotnis, Palsberg et al. 2019). Already citizens report consciously changing behaviours due to uncertainty on who is collecting what information, and for what purposes, on them and their actions. They worry about the kind of data that is being collected on them, what it can be used for, and report an overwhelming sense of loss of control. (Bitsch, Bådum, Campion et al. 2020)

In addition to data collection and processes, the sharing of information is an essential part of creating transparency. However, it is not yet clear what role AI-based technologies will play in the generation and sharing of information in the future. The hard question is: who decides what information is misinformation and what counts as “good” or “valid” information (Bitsch, Kotnis and Palsberg et al. 2019)? Furthermore, who shares the information and for which purpose?

Another challenge is that the Ethics Guidelines for Trustworthy AI focus only on the EU. At a global level, the laws, ethical criteria and definitions of what counts as social and technical ‘robustness’ vary widely. This encourages the development of different types of applications, including the use of AI for terrorism, (cyber)warfare, surveillance and more restrictive policing. These developments can undermine trust in AI, especially if they are used against the interests of citizens, consumers and societies at large (Bitsch, Kotnis and Palsberg et al. 2019).

4. Transparency

Transparency is often seen as a means to promote trust, but is also important in its own right. According to a recent review of AI ethics guidelines, transparency is the most prevalent principle in these documents and applies to domains such as data use, human-AI interaction, the use of algorithms and automated decisions, the purposes of data use and the evaluation of real-world applications of AI systems (Jobin et al. 2019, p.391). In the EU Guidelines, transparency is one of seven key requirements for the realisation of trustworthy AI and 'closely linked with the principle of explicability'. Moreover, it 'encompasses transparency of elements relevant to an AI system: the data, the system and the business models' (European Commission 2019).
But the criterion of transparency is challenging to apply to AI. Unlike with traditional software applications, the behaviour of AI systems is not predictable, in part because advanced AI is designed to learn. Hence the behaviour and results of the technology evolve in time. Considering this, a key question is how transparency can be understood and operationalised, and what the utility of the term is.

In the literature on AI, transparency has been described in various ways (Theodorou, Wortham, and Bryson 2017). Sometimes the term refers to a lack of deception. This implies that the internal workings of AI technology should be open to inspection and evaluation. At other times, the concept refers to a mechanism to report reliability, i.e. the provision of information on the system's tendency to produce errors. At still other times, it has been conceived as a means to communicate unexpected behaviour, to account for the conditions and risks when AI acts differently than expected. Most frequently, however, transparency is used to refer to the need to make decision-making processes accessible to users, so that they can understand and judge how an autonomous system has reached a certain decision.

These definitions are reflected in various policy documents. The UK Engineering and Physical Sciences Research Council's Principles of Robotics (2011), for example, specifically mandate that autonomous robots, which we consider a subset of AI, 'should not be designed in a deceptive way to exploit vulnerable users; instead their machine nature should be transparent'. The concern in the EPSRC's Principles, seems to relate especially to the human tendency to anthropomorphise animals and machines, which in the case of AIs may have a number of practical consequences that must be avoided; for example, that users disclose personal information (as they would to a human companion) when in fact they are disclosing it to a company or robot operators (Salles, Evers, and Farisco 2020).

The EU Guidelines stress in particular the importance of "explicability" or "explainability", which corresponds to the last of the above definitions (accessible forms of decision making). The Guidelines state that: "Explicability is crucial for building and maintaining users' trust in AI systems. This means that processes need to be transparent, the capabilities and purpose of AI systems openly communicated, and decisions – to the extent possible – explainable to those directly and indirectly affected" (p. 13).

The Guidelines recognise that an explanation of why an AI system has generated a particular outcome or decision is not always possible. Yet, explainable artificial intelligence (XAI) is a hot topic nowadays and conferences on XAI are burgeoning. However, mainstream algorithms in AI are based on deep neural networks that are intrinsically non-explainable. These are usually referred to as "black box" algorithms. In those circumstances, the EU Guidelines mention that, 'other explicability measures (e.g., traceability, auditability and transparent communication on system capabilities) may be

required, provided that a system as a whole, respects fundamental rights' (European Commission 2019). A promising strategy is to build models based on decision trees, decision rules or other intrinsically explainable primitives that can approximate the behavior of black-box models locally; in this way, explanations on the black-box decisions can be derived (Blanco-Justicia et al. 2020).

European citizens express a need for some form of human oversight in the application of AI technologies. More specifically, they said that it should be possible to follow decisions not only for the persons operating an AI system, but also for anyone affected by a decision or finding. The citizens were clear in their need for regulation. For example, they suggested the use of certification schemes that could help consumers and other users to judge the trustworthiness and transparency of AI applications. (Bitsch, Bådum and Campion et al. 2020)

4.1 Increasing Transparency, but how?

A problem with the 2019 EU Guidelines (and many other AI policy documents) is that they offer limited guidance on how to achieve transparency in actual practice (Naughton 2020). A response to this moral problem has recently been provided by David Spiegelhalter, the President of the UK Royal Society for Statistics. Spiegelhalter recommends that, 'when confronted by an algorithm, we should expect trustworthy claims both: (1) about the system - what the developers say it can do, and how it has been evaluated, and (2) by the system - what it says about a specific case [i.e., the results of an algorithm]' (Spiegelhalter 2020).

Transparency is a key requirement to establish the trustworthiness of these claims. However, as Spiegelhalter points out, this should not be '"fishbowl" transparency in which huge amounts of data are provided in indigestible form'. Instead, interested parties, including non-experts, should be able to assess the reliability of the claims made about and by AI systems (Spiegelhalter 2020).

In order to increase the transparency of AI, and to assure that claims around AI and the results of specific algorithms are trustworthy, Spiegelhalter proposes a set of seven questions:

1. Is a new algorithm any good when tried in new parts of the real world?
2. Would something simpler and more transparent and robust, be just as good?
3. Could I explain how it works (in general) to anyone who is interested?
4. Could I explain to an individual how it reached its conclusion in their particular case?
5. Does it know when it is on shaky ground, and can acknowledge uncertainty?
6. Do people use it appropriately, with the right level of scepticism?
7. Does it actually help in practice?

These, or similar questions, are a critical step towards the operationalisation and evaluation of transparency and trustworthiness of AI systems, on a case-by-case basis in real-world contexts. They are important also in order to identify and avoid biases. Humans and human society are biased, and therefore the technology will be as well. Examples include gender and racial biases that are reflected in the data sets that algorithms used to train AI systems. In designing AI technologies, designers must consciously decide which biases to introduce (Bitsch, Kotnis and Palsberg et al. 2019). The question is whether AI can deal with biases (make them visible) or whether it undetectably reinforces bias?

The challenge remains, however, to translate the above questions into a set of criteria and methodologies that provide consistent answers. Furthermore, although these questions are important, they ignore other aspects of the development and commercialisation of AI systems, such as marketing practices, potential forms of dual use or misuse, and the unintended effects of AI, such as unemployment and the possible impact of AI on politics and democracy. However, as the next section shows, these issues are of vital importance to citizens and other stakeholders.

5. Broader societal perspectives on AI

5.1.1 Politics, democracy and the potential for abuse

In addition to ethical principles and concerns, AI also invites questions on its implications for policy, political processes, democracy, democratic institutions, the rule of law and societal organisation in general. Trustworthiness, trust and transparency play a central role in the stability and fairness of a society (Bitsch, Kotnis, and Palsberg et al. 2019). For the political system and western democracies, a key challenge is to ensure and enforce transparency. Given the close connection between trust and transparency, AI could provide a challenge to transparency if it becomes difficult to understand how decisions are made and in which areas of decision-making AI technologies are used, and if it is no longer possible (or difficult) to protest decisions that AI systems have made (ibid).

A lack of transparency enables possibilities for abuse and manipulation, for example misinformation in elections or the strategic manipulation of public opinions which can lead to a loss of empathy or feelings of alienation. Abuse and misuse as a major concern and uncertainty. The ways in which people can be influenced and manipulated by AI technologies, was seen as particularly worrisome. Microtargeting techniques, they said, could become increasingly opaque, making it impossible for voters to judge the coherence of policies, the arguments of individual politicians and positions of political parties. Techniques for fostering a division between societal groups, for example by pushing specific messages about one social group to another, which are likely to become increasingly widespread, were also a major concern (Bitsch, Kotnis and Palsberg et al. 2019; Bitsch, Bådum and Campion et al. 2020).

A central question in this regard is who has access to the technology? The robustness of democratic institutions is a key factor for realising the positive potential of AI in our political cultures, which requires effective checks and balances and the fair(er) distribution of power. In addition, many suggest that the deployment of AI should be supported by insights from the behavioural and the social sciences, because AI is used in multiple societal contexts, and impacts depend on social contexts and behaviours. For example, social media already influence how politics is done. The misuse of social media has introduced an increased need for vetting and fact-checking information (Bitsch, Kotnis and Palsberg et al. 2019).

5.1.2 Accountability and transparency

Several factors make it challenging to ensure transparency and accountability in AI platforms and the organisations that run them. According to participants of our workshops, this is well reflected in the operation of current social media and Internet companies. For example, search engines are currently being protected by intellectual property rights (IPR) laws and proprietary rights regulation. This enables them to affect democratic institutions without accountability. Part of the challenge is that present day legal frameworks are tailored to handle traditional societal infrastructures, with a transparent development process and power structure. However, IPR and trade-secrets can be obstacles to enhance transparency, because it allows companies to be opaque about their collection, use, storage and reuse of data. The consequence ends up being that contemporary legal frameworks protect companies, but not users, data subjects or citizens (Bitsch, Kotnis and Palsberg et al. 2019).

6. Relevance for the HBP

The next section highlights some of the areas of activity within the HBP where concerns regarding trust, trustworthiness and transparency may arise and where the use of AI can require heightened levels of awareness and scrutiny.

6.1 AI and Neuroscience

There is a need for transparency regarding the underlying motivation of developing certain tools, i.e., the question whether the aim is to improve neuroscientific research or to improve AI technology. Often in HBP discourse the benefits from AI for neuroscience and from neuroscience to AI are presented as a virtuous circle (Spitzer et al. 2017). This could be questioned as it may raise ethical issues. For example, at which point do the expected benefits to AI technologies, backed by substantial amounts of public and private funding, start steering neuroscientific research agendas? This issue is relevant to all work that claims cross-benefits between neuroscientific and neurotechnological research on the one hand and AI, neuromorphic computing and neurorobotic applications on the other hand. In the HBP this means not only neuroscientific research (e.g., Bellec et al. 2019; Dickscheid's presentation "Building high-resolution models of the human brain with the help of AI and HPC" presentation at the Helmholtz AI kick-off meeting, 05 March 2020 [5] but also neuromorphic computing and neurorobotics (e.g., Bos et al. 2019; Plenary Session 4 at HBP Summit 2020, "Closed-loop AI and robotics workflows: design, test and implement robotic and AI solutions"). In the case of neurorobotics, the issue has been explicitly identified by SP12 and SP10 jointly during the workshop held at TUM in 2018 and the subsequent writing of a joint paper on ethical issues of neurorobotics (Aicardi et al 2020). For the other areas, the issue has not been as clearly expressed, but underlies all the work in the HBP that claims cross-benefits between neuroscientific and neurotechnological research on the one hand and AI, neuromorphic computing and neurorobotic applications on the other hand.

6.2 AI as Tool for Clinical Translation of Neuroscience

The Medical Informatics Platform of the HBP aims to develop, use and provide access in the clinic to clinical research and decision-making tools that incorporate various AI-related techniques such as machine learning, in particular for the diagnosis of neurological disorders such as epilepsy, Alzheimer's and Parkinson's. (e.g., Aerts et al. 2018; Cui et al. 2015; Gamberger et al. 2016; Schirner et al. 2018; Shen et al. 2019; Stefanovski et al. 2019; Venetis et al. 2015; Zufferey et al. 2017).[6] What, then, is at stake in the translation of these methods from the laboratory to the clinical setting?

6.2.1 Clinicians and Patients

For the effective clinical translation of AI-based diagnostic technologies, trust between researchers, clinicians, patients, and regulators is essential. This is less a matter of ensuring trust among "the

5 Streaming of presentation available at https://www.helmholtz.ai/themenmenue/you-helmholtz-ai/events/helmholtz-ai-kick-off-meeting/index.html, from timestamp 3:39:00

6 https://www.humanbrainproject.eu/en/medicine/ and https://www.humanbrainproject.eu/en/medicine/the-virtual-brain/, consulted 16/03/2020

general public" than of establishing and maintaining trust in clinicians, who are in many ways the gatekeepers to patients' trust. Direct interpersonal "human relationships" are instrumental for patients in trust building and trust-maintaining processes towards healthcare and medical institutions (Datta, 2018, pp4). As a result, patients' acceptance of scientific innovation reflects in large part the confidence that the administering clinician has in the use of an innovative technology in clinical settings . Here trust and trustworthiness are fundamentally relational; while belief in the epistemic validity of the diagnosis is crucial, patients' epistemic trust is most often dependent on their trust in their clinician, and his or her trust in the diagnostic technology and procedures that are used (Datta Burton et al, 2021a; Datta, 2018).

6.2.2 Inscrutable algorithms

This raises particular difficulties where algorithmic methods are used to generate diagnoses. Unsupervised learning algorithms, which aim to discover inherent structures in data without using pre-existing categories, are notoriously inscrutable even to their designers (Bender et al, 2021). In this context, the idea of using such unsupervised machine learning to discover "brain signatures" of neurological conditions or "biomarkers" of mental disorders appears problematic (Datta Burton et al, 2021a). Proponents of these methods suggest that these might not only increase the speed and accuracy of differential diagnosis of individual patients in the clinic, and hence of the accuracy of treatment decisions, but also might bring about a complete revision of the classification of mental and neurological disorders. However, such strategies are likely to fail to gain acceptability in clinical situations, unless the issues of transparency and "explainability" are addressed. Where candidate brain signatures or biomarkers result from AI-related techniques, such as machine learning and multivariate statistical analysis, they must be able to demonstrate an understandable chain of evidence and reasoning, especially if they are to be used for clinical purposes (ibid).

Another question is whether AI based machines will have the ability to interpret patient data correctly, and also whether they could be sufficiently trained in order to keep up with medical developments over time, and if AI will work in all areas of medicine. It could be that AI-based systems would work well as a diagnostic tool, but that they cannot replace physicians in understanding patients' specific needs. Moreover, AI-based diagnostics will only be as good as the data it relies on, and so poor data will lead to poor diagnosis and bad health advice. Finally, a predominant focus on the funding of AI solutions would come at the expense of other medical solutions, including low-tech solutions (such as interpersonal communication), even though these provide better or equally preferable results. (Bitsch, Kotnis and Palsberg, 2019).

6.2.3 Styles of reasoning

Research carried out by the HBP's Foresight Lab has found that there are frequent clashes between the forms of reasoning employed by clinicians in reaching a diagnosis for a particular patient, and those employed by modellers. These are not unique to modelling. We find them in other areas of medicine, for example in debates about evidence-based medicine where clinical reasoning (which is about integration of evidence, experience and knowledge of a particular patient and his or her history and circumstances) comes into conflict with protocols derived from probability based evidence derived from meta-analyses of randomised control trials in large populations.

We can point to two key issues in clashes between clinical diagnostic reasoning and the application to a particular patient of a diagnosis derived from the use of algorithms applied to large quantities of patient data concerning symptomatology and outcomes compiled from a range of sources. First, clinicians do not have the required training to critically analyse the results of new data analysis tools (Datta Burton et al, 2021a). They often distrust the results because of a lack of understanding of these models and the ways algorithms reach decisions. While unsupervised learning has been successful in diagnosis, especially in relation to the analysis of medical images (Lundervold and Lundervold, 2019), models integrating such techniques are far less interpretable than other machine learning methods. Second, modellers often perceive clinical reasoning as itself subjective and lacking in transparency, rather than the consequence of their own lack of understanding of clinical reasoning. While there is a trend towards the quantification of healthcare in recent years, and the view that clinicians' assessments are 'subjective' and 'biased', the experiential and tacit knowledge of clinicians is foundational to clinical work and to the relationships of trust between patients and physicians (Schwartz and Elstein, 2008; Datta Burton et al, 2021a).

6.2.4 Challenges for trustworthiness

Transparency is also not sufficient to ensure trustworthiness. Research with clinicians has shown that more transparency would not necessarily make the technology trustworthy to clinicians.[7] While many clinicians are pleased when model-based diagnoses confirm the diagnoses made on the basis of their own reasoning about a particular patient, they are likely to reject conclusions derived from AI technologies where these disagree with those derived from their own reasoning. Even where they understand the processes that have led an algorithm to a diagnosis, there are many reasons why they trust their own diagnoses, including, for example, concerns about the reliability of the data on which the algorithm has worked, or its applicability to this particular patient with her life history, co-morbidities and life situation. Indeed, we can see some of these concerns in relation to other diagnostic technologies, such as those based on genetic analysis and even those based on the analysis of blood or urine samples – clinicians tend to trust the results only from their own labs or those that they habitually use, and to distrust results coming from unfamiliar or distant institutions (Cousin-Frankel, 2019).

6.3 Brain Modelling, Simulation and Emulation

One aspect of brain modelling, simulation and emulation that is of high ethical and societal interest is the conclusion it may allow drawing with regards to consciousness. Detecting and proving consciousness in other subjects/agents, particularly in speechless and/or not fully/differently behaving subjects (e.g., animals and AI devices), are notoriously very challenging and yet increasingly urgent issues.

Together with a member of SP3, some researchers from SP12 have recently suggested a list of operational indicators and criteria of consciousness in order to facilitate the recognition/attribution of consciousness challenging cases like AI artifacts (Pennartz et al. 2019). The conceptual premise of the suggested criteria and indicators is the view of consciousness as a modelling activity by the brain (i.e., a multimodal situational survey) that enables the subject to survive in his/her environment through the satisfaction of his/her needs and the achievement of his goals (Pennartz 2015).

With reference to theoretically derived measures of consciousness (e.g., the integrated information

7 https://www.kcl.ac.uk/research/the-human-brain-project-the-foresight-lab

of a system), we argue that their validity should not be assessed on the basis of a single quantifiable measure, but requires cross-examination across multiple pieces of evidence, including the criteria and indicators we propose. Current intelligent machines, including deep learning neural networks (DLNNs) and agile/autonomous robots, are not indicated to be conscious yet. Instead of assessing machine consciousness by a brief Turing-type of test, evidence for it may gradually accumulate when we study machines ethologically, i.e., diachronically and in interactive contexts.

HBP is developing workflows and modelling strategies for modelling the brain at different scales, to then put together or bridge the results obtained from the different levels of organization (Amunts et al. 2016). More specifically, the HBP employs a data-driven strategy of "components models": the research is intended to model a phenomenon at a certain scale, modelling all its different components and then aggregating them to determine what happens at the higher level.

We suggest that this approach might be potentially useful for simulating consciousness, if this is operationalised in terms of neuronal correlates. Yet, in collaboration with Jeanette H. Kotaleski who works on mathematical modelling and simulation, the following challenges have been identified (Farisco et al. 2018):

- The brain is far more than an input-output machine. It can be described as a network with hidden internal layers, and its activity between the input and output layer (which seems critical for consciousness), often cannot be precisely reconstructed mathematically.
- At the local level, the properties of the brain components are relatively changeable depending on their reciprocal interaction. Modelling a single component is not sufficient to get a reliable prediction of its behaviour.
- At the global level, the brain exhibits properties and functions that supervene its different, particular components.
- In its basic form, as seen above, consciousness has been proposed to be a simulation-based interaction with the external environment, so that to simulate the conscious brain means to simulate a simulating system, resulting in a kind of second order simulation (or metasimulation).

The increasing interaction between neuroscience and AI, especially deep learning and data mining technologies, promises to offer new tools for assessing the above mentioned challenges.

6.4 Translation Pathways

Much of the promise of work within the HBP is that it will not merely lead to more and better knowledge of the human brain, but that it will also lead to potentially commercial outputs. The HBP is developing formal relationships with a number of commercial organisations and seeking to establish "translation pathways" to take the findings of HBP researchers into product development pipelines, including the commercial exploitation of emerging applications in AI and robotics. While it is challenging to introduce policies of responsible research and innovation (RRI) into publicly funded research programmes and laboratories, it is even more challenging to introduce them into organisations whose rationale is profit and shareholder value.

With the development of new techniques, technology concepts, datasets and infrastructures, the HBP plays a central role in facilitating the translation of research into commercial applications. However, product development, testing, marketing and commercialisation is largely undertaken by the private sector. AI and AI-controlled robotics that are used for political, security, intelligence

and military (PSIM) purposes, which (indirectly) result from publicly funded research, is also mainly advanced by private companies (Aicardi et al. 2019, pp16-17). This means the majority of AI and robotics applications that emerge from the HBP (and other academic research) will be developed by companies, with little oversight by the public.

Advances in the HBP have the potential to inform the use of AI and robotics across a wide range of social domains, which include manufacturing and transport, household and health care, as well as policing, surveillance, warfare and others (Rose, Aicardi and Reinsborough 2016; Aicardi et al. 2019). The aim for most AI-based applications that are currently under development is to build assistive capacities that "augment instead of replac[ing] human-led decision-making" (Datta Burton et al, 2021b, pp10). Future applications in these domains will transform (and in some instances) disrupt existing practices and generate new social, ethical, economic, political and human rights issues, some of which will become apparent only over time. These issues affect perceptions of trust and trustworthiness not only of emerging AI and robotics applications, but equally important, of the firms and organisations that develop and use them (Datta Burton, 2020).

The transformative potential of AI and robotics requires a clear commitment to responsibly develop commercial products and services, including at an international level. In the next paragraphs, we refer to two areas of particular concern.

6.4.1 Commercialisation of "mundane" AI and robotics applications

The HBP "Opinion on 'Responsible Dual Use'" has already suggested principles to distinguish between "responsible" and "irresponsible" systems of research and technology development (Aicardi et al. 2018). However, also more "mundane" forms of AI and robotics to which research within the HBP will contribute (or that will be enabled by HBP research in the longer term), can have wide-ranging consequences for societies. These can include (among others), applications that seek to optimise workflows, services or information systems (in households, firms, health care institutions, traffic systems, government units, etcetera). The fact that these technologies promise new solutions, increased efficiency and lower costs, does not mean they are inevitably beneficial or without risks and unintended disruptive effects.

Mundane forms of AI and robotics are expected to increase on the global market over the next years and decades (Chui and Manyika 2018). Government intervention at the downstream level of these applications is and will probably remain minimal, similar to other digital products and services. EU legislation such as the European Consumer Law seeks to protect consumer rights and to guarantee that commercial products are safe and ensure consumer confidence (Valant 2015). Other EU rules aim to control the international trade in dual-use items and to prevent that Horizon2020 funded research can be utilised for military applications and/or misused for unethical purposes (Aicardi et al., 2018).

In addition to a concern with consumer safety as well as dual use and misuse, it is important to engage more systematically with the potential impacts of these applications, including a concern with the broader societal and environmental implications of emerging AI and robotics applications, short-to-long term. While RRI policies seek to cover these aspects in publicly funded research, in the context of downstream translation, commercial exploitation and industry partnerships, many of these issues disappear or are inadequately considered. This undermines transparency and inhibits trust between public stakeholders and firms. It can also generate a loss of trust in science and the oversight mechanisms of governments.

What is needed, is the development of a set of criteria and a framework that aims to identify the

specific ethical and social issues that arise in the context of downstream translation on a case-by-case basis, and that can be applied in academic-corporate partnerships and by private sector innovators.

6.4.2 Transfer of AI and robotics applications from the EU to other world regions

RRI evaluation of new technology applications in EU countries does not automatically produce valid results for other parts of the world. The transfer of AI and robotics technologies between global regions takes place across significant socio-economic, political and cultural differences. As a result, the adoption and use of technological inventions in different social contexts, produces diverse outcomes. A technology that can be beneficial in one social context, can create disruptive and problematic effects in another environment. Furthermore, the transfer of technologies and products developed in the EU to other parts of the world, takes place against a broader context of global inequalities, which requires specific considerations.

Existing literature in development and technology studies has shown that the international transfer of technological products and solutions can create unexpected social and cultural consequences, which can transform local communities in problematic ways (Allen and Thomas 2000). Anticipation of these effects and the ways in which new technologies interact with different social and natural environments, is a fundamental requirement to achieve responsible commercialisation of AI and robotics products across borders.

A coherent approach and clear criteria are required to realise this. What is needed, is the development of new methodologies and interdisciplinary research that allow to examine the social, cultural, gendered, political and environmental effects of the international transfer of AI and robotics applications (cf. Gardner and Lewis 1998). Such analyses must involve the investigation of the effects of technological change at the micro level, depending on the technology and social domain in which it is applied; for example, at the level of the family, communities, schools, firms, trade networks, local government bodies, farms and regional agricultural systems, as well as sourcing and value chains for localised production systems, among others.

7. Conclusion

This Opinion draws from the broader discourse on social and ethical issues of AI. It highlights the concepts of trustworthiness and transparency as two key concerns that researchers and practitioners working with AI should be aware of. It identifies several areas of work within the HBP that have relevance to AI. It spells out some of the concerns that these HBP activities may raise.
The Opinion does not aim to provide a comprehensive answer to all of these issues. It shows that there are conceptual issues related to AI and the reasonable use of the term within the HBP.

The main outcome of the Opinion is to provide guidance for the way in which the HBP will engage with AI ethics-related questions during the SGA3 phase. The following recommendations should therefore be read as an attempt to structure relevant work accordingly and allow the relevant groups and individuals to proceed in a way that will foster trust and transparency across all AI-related work.

8. Recommendations for the Human Brain Project

Recommendation One: Provide an overview of AI-related activities in the HBP

In order to create a basis for further ethical, social and RRI reflection, we recommend that the HBP undertakes a comprehensive overview of ongoing and emerging AI activities within the project.

- The overview builds on previous HBP work on AI
- The overview will be a basis for developing an HBP approach to AI which positions itself in relation to the EU policies on AI.
- Such an overview needs to pay attention to contextual factors that may be playing a role in AI research (such as political agendas, commercialization, personal interests, and aspects such as gender, class, age, and race, among others).

Recommendation Two: Involve clinicians and other users and beneficiaries

We recommend that the HBP identifies those who are envisaged as users and beneficiaries of AI based technologies (e.g., clinicians, patients, citizens, public services, interest organisations, etc.) and involves them in the formulation of research problems and in the initial design of research projects. This would contribute to raising awareness of the needs and preferences of different groups of users.

- For clinical applications, this is a crucial step towards gaining the trust that is necessary for clinical translation. Researchers and engineers in the HBP need to understand and acknowledge the tacit, experiential knowledge involved in clinical reasoning, which has made it difficult in the past to incorporate algorithmic tools into the clinic.

Recommendation Three: Include Ethics and RRI in the HBP's AI education programme

We recommend that the Human Brain Project develops an educational programme on AI for PhD students, early and mid-career researchers, as well as corporate partners, which specifically addresses AI Ethics and Responsible AI.

- Education could help researchers understand the possible societal and behavioural contexts of implementation for AI based applications and systems.
- A better understanding of these implications can help to anticipate undesirable effects at an early stage and lead to better technologies and solutions.
- Education should include training with scholars from the humanities and social sciences to address societal and ethical issues of AI-based and AI-developing research from a more interdisciplinary perspective.

Recommendation Four: Focus on the ethical and societal implications of commercialization

We recommend that the HBP and its partner projects undertake further work on the ethical, social and RRI dimensions of the translation of research into commercial AI and robotics products and services.

- This line of work should include a concern with the societal, political, economic and environmental consequences that are caused by disruptions of existing systems of production, social organisation, administration and political control.

Recommendation Five: Examine the effects of the international transfer of AI technologies from the EU to other world regions

We recommend that the HBP and its corporate partners give careful consideration to the possible implications of the international transfer of AI and robotics applications developed in the EU to other world regions.

- This should involve a concern with the ways in which individual products and services interact with and transform different social, digital and natural environments at both micro and macro levels, including the lives and well-being of different groups of technology users and citizens, in all their diversity.
- At the same time, decisions to enable or constrain international transfer should consider the international competitiveness of AI technologies developed in the EU.

Recommendation Six: Develop new methods to integrate RRI in the HBP's strategy to facilitate commercial exploitation of project findings and inventions

We recommend the development of new methods and ethical criteria to integrate RRI evaluation in the HBP's strategy to enable the exploitation and commercial use of emerging AI and (neuro)robotics applications.

- Such methodologies must be designed for researchers in the HBP and the private sector, including for firms and organisations involved in the development and distribution of these technologies
- They must be usable in public-private sector partnerships and consider the well-being and situation of key stakeholders, in particular those whose lives will be influenced by emerging AI and robotics products (e.g. employees, citizens, consumers, patients, etc.)
- They must be context-specific and consider RRI policies enshrined by H2020 and other relevant policies, including the EU's forthcoming rules on the trade of dual use items.

9. References

- Aerts, H., Schirner, M., Jeurissen, B., Van Roost, D., Achten, E., Ritter, P. and Marinazzo, D. Modeling Brain Dynamics in Brain Tumor Patients Using the Virtual Brain. eNeuro 28 May 2018, 5 (3) ENEURO.0083-18.2018; DOI: https://doi.org/10.1523/ENEURO.0083-18.2018
- Aicardi, C., S.Akintoye, B.T.Fothergill, M.Guerrero, G.Klinker, W.Knight, L.Klüver, Y.Morel, F.O.Morin, B.C.Stahl and I.Ulnicane. 2020. Ethical and Social Aspects of Neurorobotics. *Science and Engineering Ethics* 26(5): 2533–2546. https://doi.org/10.1007/s11948-020-00248-8
- Aicardi, CA, Datta Burton, S., Mahfoud, T, Rose, N. 2019. Machine Learning and Big Data for Neuro-Diagnostics: Opportunities and Challenges for Clinical Translation. A briefing report for the Human Brain Project. https://kclpure.kcl.ac.uk/portal/files/137218687/HBPForesightLab_2019_NeuroDiagnostics_BriefingReport_PUBLIC.pdf.
- Aicardi, C. et al. 2018. Opinion on 'Responsible Dual Use'. Human Brain Project. Online resource. URL: https://sos-ch-dk-2.exo.io/public-website-production/filer_public/77/61/7761fdcd-b0a0-40a2-a6bd-904d68d52b87/opinion_dual_use_hbp_ethicssociety.pdf.
- Allen, T. and Thomas, A. 2000. Poverty and Development in the 21st Century.Oxford: Oxford University Press.
- Amunts, K, Ebell, C, Muller, J, Telefont, M, Knoll, A, Lippert, T. 2016. "The Human Brain Project: Creating a European Research Infrastructure to Decode the Human Brain." Neuron 92 (3):574-581. doi: 10.1016/j.neuron.2016.10.046.
- Baier, A. 1986. "Trust and Antitrust." Ethics 96 (2):231-260.
- Baylé, Marion. 2018. "Ethical dilemmas of AI: fairness, transparency, collaboration, trust, accountability & morality."
- Bellec, G., Scherr, F., Subramoney, A., Hajek, E., Salaj, D., Legenstein, R., and Maass, W. 2019, pre-print, in review). "A solution to the learning dilemma for recurrent networks of spiking neurons." bioRxiv/org/10.1101/738385v3, December 2019. https://www.biorxiv.org/content/10.1101/738385v3.full.pdf
- Bender, E. M., T. Gebru, and S. McMillan-Major, A., Mitchell. 2021. On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? Proceedings of the ACM/IEEE Joint Conference on Digital Libraries. Vol. 1. Association for Computing Machinery. https://doi.org/10.1145/3442188.3445922.
- Bitsch, L., Kotnis, S. R., Palsberg, A., Bådum N. B., Jørgensen, M. L. and Klüver L. 2019. AI 360 Copenhagen: Report from the workshop. [online] https://sos-ch-dk-2.exo.io/public-website-production/filer_public/a4/f2/a4f2aabd-6821-4a8d-b082-5070e2797b27/ai360_humanbrainproject_recommendations_report_final.pdf
- Blanco-Justicia, A., Domingo-Ferrer, J., Martínez, S., and Sánchez, D. "Machine learning explainability via microaggregation and shallow decision trees." Knowledge-

Based Systems (to appear).

- Boden, M., J. Bryson, D. Caldwell, K. Dautenhahn, L. Edwards, S. Kember, P. Newman, V. Parry, G. Pegman, T. Rodden, T. Sorell, Wallis, B. Whitby, and A. Winfield. 2011. Principles of robotics.
- Bos, J.J., Vinck, M., Marchesi, P., Keestra, A., van Mourik-Donga, L.A., Jackson, J.C., Verschure, P.F.M.J., Pennartz, C.M.A., 2019. Multiplexing of Information about Self and Others in Hippocampal Ensembles. Cell Reports 29, 3859-3871.e6. https://doi.org/10.1016/j.celrep.2019.11.057
- Chui, M. and Manyika, J. 2018. The real-world potential and limitations of artificial intelligence. McKinsey Quarterly. Online resource. URL:https://www.mckinsey.com/featured-insights/artificial-intelligence/the-real-world-potential-and-limitations-of-artificial-intelligence
- Coeckelbergh, M. 2011. “Can we trust robots?” Ethics and Information Technology 14 (1):53-60.
- Coleman, J.S. 1990. Foundations of social theory. Cambridge: MA: Harvard University Press.
- Cousin-Frankel, J. 2019. “Medicine contends with how to use artificial intelligence.” Science 364(6446): 1119-1120.
- Cui J, Zufferey V, Kherif F. In-vivo brain neuroimaging provides a gateway for integrating biological and clinical biomarkers of Alzheimer’s disease. Curr Opin Neurol 2015;28:351–357. DOI 10.1097/WCO.0000000000000225.
- Datta, S. 2018. Emerging dynamics of evidence and trust in online user-to-user engagement: the case of ‘unproven’stem cell therapies. Critical Public Health, 28(3), 352-362.
- Burton, S. D. 2020. Responsible use of exoskeletons and exosuits: Ensuring domestic security in a European context. Paladyn, Journal of Behavioral Robotics, 11(1), 370-378.
- Datta Burton, S., Mahfoud, T., Aicardi, C., & Rose, N. 2021. Clinical translation of computational brain models: understanding the salience of trust in clinician–researcher relationships. Interdisciplinary Science Reviews, 46(1-2), 138-157.
- Burton, S. D., Kieslich, K., Paul, K. T., Samuel, G., & Prainsack, B. 2021. Rethinking value construction in biomedicine and healthcare. BioSocieties, 1-24.
- De Melo Martin, I. Intemann K. 2018. The Fight Against Doubt. New York: NY. Oxford University Press.
- Engineering and Physical Sciences Research Council. 2011. EPSRC Principles of Robotics. Online resource. URL: https://epsrc.ukri.org/research/ourportfolio/themes/engineering/activities/principlesofrobotics
- European Commission (2019) Ethics Guidelines for Trustworthy AI, Independent High-Level Expert Group on Artificial Intelligence set up by the European Commission.
- European Commission, 2020. White Paper on Artificial Intelligence: a European approach to excellence and trust (White paper No. COM(2020) 65 final). Brussels.

- Farisco, M, Kotaleski, JH, Evers, K. 2018. “Large-Scale Brain Simulation and Disorders of Consciousness. Mapping Technical and Conceptual Issues”. Front. Psychol. 9:585. doi: 10.3389/fpsyg.2018.00585.
- Farisco, M. Evers, K., Salles, A. 2020 Towards establishing criteria for the ethical analysis of Artificial Intelligence. Science and Engineering Ethics 26: 2413-2425G20 Ministerial Statement on Trade and Digital Economy, 8 and 9 June 2019 in Tsukuba, Japan.
- Gardner, K. and Lewis, D. 1996. Anthropology, Development and the Post-Modern Challenge. London: Pluto Press
- Gamberger D, Ženko B, Mitelpunkt A, Shachar N, Lavrač N. Clusters of male and female Alzheimer’s disease patients in the Alzheimer’s Disease Neuroimaging Initiative (ADNI) database. Brain Inform 2016;3:169–179. DOI: 10.1007/s40708-016-0035-5
- Gillespie, T. (2016) Algorithmically recognizable: Santorum’s Google problem, and Google’s Santorum problem. Information, Communication & Society, 20, 1–18. Available at: https://doi.org/10.1080/1369118X.2016.1199721
- Gunning, D. 2019. Explainable Artificial Intelligence (XAI). AI Magazine 40:2.
- Hardin, Russell. 1996. “Trustworthiness.” Ethics 107:26-42.
- Holton, R. 1994. “Deciding to trust, coming to believe.” Australasian Journal of Philosophy 72 (1):63-76.
- IBM. 2018. IBM’s Principles for Trust and Transparency. Online document. URL: https://www.ibm.com/blogs/policy/trust-principles/
- Jobin, A., M. Ienca, and E. Vayena. 2019. “The Global Landscape of AI Ethics.” Nature Machine Intelligence 1:389-399.
- Jones, K. 1996. “Trust as an Affective Attitude.” Ethics 107:4-25.Jones, K. 1999. “Second-Hand Moral Knowledge.” The Journal of Philosophy 96 (2):55-78.
- Jones, K. 2012. “Trustworthiness.” Ethics 123:61-85.
- Kaminski M, Rueben M, Smart WD, Grimm CM. 2017. Averting Robot Eyes. 76 Md.L.Rev. 983.
- Kiran, Asle H., and Peter-Paul Verbeek. 2010. “Trusting Our Selves to Technology.” Know Techn Pol 23:409-427.
- Koenig, M. A., & Harris, P. L. 2007. The basis of epistemic trust: Reliable testimony or reliable sources?. Episteme, 4(3), 264-284.
- Lankton, N. K., McKnight, D. H., & Tripp, J. 2015. Technology, humanness, and trust: Rethinking trust in technology. Journal of the Association for Information Systems, 16(10), 1.
- Lundervold, A. S., & Lundervold, A. 2019. ‘An overview of deep learning in medical imaging focusing on MRI’, Zeitschrift für Medizinische Physik, 29(2), 102-127.
- McLeod, C. 2000. “Our Attitude Towards the Motivation of Those We Trust.” The Southern Journal of Philosophy 38:465-479.

- Mayer, R. C., Davis, J. H., & Schoorman, F. D. (1995). An integrative model of organizational trust. Academy of management review, 20(3), 709-734.
- Nickel, P. J. 2009. “Trust, Staking, and Expectations.” Journal for the theory of Social Behaviour 39 (3):345-362.
- Nickel, P. J., Maarten Franseen, and Peter. Kroes. 2010. “Can We Make Sense of the Notion of Trustworthy Technology?” Know Techn Pol 23:429-444.
- OECD (2019) Recommendation of the Council on Artificial Intelligence, OECD/LEGAL/0449.
- O’Neill, O. 2002. Autonomy and Trust in Bioethics. Cambridge, UK: Cambridge University Press.
- O’Neill, O. 2018. “Linking Trust to Trustworthiness.” International Journal of Philosophical Studies 26 (2):293-300.
- Pennartz, CM. 2015. “The Brain´s Representational Power. On Consciousness and the Integration of Modalities”. Cambridge, MA: MIT Press.
- Pennartz, CM, Farisco, M, Evers, K. 2019. “Indicators and Criteria of Consciousness in Animals and Intelligent Machines: An Inside-Out Approach.” Front Syst Neurosci 13:25.
- Price, WN. 2017. “Regulating black-box medicine.” Michigan Law Review 116(1): 421–74.
- Proudfoot, D. 2011. “Anthropomorphism and AI: Turing’s much misunderstood imitation game.” Artificial Intelligence 175:950-957.
- Rodriguez, J. . 2018. “Towards AI Transparency: Four Pillars Required to Build Trust in Artificial Intelligence Systems.” Medium.
- Rose, N. S., Aicardi, C., & Reinsborough, M. T. (2016, Mar 31). Foresight report on future computing and robotics: A Report from the HBP Foresight Lab.
- Ryan, M. et. al. 2019. Report on Ethical Tensions and Social Impacts. SHERPA Project. Online resource. URL: https://doi.org/10.21253/DMU.8397134.v2
- Salles, A., Evers, K. Farisco, M. (2020) Anthropomorphism in AI. AJOB Neuroscience,11:2,88-95,DOI:10.1080/21507740.2020.1740350
- Schirner, M., McIntosh, AR.; Jirsa, V., Deco, G., and Ritter, P. (2018) Inferring multi-scale neural mechanisms with brain network modelling. eLife. URL: https://elifesciences.org/articles/28927.
- Schoorman, F. D., Mayer, R. C., & Davis, J. H. 2007. An integrative model of organizational trust: Past, present, and future, Academy of Management Review, 32(2), 344-354.
- Schwartz, A., & Elstein, A. S. (2008). Clinical reasoning in medicine. Clinical reasoning in the health professions, 3, 223-234.
- Shen, K., Bezgin, G., Schirner, M., Ritter, P., Everling, S., McIntosh, AR. (2019). A macaque connectome for large-scale network simulations in TheVirtualBrain. Sci Data 6, 123. https://doi.org/10.1038/s41597-019-0129-z

- Simon, J. (2010). The entanglement of trust and knowledge on the Web. Ethics and Information Technology, 12(4), 343-355.
- Spitzer, H., Amunts, K., Harmeling, S., Dickscheid, T., 2017. Parcellation of visual cortex on high-resolution histological brain sections using convolutional neural networks, in: 2017 IEEE 14th International Symposium on Biomedical Imaging (ISBI 2017). Presented at the 2017 IEEE 14th International Symposium on Biomedical Imaging (ISBI 2017), pp. 920–923. https://doi.org/10.1109/ISBI.2017.7950666
- Stefanovski, L. Triebkorn, P. Spiegler, A Diaz-Cortes, M.A. Solodkin, A. Jirsa, V. McIntosh, A.R. and Ritter, P. for the Alzheimer's Disease Neuroimaging Initiative (2019). Linking molecular pathways and large-scale computational modeling to assess candidate disease mechanisms and pharmacodynamics in Alzheimer's disease. Frontiers Computational Neuroscience. https://doi.org/10.3389/fncom.2019.00054
- Taddeo, M., and L. Floridi, 2011. "The case for e-trust," Ethics and Information Technology, 13: 1–3.
- Theodorou, A., R. H. Wortham, and J.J. Bryson. 2017. "Designing and implementing transparency for real time inspection of autonomous robots." Connection Science 29 (3):230-241.
- Ulnicane, I., W. Knight, T. Leach, B. C. Stahl and W.-G. Wanjiku. 2020. Framing governance for a contested emerging technology: insights from AI policy, *Policy and Society*, https://doi.org/10.1080/14494035.2020.1855800
- Ulnicane, I., D. O. Eke, W. Knight, G. Ogoh and B. C. Stahl. 2021. Good governance as a response to discontents? Déjà vu, or lessons for AI from other emerging technologies. *Interdisciplinary Science Reviews* 46(1) https://doi.org/10.1080/03080188.2020.1840220
- US National Science and Technology Council 2019 US National Artificial Intelligence Research and Development Strategic Plan. Online Resource. URL: https://www.nitrd.gov/pubs/National-AI-RD-Strategy-2019.pdf
- Valant. J. 2015. Consumer Protection in the EU: Policy Overview. European Parliament Research Service. Online Resource. URL: https://www.europarl.europa.eu/RegData/etudes/IDAN/2015/565904/EPRS_IDA(2015)565904_EN.pdf
- Vayena, E., A. Blasimme, and I.G. Cohen (2018). Machine learning in medicine: Addressing ethical challenges. PLOS Medicine 15 (11).
- Venetis T, Ailamaki A, Heinis T, Karpathiotakis M, Kherif F, Mitelpunkt A et al. Towards the Identification of Disease Signatures. In: Brain Informatics and Health, pp. 145-155. Cham, Springer International Publishing, 2015.
- Whittaker, M., K. Crawford, R. Dobbe, G. Fried, E. Kaziunas, V. Mathur, S. Myers West, R. Richardson, J. Schultz, and O. Schwartz. 2018. AI Now Report 2018.
- Wilholt, T. (2013). Epistemic trust in science. The British Journal for the Philosophy of Science, 64(2), 233-253.
- Winfield, A.F.T., and M. Jirotka. 2018. "Ethical governance is essential to building trust
- in robotics and artificial intelligence systems." Phil. Trans. R. Soc. A 376.
- World Economic Forum. 2016. WEF White Paper. Digital Transformation of Industries. Online Resource. URL: https://www.accenture.com/_acnmedia/accenture/

conversion-assets/wef/pdf/accenture-digital-enterprise.pdf

- Wortham, R. H., A. Theodorou, and J.J. Bryson. 2017. “Robot Transparency, Trust and Utility.” Connection Science.
- Wright, S. 2010. “Trust and Trustworthiness.” Philosophia 38:615-627.
- Zufferey V, Donati A, Popp J, Meuli R, Rossier J, Frackowiak R et al. Neuroticism, depression, and anxiety traits exacerbate the state of cognitive impairment and hippocampal vulnerability to Alzheimer’s disease. Alzheimers Dement (Amst) 2017:7:107-114.

Human Brain Project

Ethics & Society

Co-funded by
the European Union