TechEthos D2.2: Identification and specification of potential ethical issues and impacts and analysis of ethical issues of digital extended reality, neurotechnologies, and climate engineering

Laurynas Adomaitis, Alexei Grinbaum, Dominic Lenzi

TECHETHOS

FUTURE ○ TECHNOLOGY ○ ETHICS

Identification and specification of potential ethical issues and impacts and analysis of ethical issues

D2.2

TechEthos receives funding from the EU H2020 research and innovation programme under Grant Agreement No 101006249. This deliverable and its contents reflect only the authors' view. The Research Executive Agency and the European Commission are not responsible for any use that may be made of the information contained herein.

D2.2 Identification and specification of potential ethical issues and impacts and analysis of ethical issues			
Work Package	WP 2		
WP lead partner	DMU		
Task	Task 2.3		
Task lead partner	CEA		
Lead authors	Laurynas Adomaitis (CEA), Alexei Grinbaum (CEA), Dominic Lenzi (UT)		
Contributing authors	Nitika Bhalla (DMU), Eva Buchinger (AIT), Sara Cannizzaro (DMU), Wenzel Menhert (AIT), Anaïs Resseguier (TRI), Kathleen Richardson (DMU), Lisa Tambornino (EUREC), Steven Umbrello (TUD)		
Acknowledgments	This report has benefited from the feedback from the ADIM board members of TechEthos as well as other experts listed in section 1.1.1.3.		
Due date	30/06/2022		
Submitted date	30/06/2022		
Version number	3	Status	FINAL

How to cite

Adomaitis, L., Grinbaum A., Lenzi, D. (June 2022) *TechEthos D2.2: Identification and specification of potential ethical issues and impacts and analysis of ethical issues of digital extended reality, neurotechnologies, and climate engineering.*

Project Information

Project number	101006249
Start date	01/01/2021
Duration	36 months
Call identifier	H2020-SwafS-2020-1
Topic	SwafS-29-2020 – The ethics of technologies with high socio-economic impact
Instrument	CSA

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No.101006249.

Dissemination Level	
PU: Public	☒

Quality Control	
Reviewed by:	Review date:
Nicole Santiago (TRI)	28/06/2022
Michael Bernstein (AIT)	28/06/2022

Revision history

Version	Date	Description
0.1	10/12/2021	First plan and outline
0.2	11/02/2022	Final plan and outline
0.5	15/04/2022	First incomplete draft
1	20/05/2022	Complete draft
2	24/06/2022	Pre-final draft
3	30/06/2022	Final version

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No.101006249.

Table of contents

This project has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement No.101006249.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No.101006249.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No.101006249.

Executive summary

This Deliverable 2.2 is produced as part of Work Package 2 of the Horizon-2020 project TechEthos. Based on literature studies, original research, expert consultation, and digital ethnographies, this report provides in-depth analysis of ethical issues raised by three technology families, formerly selected in Work Package 1 (Deliverable 1.2):

- Digital eXtended Reality, including the techniques of visually eXtended Reality (XR) and the techniques of Natural Language Processing (NLP);
- Neurotechnologies; and
- Climate Engineering, including Carbon Dioxide Removal (CDR) and Solar Radiation Management (SRM).

For each technology family, this report:

- Briefly presents various technologies belonging to the technology family;
- Describes key applications and use cases;
- Identifies core ethical dilemmas and provides conceptual arguments for understanding the nature, history and significance of these dilemmas;
- Identifies ethical values and principles in line with the "ethics by design" methodology and provides a contextualized discussion of the impact of the technology family on each value and principle;
- Outlines arguments for possible mitigation strategies with regard to each value or principle; and
- Provides operational checks and balances with regard to each value or principle, in the form of questions to be asked by designers, policy makers, and users of particular technologies.

The three technology-specific studies in this report are preceded by an introductory chapter identifying cross-cutting topics relevant for the ethical analysis of new and emerging technologies. These cross-cutting issues form a foundation of technology ethics and help to understand the connections and interdependencies between the ethical analyses of Digital Extended Reality, Neurotechnologies, and Climate Engineering.

Chapter 1: Introduction and cross-cutting topics

The connections and interdependencies in technology ethics rely on common issues that cut across all technology-specific studies. These include: the narratives of lay ethics; the motives of irreversibility and novelty that are present in cultural narratives and shape the temporality of innovation; the problems of justice, vulnerability, security, and the structures of power; governance and perception of uncertainty; and the concern about ethics washing. The deliverable provides a concise presentation of each topic.

Chapter 2: Ethics of Digital eXtended Reality

Technologies of virtual, augmented and extended reality, avatars and the metaverse, digital twins and affective computing can be applied in areas such as training and education, health, and remote work. They influence social relations, e.g. romantic relationships or social networking, and have an impact on practices such as gaming. Ethical analysis of these technologies is informed by two core dilemmas bearing, respectively, on the question of preference for material reality and on the Equivalence Principle between material and virtual phenomena. After a brief discussion of the underlying conceptual arguments, we analyse the

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No.101006249.

impact of eXtended Reality on the values and principles of Transparency, Dignity, Privacy, Non-manipulation, and Responsibility, as well as their relevance for the analysis of risk reduction, environmental impact, dual use and misuse, gender bias, and power and labour relations. Each value/principle discussion leads to a list of ethical questions to be asked by designers, policy makers, and users of XR technologies.

In Natural Language Processing, technologies of text generation, conversational systems, including Large Language Models, and emotion analysis and generation can be applied in areas such as education, long-term care and psychiatry, human resources, journalism, legal advice, and creative writing. Ethical analysis of these technologies is informed by three core ethical dilemmas: the lack of human-like reasoning or understanding in NLP systems, spontaneous anthropomorphisation of chatbots, and the influence of artificial emotions on human users. After a brief discussion of the underlying conceptual arguments, the report analyzes the impact of NLP on the values and principles of Autonomy, Dignity, Decency, Non-manipulation, Respect for cultural differences, Avoiding bias, Responsibility, Privacy, and Security. Each value/principle discussion leads to a list of ethical questions to be asked by designers, policy makers, and users of NLP technologies.

Chapter 3: Ethics of Neurotechnologies

Technologies of adaptive deep brain stimulation, optogenetics, functional magnetic resonance imaging using machine learning, brain computer interfaces, and functional near infrared signal can be applied in areas such as medicine, predictive diagnostics, criminal law, entertainment, military and intelligence spheres, and education. Ethical analysis of these technologies is informed by two core ethical dilemmas related, respectively, to neurodeterminism (free will, autonomy and responsibility) and enhancement. After a brief discussion of the underlying conceptual arguments, the report analyzes analyse the impact of neurotechnologies on the values and principles of autonomy, responsibility, privacy, risk reduction, and informed consent. Each value/principle discussion leads to a list of ethical questions to be asked by designers, policy makers, and users of neurotechnologies.

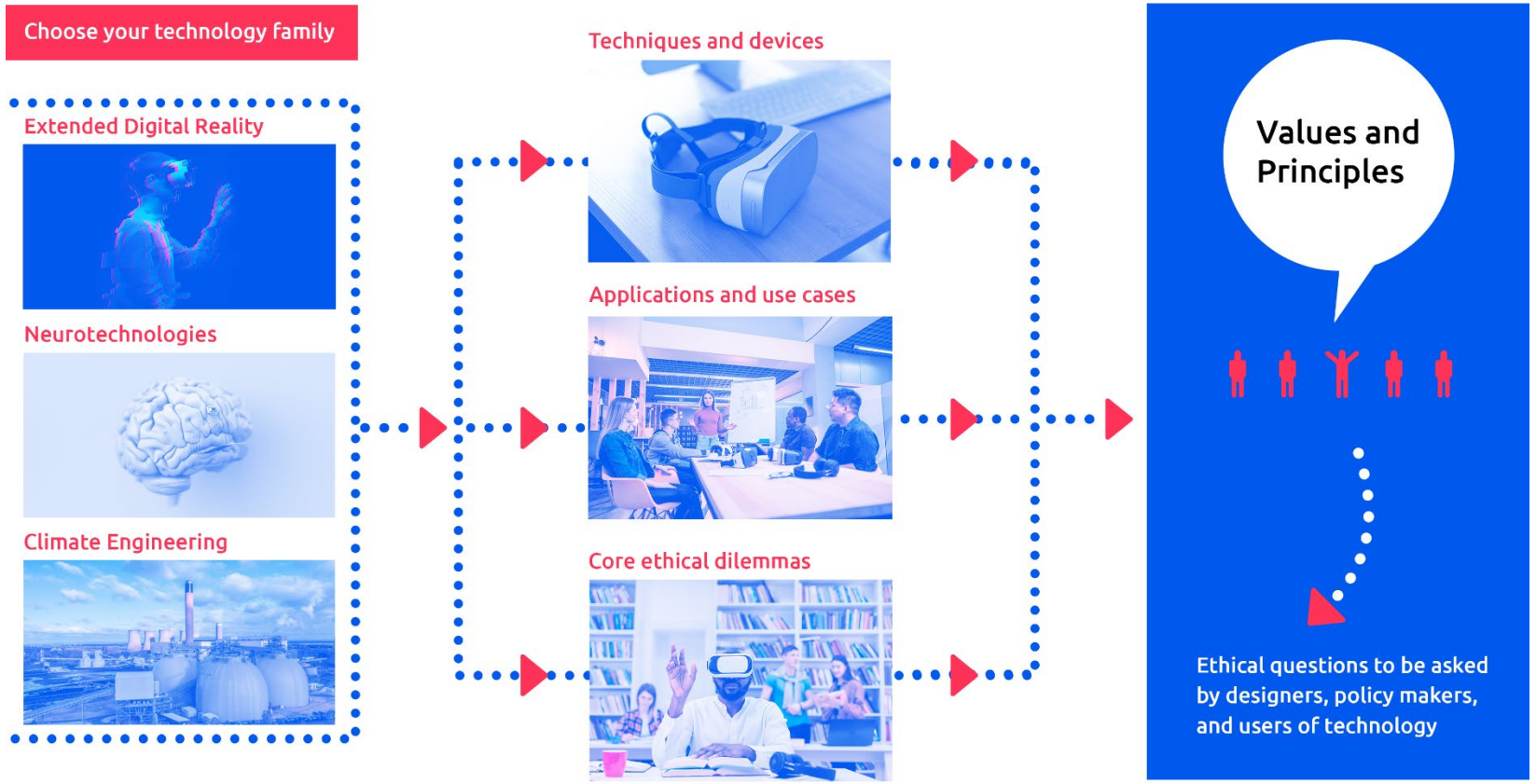
Chapter 4: Ethics of Climate Engineering

Technologies of climate engineering are discussed in two parts, one dedicated to carbon dioxide reduction (CDR), another to solar radiation management (SRM). The areas of application are not discussed since climate engineering is not a generalist technology and has a common aim – to revert climate change. Core ethical dilemmas of climate engineering apply to both CDR, and SRM. The first dilemma questions whether the less costly, though less certain methods of changing the climate can steer away from the more sustainable, yet more costly solutions. The second dilemma adds to the first one by asking whether climate engineering can encourage more even wastefulness. The third dilemma addresses hubris and the limited human knowledge and control of the full scope of changes that might be inflicted to the climate.

After discussing the common core dilemmas, more specific values and principles are introduced for CDR and SRM separately, although some values appear in both. CDR technologies invoke the considerations of distributive justice, procedural justice, future responsibility, and side effects. SRM raises slightly different concerns regarding the distributive and procedural justice than CDR and introduces two new issues of research ethics and termination shock. Each value/principle discussion leads to a list of ethical questions to be asked by designers, policy makers, and users of climate engineering technologies.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No.101006249.

Three roads to arrive at Values and Principles



This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No.101006249.

1. Introduction

1.1. Methodology

This report is based on literature studies, original research, expert consultation, and digital ethnographies. Original research in TechEthos follows the ethics-by-design approach. We have added an operational dimension to the standard approach (section 1.1.1), striving to make the meaning of fundamental values and principles clear and comprehensible for engineers, designers, researchers, users, manufacturers, and policymakers.

Additionally, we have identified cross-cutting topics in the ethical analysis of new and emerging technologies (section 1.3). This provides a methodological foundation for the study of ethical issues in TechEthos, which may also be used in future European research projects in the field of technology ethics.

1.1.1. What is "ethics by design"?

The notion of "ethics by design" (Brey and Dainow 2021; Jansen, Philip et al. 2021; Jensen et al. 2018) is based on the idea of respecting fundamental values when designing a technical system and designing for those fundamental values. Ethics by design can be understood within different theoretical and methodological frameworks (Van den Hoven et al. 2015), most notably "value-sensitive design" and "technology assessment" (Friedman and Hendry 2019; Grunwald and Hillerbrand 2013). These approaches have been in development for more than three decades and they aim to integrate human values into the design process of technical systems. Ethics by design, value sensitive design, and technology assessment each in their own ways recognize that the creation of large technical systems necessarily include human values in the process. Integration of a broader array of human – and environmental – values requires critical reflection on complicated design processes involving designers, entrepreneurs, users, and policymakers, all of whom may need help realizing the ways in which values influence and are influenced by technological designs.

The process of evaluation, which contains within its concept an etymological link to the notion of value, is an integral part of "ethics by design." If an ethical framework is formulated in terms of values, it aims at determining the degree of correspondence or fit between our current understanding of a value and the way that a system operates. An immediate example is the evaluation of biases involved in the development and training of artificial intelligence (AI) systems that rely on statistical learning from large data sets. An AI system should not merely claim to not discriminate against user-groups; bias must be measured with specific quantitative indicators and its meaning discussed. A large body of scientific work on bias assessment already exists as a part of the "ethics by design" approach. Several enterprises, including some digital giants, integrate tools for measuring explicit or implicit biases into the design process of their products (Bellamy et al. 2018).

Philosophically speaking, "ethics by design" is predicated on, among other domains, applied ethics. Like many fields, applied ethics has been marked by several paradigm shifts or "turns", in which the focus of the discipline changed towards new paradigms. Up until the early 1980s, this focus was primarily on empirical work and its translation into regulation (Florman 1994). However, Langdon Winner's work, most notably his paper "Do artefacts have politics?" (Winner 1980), could be argued to mark a new turn in applied ethics: the design turn (Van den

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No.101006249.

Hoven et al. 2017). Since then, an active campaign has been going to implement these human values as a function of design: what is needed, according to this paradigm, is a principled focus on designing technologies and technological artefacts to fit fundamental human values. However, discussions of values, e.g. sustainability, human autonomy, equity, privacy, security, etc., often run counter to dominant political economic narratives and are therefore discouraged or difficult to integrate with "conventional" design requirements. Various approaches to accomplish this in practice have been proposed. Techniques like participatory design, midstream modulation, technology assessment, and most notably value sensitive design are means by which the essential values of impacted stakeholders, both present and future, can be accounted for early on and throughout the design process.

As design decisions impact the way in which technologies diffuse and become pervasive in societies, design decisions that shape future stakeholders are limited by currently available choices. "Ethics by design" is predicated on the paradoxical notion of ensuring that the ethics of potential future stakeholders can be comprehended and inserted in the core of current technology development. The relationship between the current generation and future people, whose judgment is unfathomable and can be approximated only by imaginative projection, is at the core of many debates in applied ethics (Dupuy 2012).

1.1.2. Digital ethnography and future studies

A classic definition of ethnography by Ingold defines it as an approach with the objective to "describe the lives of people other than ourselves, with an accuracy and sensitivity honed by detailed observation and prolonged first-hand experience" (Pink and Morgan 2013). Hence *information, emotions, observership* and *subjectivity* appear to be key traits of this research methodology. Another key trait is the focus on *context* by means of reference to the concept of thick "description", borrowed from anthropology (Geertz 1977). Context here refers to the web of meanings, which constitute a culture and within which objects as cultural signs are situated. Prasad (Prasad 1997) argues that it is the ethnographer's task to uncover and present these multiple meanings and their complex connections with each other in the course of analysing any social event. He reminds readers that meanings are sometime shared but other times contradictory and contested. Greenhalgh and Swinglehurst (Greenhalgh and Swinglehurst 2011) refer to three more concepts characterising ethnography, which they term key interpretive criteria i.e authenticity, plausibility, criticality. Authenticity is gained through immersion of the ethnographer within the culture, plausibility amounts to developing explanations, which make sense to participants and are arranged in a coherent narrative, and criticality refers to questioning assumptions.

Time is a central concept within ethnography. This approach to research is usually intense and long, for example it would require a one year of fieldwork immersion at least during PhD studies. However, there is such a thing as short-term ethnography where the "immersion" of the ethnographer is for only a short period. These ethnographies are characterized by research activities being undertaken in a shorter time frame (Pink and Morgan 2013). This approach has also been dubbed "quick and dirty" as it recognises the impossibility of gathering a complete and detailed understanding of the setting at hand" (Pink and Morgan 2013). Alongside a compressed notion of time, place and space are a key feature of short-term ethnography. Ethnographic places are not simply fieldwork localities, but rather entanglements through which ethnographic knowing emerges. This is significant for the

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No.101006249.

purpose of this project because in times of Covid-19 pandemics, lockdown and remote working conditions, places, including ethnographic have become virtual, hence the emergence of short term *digital* ethnographies. This type of ethnography considers how humans live in a digital sensory environment. Pink et al (Pink et al. 2016) define digital ethnography as a way to research practices that are reported or demonstrated, for example through participants" own digital media biographies and capturing the language that is used when speaking about their area of concern. Referring to Algorithmic ethnography, during and after COVID-19, Christin (Christin 2020) defines digital ethnography as a collection of methods that entail identifying, gathering, and analysing digital data.

Ethnography tends to become shaped by the discipline it is being engaged through, and the research evolves in dialogue with theory rather than being led or structured a priori by it (Pink and Morgan 2013). In the case of our project, the theory that shapes but does not dictate or determine the ethnography is provided by different subject areas: ethics, technology and future studies. Ethics provides the main backdrop for the short-term ethnography of new and emerging technologies with high socio-economic impact. Technology and particularly, technological innovation, can be investigated "in-the-making" through ethnography (Petschick 2015); future studies represent a rich humanist perspective providing critical lens through which to investigate ethics of new and emerging technologies.

Pink and Morgan (Pink and Morgan 2013) explain how in long-term research the dialogue between theory and data-collection might be less intense, and may indeed take place largely at the end of the fieldwork, or at certain points of review, but in short-term ethnography the focus is sharper, the research questions need to be responded to more firmly and data collection and analysis intertwined.

As for the ethnography of technology Prasad (Prasad 1997) explains how the anthropological tradition within which ethnography is situated treats technologies as a cultural artifact accomplishing specific social functions as well as both reflecting and structuring social practices. In other words, in the ethnographic approach, technologies are seen as more than merely functional instruments fit for specific purposes but they are seen as cultural and symbolic object/artifact e.g., they may be ceremonial, embedding the myths of the culture in which they are situated or they may exert social control (Prasad 1997). In other words, ethnography can uncover the symbolic function of a technology within the context of the culture in which it is embedded, much like the palaeolithic hand-axe of the Ficron type represented male sexual prowess rather than just a means to butcher animals, or the polished neolithic axe fit for tree-felling was also often used to accompany individuals in their journey through death, and hence found in burials. The symbolism of the symbolic and cultural objects is not just defined by the place but also in part by the historic moments in which they are situated (Graeber and Wengrow 2021).

Table 1 is a sample of ethnographic objects we analysed, comprising of the material for analysis. A search for businesses' proposing applications within the technology families has been made from the business platform Linkedin. This was reputed a better source than Google for search thanks to its filters which helped to gauge the relevance of the results rather efficiently as it contains filters such as companies, people, region, industry and company size. We selected a mixture of webpages and YouTube videos to use as ethnographic objects of analysis. YouTube videos were selected when they included talks at a conference or interviews

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No.101006249.

by media agent rather than solely promotional videos which are more staged and may have hindered the detection of any spontaneity of emotions triggered when talking about the future.

Company's Reference Number	Ethnographic Object Type	Technological application	Country in which the company is based
1	Website page	Electroencephalography (EEG) and Brain-Computer Interface	Lithuania
2	Website Page	Wearable medical Device for monitoring Parkinson's disease	Greece-UK
3	YouTube Video	Neuromodulation through prismatic lenses	Italy
4	YouTube Video	Brain-Computer Interface (implant)	US
5	Website page	Carbon Dioxide Removal and Utilisation	Sweden
6	Website page	Carbon Dioxide Removal	US
7	YouTube Video	Carbon Dioxide Removal and Geological storage	Switzerland
8	YouTube Video	Carbon Dioxide Removal	US
9	Website page	XR – holographic display	Denmark
10	Website page	XR - extended reality experience	UK
11	YouTube Video	VR social platform	US
12	YouTube Video	VR and AR	Portugal

Table 1. List of references to the digital ethnographic objects (video or company website) systematically- selected for the digital ethnographies. The names of the companies have been withheld to ensure anonymity.

1.1.3. First round of expert interviews

During expert interviews, ethical dilemmas, questions informed by epistemological analysis as well as the 'guiding questions' method suggested by Stahl et al. (Stahl et al. 2017) have been used in order to open ethical reflection on new and emerging issues. In addition to this, the interviews have followed a similar structure to that of the literature review where questions around future ethical issues and impacts have been explored, as well as the ethical principles and values that arise when analysing each technology family.

The TechEthos project focuses on the ethical issues associated with the three technology families, therefore the criterion for interviewee selection was technical and ethical expertise associated with Climate engineering, Extended Digital Reality and Neurotechnology. Eight interviews have taken place online using MS Teams.

The contact details of the interviewees was identified through collaboration with the TechEthos project partners. The interviewees was contacted via a template email, after agreeing to an interview each interviewee have been sent a TechEthos information leaflet and, a consent form to complete, sign and return as their acceptance to participate in the interview.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No.101006249.

Follow-up email contact have been made with all potential interviewees who have not responded by return of the completed and signed consent form, within seven days of the original email being sent.

The semi-structured but flexible interviews were approximately 30 minutes duration with anticipated scope for extension beyond, given interviewee active/engaged participation and willingness to continue. Accordingly, the infterview protocol consisted of a minimum of eight essential, open questions. The interviews were audio and video recorded via MS Teams, and the insights have been captured as a summary of each question. Later, the insights generated during the interviews have been inserted into the text of the present report, noting that the insight is coming from an interviewed expert.

The interviewees were asked the following questions in a semi-open format:

1. Can you tell us about your area of expertise, how many years have you worked in your field of interest?
2. As a result of technological innovation in the area of (technology family) how do think the world will change by 2045?
3. In your view, what do you think are the benefits associated with this technology by 2045?
4. Can you anticipate what risks and harms might arise?
5. Who are the main beneficiaries of this [technology family]? And who will be excluded in your view?
6. Considering the global interest in the issue of ethics what do you predict to be the ethical issues that could arise by 2045?
7. Do you think we have gone past the point of reversibility & irreversibility of this technology? And please explain why?
8. Is there anything else you would to add which we have not covered already?

1.1.4.Second expert consultation and interviews

The second round of consultations with experts was conducted after the early draft of the current report was ready. Different parts of the draft were submitted to different experts and qualitative interviews and workshops were set up to receive feedback on the following questions:

- Clarity: Is the meaning of the value in the context of this technology family clear and comprehensible?
- Completeness: Is the main argument in the subsection complete? What should be added?
- Operationalization: Are the questions at the end of the subsection helpful operationally? Is anything missing in that aspect?
- What else do you find interesting and worth mentioning about this technology family?

The consultations took part as a form of workshop with the ADIM board members on June 13th, 2022, organized by TechEthos Work Package 6 in close collaboration with Work Package 2. The entire draft report, in particular sections 2.2.5, 2.4.4, 2.8.5, 3.1.1, 3.4.2, 4.4.1, 4.5.1, have been addressed in a discussion on the overall structure, then in three plenary sessions dedicated to the three TechEthos technology families. The workshop was attended by:

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No.101006249.

Florin, Marie-Valentine	Executive Director, International Risk Governance Center, Ecole polytechnique fédérale de Lausanne (Switzerland)
Gefenas, Eugenijus	Professor, Center for Health Ethics, Law and History at the Medical Faculty of Vilnius University (Lithuania)
Guston, David	Associate Vice Provost for Discovery, Engagement and Outcomes, Global Futures Laboratory (USA)
Hiney, Maura	Chair of the ALLEA Permanent Working Group on Science and Ethics (Ireland)
Mocchio, Elena	Head of Innovation and Development, UNI – Ente Italiano di Normazione Milan (Italy)
Parker, Andrew	Project Director, SRM Governance Initiative (UK)
Philbeck, Thomas	Managing Partner, SWIFT Partners (Switzerland)
Rementeria, Maria José	Social Link Analytics Team Leader, Life Science, Barcelona Supercomputing Center (Spain)
Renn, Ortwin	Scientific Director, Institute for Advanced Sustainability Studies (Germany)
Vakhshtayn, Victor	Senior Researcher, Center for Fundamental Sociology, Higher School of Economics (Russia, in exile)

All feedback has been documented. Additions and corrections suggested by the ADIM Board were implemented in June 2022.

Additionally, the drafts of entire chapters of each respective technology family have been offered to the following experts, who returned detailed commentary, which was duly considered and implemented. Some of the interviews were conducted by email, while others were conducted during a specially organized workshop on June 27th at Sorbonné Université in Paris.

The experts on DXR were:

Chatila, Raja	Sorbonné Université (France)
Nordmann, Alfred	TU Darmstadt (Germany)
Pelachaud, Catherine	CNRS-ISIR (France)

The experts on Neurotechnologies were:

Gaillard, Maxence	Université catholique de Louvain (Belgium)
Torrence, Steve	University of Sussex (UK)
Chneiweiss, Hervé	CNRS and INSERM (France)

Thus, the deliverable has greatly benefited from the attention of leading experts.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No.101006249.

1.2. Relation to other TechEthos work

The scenario development process in Work Package 3 involved ethical questions from the start[1]. This process started with (i) identification of trends and drivers, (ii) followed by the identification of key factors, and (iii) creation of future projections (according to Social, Technological, Economic, Ecologic, Politics and Values (STEEPV) considerations). Thereby the key factor identification was based on an impact-uncertainty analysis.

The participatory process in WP3 and WP7 also includes ethical considerations. The discussions on the three TechEthos technology families (climate engineering, digital extended reality, neurotechnologies) involve a discussion on cross-cutting ethical issues as well as on specific values and principles. The stakeholders representing various function systems in the sense of Luhmann (see section 1.3.5) and the related organizations (universities, enterprises, public administration etc.), as well as the public (NGOs[2], citizens[3]) were consulted by the TechEthos team on ethical, societal and legal issues.

1.3. Cross-cutting topics in the ethics of new and emerging technologies

1.3.1. Narratives of lay ethics

Technological innovation implies more than a new set of techniques. It ultimately creates "new social practices and even institutions that transform the ways in which human beings interact with the world around them" (Grinbaum and Groves 2013). Innovation is a future-creating activity: by bringing something new into the world, it changes the world itself – often incrementally, sometimes more radically. Ethical reflection must therefore acknowledge that the responsibility associated with innovation is a responsibility for the future it helps to create. This future does not yet exist in actuality, yet it can be thought of. As laypeople and oftentimes the experts themselves think, tell stories, and make judgements of such technological futures, their imaginaries are structured by several recurrent narratives.

In a seminal publication in 2010, Davies and Macnaghten analyzed the debate on the then-emerging field of nanotechnology (Davies and Macnaghten 2010). Based on sociological studies involving focus groups in the UK, they claimed that the dominant technoscientific narratives of control and mastery were countered in public perception with alternative narratives belonging to what they have deemed "lay ethics". In the same year, Jean-Pierre Dupuy contributed a detailed analysis of three such narratives (Dupuy 2010):

1) "Be careful what you wish for", based on the motifs of exact desire and too big a success;

2) "Messing with Nature", based on the motifs of irreversibility and power;

[1] Adopted from (Haraldsson and Bonin 2021; Schoemaker 1995; Sessa et al. 2021; Theis and Köppe 2018; Walton et al. 2019)
[2] For example, EUREC (the European Network of Research Ethics Committees) and ALLEA (the European Federation of Academies of Sciences and Humanities | All European Academies) are partner in TechEthos, bringing in the perspectives of a network of national Research Ethics Committees (RECs) and a network of more than 50 academies of sciences and humanities.
[3] Six science centers across Europe are associated with TechEthos as linked third parties (LTPs), each of them conducting local citizen engagement workshops.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No.101006249.

3) "Opening Pandora's box", based on the motifs of irreversibility and control.

In the original work of Davis and Macnaghten, these three narratives are complemented with two more:

4) "Kept in the dark", based on the motifs of alienation and powerlessness;

5) "The rich get richer, the poor get poorer", based on the motifs of injustice and exploitation.

The key conjecture formulated by Davis and Macnaghten and confirmed through Dupuy's research: lay ethics with its recurring narratives is not limited to the perception of nanotechnology, but applies to all new technologies as they emerge in the public eye and become a subject of broad societal debate involving non-experts and lay users (Swierstra and Rip 2007).

1.3.2. Irreversibility

The motif of irreversibility cuts through several fundamental narratives of lay ethics. Initially analysed with regard to nanotechnology (Grinbaum 2010), it is also present in the public discourse with regard to the three TechEthos technology families.

Digital Extended Reality may lead to irreversible anthropological change in humans and in the relations between them. The blending of the material and the virtual inaugurates a new kind of existence with far-reaching anthropological and cultural implications: "I believe that there's something called a metaverse generation. We see digital and physical reality as distinct from each other, and we see digital as less than physical. But this younger generation sees them as not only equal but as not separate. They live their lives both at the same time. They view it differently. They naturally socialize. They understand how to get around, and understand the social norms" (Takahashi 2022)

Natural Language Processing may alter the nature of language and social communication by adding a class of non-human agents (in a metaverse or in the new merged material-virtual reality) that perceive and generate natural language on a par with humans. Neurotechnologies may provide tools that will directly and irrevocably influence one's brain, modifying the concepts of agency and responsibility beyond what humankind has known until now. Most evidently, climate engineering may intentionally unleash irreversible modifications of the unique climate system of the planet Earth.

Ethically speaking, all such foreseeable irrevocable changes need to be analysed on the case-by-case basis and placed in context. However, the motif of irreversibility that is contained in each of them leads to similar patterns of ethical concern and judgment. The types of ethical concern and judgment that directly follow from the motif of irreversible change provoked by one's technological action can be illustrated by these instances of religious text:

1) An action may lead to irrevocable change because the force has been applied that is too strong. This excess of power leads to damage and is morally condemned as hubris (Talmud Sotah 47a:15).

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No.101006249.

2) An action may lead to irrevocable change because it is too hasty (Gen 49:4) or unreflected. The agent may then be condemned as unstable, arrogant, unbounded in their desires, or boastful.

3) When irrevocable change has already occurred, the action that had produced such change may be judged, depending on the nature of its consequences but because of their irreparable character, as wrong, crooked, or twisted (Ecc 1:15).

The notion that excess of power or too much success incurs a supernatural danger, especially if one brags about it, has appeared independently in many different cultures and is deeply rooted in human nature: "Ancient Greek mythology and later Greek thought distinguish between four different kinds of circumstances: successful action may provoke jealousy of the gods (*phthonos*), it may lead to divine retribution (nemesis), it may cause complacency of the man who has done too well (*koros*), or it may lead to arrogance in word, deed or thought (hubris). Hubris is condemned by the Greek society and punished by law, but reaction to the other three is more subtle. Phthonos and nemesis are dangerous and must be feared. The attitude that the Greeks have towards koros is rather ambivalent: the complacency assumed in this notion makes someone's life untenable, however koros can hardly be avoided, for it goes hand in hand with ambition, or the inability to put an end to one's desire of great achievements, called philotimia. […] Thus perfect success forbids peace of mind, and, by way of analogy between ancient and modern ethical thought, this is at the same time a part of the innovator's human condition and a moral problem of its own" (Grinbaum and Groves 2013).

In ethics the property of irreversibility is not by itself morally condemned in all circumstances. Provoking irreversible change is inherent to human action (Arendt 2013). It may be a daring course of events but it is not always wrong to enable it. For a negative judgment to emerge, irreversibility is accompanied by a nostalgic look at the past and at the situation that is forever lost, eliciting the desire of return based on the "Golden Age" motif present in mythology and literature, e.g. in Virgil's Fourth Eclogue. On such a Golden Age background, the contemplation of irrevocable change becomes a call for repair, a hope for a better future even when the return to the mythologized past is no longer possible, and a source of melancholy (Klibansky et al. 2019).

A classic metaphor to express this morally ambiguous stance on irreversibility is cutting off the salt from a salt mine. The absoluteness and irreversibility of cutting the salt are underscored in the use of the salt metaphor to describe the covenant between the people of Israel and their Lord as "an eternal salt-like covenant before the Lord" (Numbers 18:19). This use of the salt metaphor is obviously positive. At the same time, the same metaphor of irrevocable change is used for markedly negative judgment, for example to characterize the forever fruitless and dead land as "salt-sodden soil that will not be settled" (Jeremiah 17:6). Thus the irreversibility of action invariably brings dramatic consequences, be they positive or —as it is most often the case in narratology— negative.

All these patterns of judgment are transparent in the ethical debate on the three TechEthos technology families. As in the previous examples, irreversibility is linked with a judgment of an actor's boldness beyond the ordinary, which in turn calls for heightened comparisons, added caution, and diminished trust. In terms of virtue ethics, the scientist or the innovator, e.g., a climate engineer, may be judged as too bold and his action, as too powerful. In terms of the consequences, the irrevocable effects, e.g., of XR on children or vulnerable individuals, may

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No.101006249.

lead to the condemnation of the entire technology. The speed of change, conjoined to its irreversibility, is a constant source of concern in the digital sphere as well as in technological innovation in general.

1.3.3. Novelty and speed of change

The narratives of novelty are some of the defining features of technological innovation. They can elicit a variety of reactions. An immediate impression upon the perception of novelty leads to a quick rationalization and spontaneous ethical judgment. The spectrum of such spontaneous reactions typically lies within the five narratives of lay ethics. As time goes by, spontaneity yields a more seasoned, empirically grounded and reflected assessment. The motif of novelty and the associated judgment evolve in the historic time of technology development, but they also depend on the past and future projections made by the present generation of technology developers and users: What would our ancestors say if they lived to see it? Are we breaking away from their tradition? What will future generations say and how will they judge us? (Dupuy 2012; Jonas 1985). From this temporal perspective, ethics of technology essentially amounts to questioning the speed at which technology brings novelty into the world: Is the change in human condition induced by technology occurring too quickly? In other words, are 'we' going too fast? If 'we' were to 'slow down', what would it mean in practice? In other words, who should take the decision to slow down and what does 'taking time to adapt' amount to?

Novelty evokes curiosity and, beyond it, all sorts of fantasies about possible futures. It is represented in the figure 1 below of the Gartner's curve. At an early stage technological innovation typically provokes polarized perceptions: on one hand, hopes and promise a better life; on the other hand, fear and feelings of an imminent catastrophe. The dichotomy between salutary and apocalyptic technology often dominates social debate and adoption.

The three TechEthos technology families can be placed at different locations on the Gartner curve. On the basis of a metaverse hype that started in 2021, one may reasonably argue that digital extended reality is currently close to the peak of inflated expectations. Neurotechnologies, on the contrary, have arguably passed this peak, moving to the zone of what Gartner calls "enlightenment". Their applications and associated narratives are becoming more sober and realistic. As for climate engineering, it can be located at the very outset of the curve within the "technological trigger" area.

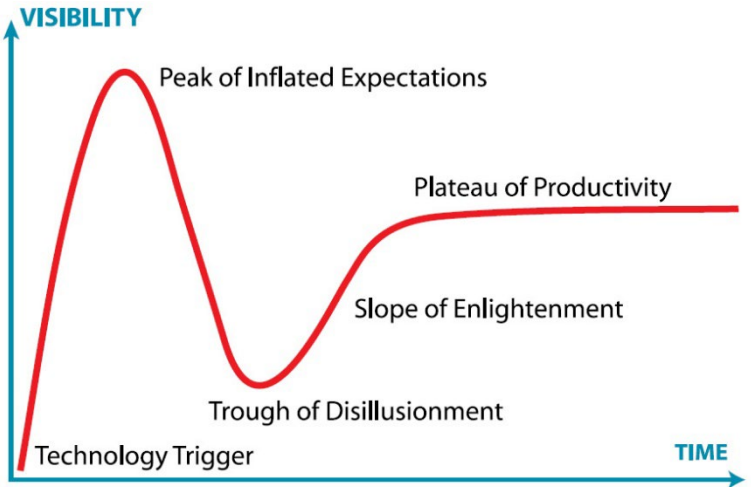


Figure 1. Gartner's hype cycle (gartner.com)

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No.101006249.

As noted in section 1.3.1, novelty is present in all the narratives of lay ethics. More deeply, this motif relies on the core cultural narratives that determine our civilizational approach to technology. Two such cultural narratives put forward the various aspects of ethical judgment with regard to techniques previously unseen and unheard of.

Firstly, the biblical myth of *Tobias and the fish* addresses the fear of novelty and the problem of uncertainty in a new, emerging reality (Grinbaum 2012). At the beginning of the book of Tobit, a young Tobias leaves the home of his parents for the first time. On his journey he's accompanied by an angel. At the end of a full day of walking, Tobias is tired and finds himself on the shore of a big river:

"And he [Tobias] went out to wash his feet, and behold a monstrous fish came up to devour him. And Tobias being afraid of him, cried out with a loud voice, saying: Sir, he cometh upon me. And the angel said to him: Take him by the gill, and draw him to thee. And when he had done so, he drew him out upon the land, and he began to pant before his feet. Then the angel said to him: Take out the entrails of the fish, and lay up his heart, and his gall, and his liver for thee: for these are necessary for useful medicines" (Tobit 6:2-5)

How could it be that a simple fish in the river frightened Tobias? How could he take this threating fish out of the water so easily? If the fish was small, why such fear? A mythological story always allows for a variety of interpretations. One such reading is that Tobias had never seen a live fish before. Sheer novelty made him cry with a loud voice. Fear of novelty engulfed him entirely.

This fear of novelty only disappears when the angel gives Tobias a protocol to follow, leading to him producing useful medications. One should note that the angel in this myth does not say "don't be afraid" or "everything will be fine", but gives purely procedural instructions with a manifestly useful purpose. Tobias's spontaneous fear of novelty dissolves thanks to this obvious usefulness and the simplicity and clarity of the instructions given by the angel. These are no verbal reassurances of trust but a user's manual.

Secondly, the well-known Greek myth of Prometheus exhibits a connection between stealing fire from the gods, giving it to the humans in an act of technological innovation, and eliciting Zeus's rage. Fire is a novum given to humans to use as they please. However beneficial it may be for humans, the rage of Zeus cannot be restrained by such planned increase in the human well-being. The god's reaction is solely dictated by *koros*, a feeling of jealousy, viz. that men will perform better than the gods themselves (Grinbaum and Groves 2013). On this interpretation, Prometheus, like Ulysses later on, acts on his *philotimia*, the love of great achievements, and the Promethean myth can be used to describe moral ambiguity necessary to achieve technological progress (Politi and Grinbaum 2020).

The virtue ethics approach takes the figure of Prometheus as a paradigmatic innovator. Much like paradigmatic innovators in today's technology, e.g. Mark Zuckerberg for a metaverse or Elon Musk for brain-machine interfaces, this figure is morally ambiguous. Bachelard claims that "the Prometheus complex is the Oedipus complex of the life of the intellect" (Bachelard 1964). Although Prometheus could be portrayed as a champion of humankind, he could also have tricked and challenged the gods. At a close inspection, technological innovators may be driven by personal ambition, the desire of recognition, prestige, or career advancement.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No.101006249.

According to the virtue ethics approach, ethical judgment with regard to their innovation depends on the judgment made of their individual qualities and of their person as a whole.

The anxiety of missing out emerges as a motif from the study of digital ethnographies. A webpage presenting an extended digital reality business reads: "The future of storytelling: capture your audience's imagination by blending virtual and physical worlds to immerse them within your narrative." This reference to "the future of storytelling" foretells the company's view of a single monolithic future. The company implies that anyone not joining in the innovation will be missing the ever-changing modernity and excluded from "the future".

The mythological narratives of Tobias and Prometheus, as well as the imagination of the future that emerge in the TechEthos digital ethnographies, show the relevance but also the limitations of the TechEthos approach to technology ethics. The sheer novelty of the three technological families chosen by the TechEthos consortium elicits radical reactions akin to Tobias's fear of novelty, or the anxiety of missing out, in itself a form of fear for the future. In these circumstances, as the narrative suggests, it is of little help to build verbal reassurances of trust or global benefit to humanity. Polarization of perceptions between catastrophic fears and unbounded hopes of salvation can be mitigated by a series of simple and clear instructions on how to use these technologies for the user's immediate and perceptible benefit. Such instructions may not be available at an early stage; thus, fear may remain with the users despite all the best efforts of the technologists and the innovators.

Furthermore, personal qualities of the technologists and innovators have a major bearing on ethical judgement. The virtue ethics approach brings home the idea that, under uncertainty about the future use of technology, ethical judgment is dominated by personal trust into the symbolic figures associated with this technology ("the angels" in myth or "technology evangelists" in the technological world). Such engineers-turned-evangelists are always ambivalent figures, like Prometheus, rather than godly incarnations of goodness. Analyzing the families of emerging technologies in separation from the judgment on these individuals who, in the public eye, act as flag-bearers for these technologies is necessarily a limited approach to technology ethics.

Among the three TechEthos technology families, XR has a set of individuals associated with the current public discourse on the metaverse. Public perception of a metaverse depends on one's opinion with regard to these figures. Neurotechnology is currently as an early stage of acquiring its own circle of evangelists, e.g. Elon Musk is increasingly associated with BCI, as shown in the TechEthos deliverable 3.3. This process has also begun for climate engineering with people like Russ George. Hence, one may expect significant temporal and methodological divergences in the construction of ethical reflection and judgment between the three TechEthos technology families.

1.3.4. Vulnerability and the structures of power

Technologies developed today engender concerns with distributive justice, inexorably, as a result of political, cultural, economic milieux of the age.

Whether the instrumentalization of ethical initiatives is intentional, such as in Big Tech companies' push for ethical responses to the regulation of AI. There we can observe a *common failure to account for relations of power and structures of inequalities*. This is what the ethics of

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No.101006249.

new and emerging technologies needs to improve on to avoid its misuse and to build on firmer grounds.

What is required is an approach to ethics that stays away from "fact- and reality avoidance of ideal theory" dominant in mainstream ethics (Mills 2005, p. 179), one that "gloss[es] over the power conflicts characteristic of politics" (Delacroix and Wagner 2021a, p. 1). Rather, we need an ethics that accounts for power relations, an ethics that situates its analysis "in the world", including its structures of inequalities and injustices (Blodgett et al. 2020). This is what will prevent its manipulation. Although one can highly praise the Crawford's book *Atlas of AI* as it convincingly "challenge[s] the structures of power that AI currently reinforces", we do not agree with her claim that "we must focus less on ethics and more on power." (K. Crawford 2021, p. 224) Ethical analysis and responsible initiatives *can* and *should* consider power relations and systems of oppressions. Alternatives to mainstream ethics exist and offer great theoretical resources in this regard, such as the ethics of care that starts from "the reality of inequality before the idealness of principles" (Laugier 2011, pp. 185–186).[4]

What does it mean to situate new and emerging technologies in the world, accounting for its power structures and historic inequalities, in the ethical analysis of these technologies?

The effort to account for power structures in the ethical analysis of technologies requires one to ask the "who question". This applies throughout the life-cycle of a technology, from research to impact. Which actor/who is behind such an initiative? How does this affect the power of this actor? Who is affected the most adversely and how? What are the capacities and vulnerabilities of the actor? Furthermore, the "who question" is not only relevant at the individual or group level (or company level) but also matters at the international level, such as between the Global North and the Global South[5]. An ethical analysis should look at the way risks and benefits are distributed along several dimensions, including gender, race, sexuality, social class, age, ability, origin, and North/South relations. The intersectional approach is of key value here. The main point is to look closely at the reality under study and to avoid being blinded by what Mills calls the classic liberalism ideology of "abstract and undifferentiated equal atomic individuals" (Mills 2005, p. 168). Below we highlight a few aspects to consider for the ethical analysis of digital extended reality, neurotechnology, and climate technology.

For example, Blodgett et al. (2020) examined 146 papers published between 2015 and 2020 that analyse biases in NLP. They showed that these studies fail to address the root cause of discriminatory impacts of NLP due to a neglect of the historic inequalities and the system of oppression that have shaped perceptions toward languages, in particular African-American English (AAE) in the US context. Blodgett et al. demonstrate that one cannot understand and address the issue of biases in NLP without accounting for the historical conditions that shape how AAE is viewed today. As they state: "AAE as a language variety cannot be separated from its speakers— primarily Black people in the U.S., who experience systemic anti-Black racism— and the language ideologies that reinforce and justify racial hierarchies." (Blodgett et al. 2020) Blodgett et al. ask: "Who are the speakers of AAE? How are they viewed?" (Blodgett et al. 2020) These considerations are not external to the subject matter; they are fundamental to

[4] Translation by Anaïs Resseguier. The value of the ethics of care for the ethics of AI is further developed in (Resseguier and Rodrigues 2021).
[5] The Global North and the Global South refer to two groups of countries worldwide that have similar socio-economico-political characteristics. The Global North groups together mostly Europe, North America, and Australia, while the Global South refers to regions of South America, Asia, Africa and Oceania.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No.101006249.

understanding the biases and discriminatory outcomes of NLP, and eventually to finding mitigating measures to address this problem.

Moreover, such attempts to promote fairer systems, not only fail to address the key issue at stake, but they risk being counter-productive as they tend to further veil the systems of oppression that they neglect to account for, and as such contribute to entrenching these systems. This is what D'Ignazio and Klein argue when they point to how data ethics initiatives tend to work with "concepts that secure power" as opposed to challenging it (D'Ignazio and Klein F. 2020, p. 60). This critique directly echoes that of Mills toward "mainstream ethics" that tends to "rationalize the *status quo*" of structural injustices and existing power structures rather than seeking more just systems for all (Mills 2005, p. 181).

This section has shown that the ethical enquiry on the ethics of new and emerging technologies requires engagement with questions of power, i.e., that power is a crucial component of ethics. It shows how an approach that places power at its heart allows for a more convincing ethical study of these technologies. This does not mean giving up on the value of ideals or high-level principles – such as those of dignity, fairness, transparency, or privacy – but rather to ensure they are realised, and for all. As Mills puts it, "the best way to bring about the ideal is by recognizing the nonideal" (2005, p. 182).

1.3.5. Governance of uncertainty

Uncertainty is accompanying the evolutionary path of new and emerging technologies. Following Kant's categorial considerations, uncertainty is related to the mode of contingency and thus distinguished from modes such as necessity and possibility[6]. In an early investigation of uncertainty Knight expressed his "conviction that contingency or 'chance' is an unanalysable fact" (Knight 1921, p. lx) insofar as dealing with it on the basis of probability theory and the mathematics of the distribution of possibilities is insufficient. In other words, if contingency "results from excluding necessity and impossibility" (Luhmann 1984, p. 106), mechanism of uncertainty absorption must exist.

According to the theory of social systems, the absorption of uncertainty is provided by meaning-constituting systems, i.e., including psychic systems (i.e., persons as unity of a meaningful related psychic-organic complex of actions and experiences), as well as social systems (Luhmann 1990). Thereby, meaning refers to the principle of ordering human experience taking into account that "all orientation is construction, is difference reactualized from moment to moment" (Luhmann 1997, p. 18), which is framed by an interconnected complex of meaning (*Sinnzusammenhang*) or horizon of meaning, as Husserl puts it (Husserl 1936, pp. 48, 249)[7]. Meaning is structured by expectations and various social systems are providing structures of expectations at least on three interrelated levels (Luhmann 1984, pp. 2, 293, 408)

- on the micro level interaction systems (personal communication) are structured by centering related to the persons present, and their themes of communication

[6] Kant's table of categories includes the modes of possibility/impossibility, existence/non-existence, and necessity/contingence (Kant 1998, p. 212)
[7] See for the discussion of the central role of Husserl's phenomenology in Luhmann's theory of social systems (Buchinger 2012).

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No.101006249.

- on the meso level, organizational systems are structured by decision premises, such as job descriptions (roles), communication channels (who reports to whom), and decision programs (if-then decision programs/strict coupling, purposive decision programs/loose coupling)
- on the macro level "encompassing societies" (which under modern conditions converge to a world society) are structured by programs related to function systems such as economics (investment and consumption programs based on prices and market mechanisms), science (theory/method programs), and politics (ideologies, political programs)

Personal communication itself is seen as a process which is potentially a situation of not knowing, because of the double contingency (one person is a black box for the other; ego/alter scheme): "The basic situation of double contingency is then simple: two black boxes, by whatever accident, come to have dealings with one another. Each determines its own behaviour by complex self-referential operations within its own boundaries" (Luhmann 1984, p. 109). Structured expectations provide a framing, i.e., they are unpacking the black box (Glanville 1982). They also help to start a conversation and provide space for unexpected, contingent communicative actions and reactions.

Governance of uncertainty with regard to technologies can be described as making systems resonate (Luhmann 1986, p. 12; Rosa 2016). In this sense, TechEthos aims at facilitating resonance on the level of personal systems (individuals) such as researchers, on the level of organizational systems such as research organizations (universities, enterprises), research ethics bodies, research funding bodies, and policy bodies; and eventually on the level of function systems (science, economy, law, politics). Governance in the form of media of steering[8] comprises "knowledge" (i.e., guidelines using the ethics by design approach, enhancement of ethical frameworks), "money" (as far as research funding bodies are concerned), and "law" (enhancement of legal frameworks).

1.3.6. Perception of uncertainty

Prediction, foreseeing of the future is a key feature of all human cultures and was traditionally expressed by oracles and other mythical creatures. In modern scientific societies, prediction moved from the professions of clairvoyants, fortune tellers and prophets to professionals, academics who would develop techniques and methodologies for "seeing the future". A dictionary definition of the future is "going or expected to happen or be or become" (Sardar 2010, p. 178). The term "futurology" was first used by Flechteim in 1966 with the publication of History and Futurology (Sardar 2010, p. 178) a new field that would explore the "destiny" of humans. He regarded the subject as a branch of "historical sociology". In socio-technical capitalism societies, technologies are reshaping the present, offering up new possibilities, and problems. For this reason, scholars have begun to regard the current period of geological time as the *Anthropocene*, an epoch shaped by human activity, it "describe(s) a connection that reaches back into the past and far into the future" (Schwägerl and Crutzen 2018, p. 6).

As Sardar states "it is a technocratic misconception to assume that knowledge of the future in a singular, monolith, scientific sense is possible we need to abandon the idea that

[8] See for a discussion of media of steering according to the theory of social systems (Buchinger 2007, 2010).

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No.101006249.

futures studies is a "discipline" with rigid boundaries, fixed theories, esoteric terminology and "great men" always men who have laid the foundations of this entity" (Sardar 2010). A critical problem for thinking about uncertainty is the role of *time* – not understood as linear and singular but, as Schneider (Schneider 2019) explains it, as the future seen as an outcome of gestures and properly studied as "interval crossers" and "interval openers" (Schneider 2019, p. 147). Time is linked to motion, in so far as it can be considered as a measure of movement (Darwiche 2019).

Because time is complex, there can be possible or even alternative futures. In regard with possible futures, Dupuy & Grinbaum explain that the future is made of "futuribles", meaning precisely the open diversity of possible futures (Dupuy and Grinbaum 2004). In this view, the future cannot be predicted but "alternative futures can be "forecasted" and preferred futures "envisioned" and "invented''-continuously" (Sardar 2010). In this view, potential, possibilities, and mutual diversity are central to the future.

Understanding uncertainty requires imagination, and "the imaginary as resources for reshaping our world and imagining new relations" and prioritising the role that stories play in constructing human existence (Spengler 2019, p. 168). Studying the imagination implies facing uncertainty about what the future will *look* like. According to Dupuy and Grinbaum (Dupuy and Grinbaum 2004) tackling uncertainty should include a study of the linguistic and cognitive channels through which descriptions of the future are made, transmitted, conveyed, received, and made sense of, because "the very description of the future is part and parcel of the determinants of the future" (Dupuy and Grinbaum 2004). That is, the description of the future yielded by those who create the technologies which can potentially shape it, is expected to "cause" change much like a teleological force or final cause. As Spengler (Spengler 2019) states, the possibility for a viable future depends on the imagination and on the imaginary as resources for (re-)shaping our world and imagining new relations. Hence the reason for embracing future studies as research framework for studying the ethical impact of emerging technologies.

What is the future – is it anytime that is beyond the present, or a place that is always shaped by fictional imaginaries and any prediction must consequently be partly, a work of fiction. Moreover, artists, including novelists have shaped future predictions, from Isaac Asimov to Arthur C. Clarke (Potts 2018). Science fiction writer Ursula Le Guin warned against calling upon artists to predict the future, as she claimed they do the opposite – they tell lies (Guin 2015). Le Guin also notes the struggles she had to write believable female characters into her science fiction in a male dominated field - as the human is often reproduced as the *default male* - woman has to be explicitly stated if they are to be included in future forecasts as noted by feminist writer Caroline Criado Perez in her book *Invisible Women: Exposing Data Bias in a World Designed by Men* (Perez 2019).

Forecasting is a deterministic process. Generating a forecast means understanding why or how a process unfolded (Ghosh 2019). Forecasting in fact treats the system as if it were a purely physical system, and implies that anticipating the future of the system has no effect whatsoever on the future of that system (Dupuy and Grinbaum 2004). For example, forecasting the weather will not have an impact on how the weather will play out.

On the other hand, foretelling implies, and indeed counts on the effect that imagining the future may have on its realisation. A tarot card reading is indeed interpreted (and meant

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No.101006249.

to act) as a call for action on the side of those whose future is being foretold. Dupuy and Grinbaum (Dupuy and Grinbaum 2004) explain how foretelling works: To foretell the future in projected time, it is necessary to seek the loop's fixed point, where an expectation (on the part of the past with regard to the future) and a causal production (of the future by the past) coincide." Foretelling is close to prophecy in its reference to final causality also commonly known as self-fulfilling prophecy. The prophet is the one who seeks out the point where voluntarism can achieve the very thing that fatality dictates (Dupuy and Grinbaum 2004). A distinction between foretelling and prophecy can be made in terms of the magnitude that the latter can bear, e.g., in the Bible, Moses" prophecy of 7 years of calamities is prophecy insofar as it helps the Christian God to free the Jewish population from their Egyptian enslavers – hence an impact of great magnitude.

Much like foretelling and prophecy, revolutions realise/open up the possibility of an unimaginable future, in other words, make a possible, desired future manifest. Revolutions "introduce futures that their subjects and observers alike are unable to imagine until the event itself changes the course of history" (Millar 2019).

Sociologist Zygmunt Bauman noted that the future is not always a desired goal, and he coined the term "*retrotopia*" as an umbrella term for those movements and trends that seek to get back to something, rather than moving somewhere else. Hence ideas of the future are intrinsically connected to the past and present, imagined and factual, as opportunities, and destruction are feasible outcomes of any process. The past can be changed in the future, each time it is recalled, as "the gestic materiality of call and response is how the past changes" (Schneider 2019). In fact, as well as multiple futures, there can be multiple past(s) because as Bergson in (Darwiche 2019) state, duration is unity and multiplicity. Every time past memories are evoked new details surface whilst others disappear. For this reason, Bergson argues, the past can be never a homogeneous monolith (Darwiche 2019).

The speakers in one video analysed in the ethnographies appear extremely ambitious and very optimistic in their self-outlook, saying that they are "rewriting the future" because they are changing how energy is used. All of this is said to be done while both gaining profit and creating jobs. The two individuals one the director or creator of this video and the other the CEO of the company, framed the technology as being absolutely disruptive using the phrase "the great rewrite", in and of itself is transformative in its framing. As Larson (Larson 2019) states, it is a fundamentally subjective and temporal concept. However, the future's imagination can be characterised by a lack of optimism too. Future Studies is not without its critics, for to have a future must imply a desired or imagined state of existence – calling into question who decides this future? Who is left out or excluded from future imaginings? The question is whether technology innovation is the solution to the problems developed in tech-capitalist societies? Technology, is the engine of capitalism innovation, opening up the possibilities of creating new products, processes and practices, underlying a belief in unfettered creativity and flexibility of the human species to adapt to any technologically inspired living arrangement.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No.101006249.

Höjer and Mattsson (Höjer and Mattsson 2000) have identified key questions about uncertainty of the future:

1) identifying "cyclic behaviour in socio-technical changes";

2) viewing one technology to be crucially reliant on the development of another (in their case it was transport and communication that entangled and connected);

3) interrogating basic assumptions about a field (in their case it was the "hypothesis of constant travel time" as a stable);

4) human and resource relationships (Höjer and Mattsson 2000, p. 685).

The future is a "fiction" of sorts, shaped by practices, ideas and, extrapolated into some undefined future point – problematically producing a determinism – if this, then that – view. Moreover, they suggest that "backcasting" as an alternative and better predictor than "forecasting" in cases where future scenarios are seen as detrimental, and harmful. Sardar prefers the term "alternative futures" due to the possibility of plurality, identity crises and meaning (Sardar 2010).

Ethically speaking, certainty about the "future", if it exists at all, is a contested domain, heterogenous, and diverse, while ethics proposes a set of standards to be recognised and incorporated into technological practices and artefacts.

1.3.7.Security

In relation to digital technologies, security is mostly understood as information security, which comprises different fields, like access control, cyber security, cryptography, anonymization, etc. Ethical concerns about security arise when there is a conflict between entities (Whitman 2003). For example, there can be sensitive information about a user on a defender platform and some attacker wants to gain access to it. The intentions of the attacker and the defender are at odds. These intentions in themselves carry an ethical interest, for example the attack can be malicious or causing intentional damage (black-hat attack) or it can be a testing attack to increase security (red-hat attack), or even a genuinely well-intentioned attack, geared to defeat a malicious opponent or to preserve some ethical value, like privacy (white-hat attack, or ethical hacking). In any case, information security captures the array of methods that a defender can employ to parry the attack or to make the information useless to the attacker.

Security is an essential ethical concept because it is necessary to preserve the ethical design of any application. Systems and algorithms that have been designed ethically have been exploited because they were not secure and attackers managed to steal or replicate them, then use for malicious purposes. Being negligent with regard to security can negate all ethical design by enabling damaging consequences. Thus, security should be considered as a core value in an ethical approach to technology.

Security situations are best analysed as conflict situations. To understand whether security has been compromised and what security measures are necessary, one needs to analyse several critical aspects to understand the ethical implications of an attack. As Polybius advised: "There is no more precious asset for a general than a knowledge of his opponent's guiding principles and character, and anyone who thinks the opposite is at once blind and

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No.101006249.

foolish" (Polybius 1980, p. 81). The Greek historian had in mind Hannibal, who defeated multiple Roman generals by exploiting their personal weaknesses. The power balance in a security situation always depends on the resources available to both sides. However, if the attackers have good information about the target, they can leverage simple attack tactics to defeat a significantly more powerful defender. For example, there are obfuscation techniques available to an individual that can help that person evade commercial or government surveillance (Brunton and Nissenbaum 2015). Sometimes attacks are carried out purely for financial gain (e.g., financial scams), but often with a political agenda (like pre-election data poisoning), or both (e.g., ransomware attacks on Law Enforcement Agencies). The intentions of the conflicting parties are the key ethical determinant in security situations.

1.3.8. Ethics washing: lessons learned

From the 2010s, technology ethics has turned its attention to the ethics of AI. A number of lessons can be drawn from AI ethics for technology ethics more generally. In particular, there are good lessons learned from critiques the field of AI ethics has received on risks of "ethics washing", I.e., pushing for an ethical governance of AI in order to avoid hard laws that could limit technological innovations.

AI governance experts Delacroix and Wagner have asked: "Why has ethics come to acquire such a bad name in AI and data governance?" (Delacroix and Wagner 2021b, p. 1). What has happened in the ethics of AI that may explain such severe "ethics bashing" (Bietti 2021)? And what can be done to avoid it for other fields of technology ethics? This section seeks to answer these questions. It draws lessons learned in the ethics of AI and applies them to future developments in the ethics of new and emerging technologies.

The ethics of AI, and more generally, many responsible AI initiatives, have been subject of strong criticism since about 2018.[9] Experts have pointed out that these fail to achieve their aim of identifying and mitigating potential harms related to the development, deployment, and use of AI. Some experts have shown that ethics initiatives are being misused by industry actors to avoid the development of regulations that would limit their business. This misuse of ethics has been called "ethics washing" (Hao 2019; Mittelstadt 2019; Rességuier and Rodrigues 2020; Wagner 2018). Other experts have shown that these ethics initiatives contribute to maintaining the *status quo*, i.e., they serve the interest of the privileged members of the society at the expense of the marginalised ones who are precisely those the most exposed to the negative impacts of AI (D'Ignazio and Klein F. 2020). How does one explain these strong critiques? What has made such manipulation of ethics in the ethics of AI field possible?

This is due to a certain vulnerability in the approach to ethics in these AI ethics initiatives. There is an inherent pitfall in this approach that makes it prone to its instrumentalization, whether this manipulation is deliberate (i.e., to serve interests such as

[9] These ethics and responsible AI initiatives include documents listing principles for the development, deployment and use of AI, such as, the High-Level Expert Group on Artificial Intelligence, "Ethics guidelines for trustworthy AI", European Commission, Brussels, 2019, https:// ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai; OECD, "Recommendation of the Council on Artificial Intelligence", adopted on 22 May 2019, https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL- 0449; Google, Artificial Intelligence at Google: Our Principles, https://ai.google/ principles/. They also include work that seek to promote responsible AI, such as the work conducted by the Responsible AI team at Facebook, now Meta, as described in (Hao 2021).

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No.101006249.

those of Big Tech) or unintentional (i.e., committed out of ignorance or naivety). This weakness is due to ethics often being blind to power relations and structures of inequalities.

Philosopher Charles Mills' work on what he calls "mainstream ethics" is quite helpful to clarify the shortcomings of AI ethics. Mills defines "mainstream ethics" as an approach to ethics dominated by "ideal theory", i.e., a theory that relies on "idealisation to the exclusion, or at least marginalization, of the actual." (Mills 2005, p. 168) John Rawls' theory of justice is one of the most renowned illustrations of such approach.[10] According to Mills, this ideal theory of ethics is "obfuscatory" in "crucial respects", those related to "actual historic oppression and its legacy in the present, or current ongoing oppression" (Mills 2005, pp. 166, 168) He highlights the paradox of a theory of justice that neglects actual conditions of injustices and inequalities, precisely what should be the core concern of such theory. Mills argues that such theory is founded on a "tacit social ontology" that "will typically assume the abstract and undifferentiated equal atomic individuals of classic liberalism" and "will abstract *away* from relations of structural domination, exploitation, coercion, and oppression" (Mills 2005, p. 168). The "obfuscatory" nature of mainstream ethics in "crucial respects" that Mills highlights helps shed lights on the critical elements that AI ethics tends to be blind to, i.e., relations of power and historic structures of inequalities. The efforts towards fairer natural language processing (NLP) systems illustrate such blindness.

[10] This is what we have qualified elsewhere as the principled approach or legal approach to ethics (Resseguier and Rodrigues 2021; Rességuier and Rodrigues 2020)

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No.101006249.

2. Ethical Analysis: Digital Extended Reality

Digital Extended Reality is a technology family, i.e., a collection of technologies that are related to each other, with a common functionality to emulate and imitate human traits and social circumstances: language, appearance, lived spaces, objects, experiences, etc. Two main technologies of the Digital Extended Reality technology family are Extended Reality (XR) and Natural Language Processing (NLP). They will be characterized in the subsequent sections by their main constituting techniques. Although XR and NLP can be considered as standalone technologies and are self-sufficient, there are particular ethical issues raised by their combination. A resulting artifact can have both an XR and an NLP implemented as part of its production. For example, GTBX-100 produced by Gatebox Inc. simulates a virtual partner experience that involves both a holographic image, which is part of XR, and advanced chatbot capabilities, which is part of NLP (Aronsson 2020; J. Liu 2021). Such combinations warrant an ethical analysis on Digital Extended Reality that considers the XR and NLP technologies separately but with the prospect of their combination in certain artifacts (Reiners et al. 2021).

Digital extended reality can change how people connect with each other and their surroundings by combining advanced computing systems (hardware and software) with virtual environments. Potential ethical issues in digital extended reality center on the cognitive and physiological impacts as well as behavioural and social dynamics, such as influencing human actions, misinformation, monitoring and surveillance, privacy, security, sensible data management, etc., as outlined below.

2.1. Technologies of extended reality (XR) and the metaverse

XR is a term that is used in the literature to cover the broadest range of techniques and artifacts that relate to virtual and simulated experiences using digital technologies. XR (also called “mixed reality,” (Speicher et al. 2019)) is sometimes conceptualized as a continuous spectrum between a completely real environment and a completely virtual environment (Milgram et al. 1995; Skarbez et al. 2021). In this sense, XR is a classification. However, XR can also be considered as a particular technique that “allows interaction of real objects with synthetic virtual objects and vice versa” (De Guzman et al. 2019). This section differentiates among techniques prominent in the technology of XR.

2.1.1. Virtual reality

Virtual reality refers to the environment that is completely simulated by digital means with the goal of creating a synthetic experience for its user (Doerner et al. 2022). Simulating virtual reality currently focuses on visual aspects and digital graphics, although other senses are being incorporated into building virtual reality environments, also referred to as virtual realities (McLellan 1996). Extreme examples of complete virtual reality environments have been the subject of science fiction in the works of Stanislaw Lem and the Wachowski brothers. They describe a reality that consumes the user and produces complete immersion without the ability of the user to distinguish it from the real environment (Warnke 2016). However, a perfect simulation of the environment is not necessary for users to engage in behaviours of playing, learning, communicating, etc. Users suspend their expectation of perfect immersion and continue to use virtual reality systems while being conscious of the fact that they operate within a virtual environment.

This project has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement No.101006249.

Virtual reality system is a computer system consisting of suitable hardware and software to implement the concept of a virtual environment. Virtual environments are usually defined through their immersive aspects, e.g., a virtual environment is an "immersive, interactive, multi-sensory, viewer-centered, three-dimensional computer-generated environment" (Cruz-Neira et al. 1993). Immersion is a debated topic in both virtual reality studies, and psychology of media in general (Bormann and Greitemeyer 2015; Southgate et al. 2017). Immersion distinguishes VR from other kinds of human-computer interfaces. However, the use of this term does not benefit from a consensus in the literature. A technological definition centers on the qualities of the virtual reality system. (Slater and Wilbur 1997) identify the following elements as contributing to immersion:

- *Inclusivity* indicates the extent to which the user's senses are obfuscated by the virtual reality system, i.e., the more the user is isolated from the real environment, the more immersion they experience.
- *Extensivity* refers to the range of different sense experiences.
- *Surrounding* indicates the extent to which the experience is panoramic and wide ranging rather than limited in one area or field of view.
- *Vividness* indicates the resolution, representation quality, and dynamic range of graphics or other media within a sense modality.

In addition to the technological definition, immersion is sometimes referred to as a subjective quality of the users' experience when they enter a virtual environment (Witmer and Singer 1998). The latter understanding of immersion is also referred to as the feeling of presence. Generally, virtual environments seek to elicit the feeling of presence in its users through various means of immersion.

One feature that XR introduces in comparison to other current digital systems is multisensory and subliminal data collection from various interfaces. XR hardware devices include sensors that are able to detect eye movement or bodily gestures. The movements can be recorded and analysed. The potential interfaces that can supplement the data from eye or hand movement include height and weight of the user, heart rate, including electrocardiography (Condon and Willatt 2018), perspiration rate, oxygen concentration, temperature, facial features and facial expressions, body dynamics, gait, voice, and others (De Keyser et al. 2021).

The collection of such data is significant because it allows analyzing user behaviour that is not conscious. For example, users cannot control their perspiration, which can be used to infer arousal or other cognitive or emotional variance in users (Bryant and Howard 2019). In turn, the resulting data can inform marketing decisions and be used to create nudging patterns that function without conscious awareness from the user (Sethumadhavan and Phisuthikul 2019). It is significantly different from the type of data being collected on social media platforms at the moment, which mostly consists of conscious user actions, for example, views, clicks, reactions, and viewing time (Arora et al. 2019).

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No.101006249.

2.1.2. Augmented reality

Augmented reality is part of the XR technology family that combines elements of real and virtual environments instead of trying to obfuscate senses and achieve complete immersion. While immersed in a virtual reality, the user cannot experience the real environment around them. In contrast, augmented reality allows the user to see the real world, with virtual objects superimposed upon or combined with the real environment. Therefore, augmented reality supplements environments, rather than completely replacing them (Azuma 1997).

Some scholars argue that the developers of augmented reality usually seek to achieve the immediate and seamless perception of the real environment enriched by virtual content in real-time (Doerner et al. 2022). This would imply that the added elements of augmented reality should be resembling reality to the largest extent possible, so that perception of real and virtual objects would be seamless and even indistinguishable.

However, like with virtual reality, complete indistinguishability is not part of the state of the art. Many augmented reality applications today consist of imaginative additions to reality, like selfie filters, games, or animation. One of the most successful artifacts of augmented reality to date has been a mobile application game called Pokémon Go developed by Niantic and Nintendo (Rauschnabel et al. 2017). The game combined real geographical data of players with animated avatars. It did not rely on indistinguishability and seamless incorporation of virtual elements.

Magic Leap Inc, one of the technological leaders in augmented reality systems, refers to augmented reality in their patent applications as a "scenario [that] typically involves presentation of digital or virtual image information as an augmentation to visualization of the actual world around the user" (Bradski et al. 2016). This represents an inclusive yet definitive characterisation of augmented reality. It leaves room for both recreational and informational applications, which may have different requirements as to the realism and seamlessness of virtual objects in real environments.

2.1.3. Avatars and the metaverse

A metaverse is an artefact of extended reality that includes and often emphasizes the social element of immersion by allowing multiple users to interact in one virtual or augmented environment. It differs from a multiplayer game in that there is no generally agreed purpose or mission to complete it. Rather, a metaverse generally consists of three key characteristics, although they may depend on a company that operates it.

- Social engagement - avatars that participate in a metaverse generally represent real people or at least an animated version of them. A leading metaverse company has patented a "skin replicator" that reads the facial features of the user and feeds the data to an "avatar engine" that produces an animated version of the avatar that remains based on the biometric data of the real person (Albuz et al. 2022). Some metaverse participants have described the experience of being "like you but not you" (CoinYuppie 2022).
- Personalisation - despite representing or resembling a real person, metaverse avatars can be customized according to the preferences of the user. It is

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No.101006249.

designed to be a "representation with customization" (U. Sharma 2022). Various customization features are expected to become a commodity that is sold and traded within the metaverse. Some experts estimate that the retail of customization features along with personalized advertising will ground the economy of a metaverse (Murphy 2022).

- Persistence - in order to create ownership and scarcity in the metaverse, some experts expect non-fungible tokens (NFTs) to be used to sign virtual contracts in a metaverse for buying, selling, owning, and any other contracting (Wang et al. 2021). The use of NFTs is expected to ensure that certain transactions in a metaverse remain persistent. However, the precise mechanism of how that should work is unclear. There are critical responses as well that claim NFTs are not suitable to achieve that (Olson and Che 2022).

Despite a metaverse imitating certain features of material reality (e.g., persistence), it changes their meaning and temporality or the relation with time. There can be multiple setups on how time passes in the metaverse.

In one setup, called heterochrony, avatars in a metaverse might not age or age differently from the material person. They may persist session to session but they do not continue to evolve and interact during offline hours. Yet a metaverse as a whole continues to run despite one person being offline. This disconnect can put pressure on the person behind the avatar to spend as much time as possible online, so as not to miss out or fall behind virtual events. Similar mechanisms have been shown to be involved in the addiction to video games (Duman and Ozkara 2021).

An alternative set up, synchrony, keeps up the activity of the avatar by performing actions in a metaverse time on the part of the user while they are offline. This maximizes interactions, since other users can continue to interact with an offline user, but it diminishes the autonomy of the avatar since not all actions performed by the avatar will be traceable to the human actions. Several temporal frameworks or ontologies exist across video game genres that can become exemplary to a metaverse (Zagal and Mateas 2010). Manufacturers of a metaverse will have to decide on the ontology of time. It is possible that multiple versions of different time ontologies will coexist.

A metaverse is evolving quickly and has a potential to produce high socio-economic impact. It provides the environment for realistic human social problems that do not lend themselves to common ethical solutions. It is thus important to include a metaverse as a category in this present analysis.

2.1.4. Digital twins

A digital twin is a digital replica of a physical object that can possess dynamic features, like the synchronizing of data between the physical twin and the digital twin to monitor, simulate, and optimize the physical object (El Saddik 2018). Digital twins are largely used in training engineers and their professional tasks. For example, implementing digital twins is discussed in the context of energy supply or car manufacturing (General Electric 2021).

As described in WP3 (D3.1), a digital twin is a virtual model designed to accurately reflect a physical object. The object being studied — for example, a wind turbine — is outfitted

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No.101006249.

with various sensors related to vital areas of functionality. These sensors produce data about different aspects of the physical object's performance, such as energy output, temperature, weather conditions and more. This data is then relayed to a processing system and applied to the digital copy. Once informed with such data, the virtual model can be used to run simulations, study performance issues and generate possible improvements, all with the goal of generating valuable insights — which can then be applied back to the original physical object.

Educational, training, engineering, commercial, farming, organizational, and other artifacts of augmented reality or virtual reality can require very precise replicas of real world objects. (Chi et al. 2013) have determined that "with the rapid development and adoption of augmented reality (AR) applications, there are numerous opportunities for integrating AR and improving conventional methods used in the fields of architecture, engineering, construction, and facility management." Currently, aviation and medical education is largely embracing training with augmented and virtual reality. This shows the applicability of XR that uses precise representations of real objects and seamless experience (Borgen et al. 2021; Yeung et al. 2021). However, that is not the exhaustive characterisation. At least luditive artifacts remain reliant on animated objects. This is in part due to limited computing power of mobile augmented reality applications, which need to be addressed for more precise training application via cloud computing, as (Chi et al. 2013) also note.

2.1.5. Affective computing in XR

Data generated through XR applications can include biometric and personal data as well as other information allowing analysing and influencing human emotions. For example, an advertising algorithm can leverage the particular movement of the user's eyes to determine not only that they have opened an advertisement but also that the user has looked at it, which parts of it stood out, how long it was looked at, and similar characteristics (Chatellier 2022). Current governance mechanisms of personal data are not fully equipped to deal with the new challenges brought about by the involvement of XR in affective computing.

2.2. Core ethical dilemmas in XR

Core ethical tensions in XR concern the status of virtual objects and actions performed while engaged in one of the modalities of XR. One view holds that virtual actions should be equivalent to material actions and that the ethical analysis of material reality applies to the virtual circumstances. An opposing perspective describes XR as a different medium with inequivalent norms, values, and conflict resolution approaches.

2.2.1. Is there a preference for material reality?

The core ethical dilemma raised by the emergence of virtual reality is whether virtual experiences mediated via XR are equivalent to material experiences gained in the real world. It is not only a question of whether the same visual stimuli can be recreated, but whether virtual experiences evoke similar or equivalent emotions, behaviour patterns or judgments in XR.

A useful dilemma was formulated by Robert Nozick in 1974 It is called the "Experience Machine": "Suppose there was an experience machine that would give you any experience you desired. Superduper neuropsychologists could stimulate your brain so that you would think

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No.101006249.

and feel you were writing a great novel, or making a friend, or reading an interesting book. All the time you would be floating in a tank, with electrodes attached to your brain. Should you plug into this machine for life, preprogramming your life's experiences?" (Nozick 1974, p. 264).

Nozick's example relies on a brain-computer interface rather than mediated experience, and thus lies on the intersection of XR and Neurotechnologies. However, his intuition was that people would not prefer even perfectly simulated experiences. The synthetic character of the experiences would make them less than real and that would discourage people to plug in. However, other philosophers have questioned Nozick's intuition (De Brigard 2010), saying that "many people choose to use drugs they know are dangerous, such as alcohol, in spite of the fact that they know that it is difficult to give up the habit of using them. So why not opt for a perfect experience machine (that you can opt out from if you like) with no bad side effects—and stay plugged into it?'" (Tännsjö 2007).

The arguments for "plugging in" an experience machine are hedonistic, claiming that enjoyment should be the key internal justification of action (Crisp 2006). The arguments against entering the experience machine are often eudaimonistic, claiming that meaning and purpose are key justification for human happiness and merely experiencing enjoyment would not constitute a meaningful life (Waterman 2008).

This debate raises important questions regarding the status of experiences in XR. Are they the same as they would be in material reality, given a high immersion and resolution? Is there a preference for material reality despite pleasures offered in virtual environments? Is the preference constant or evolving with time? Is it based on desire or rationality? How do the ethics of real environments relate to the ethics of virtual environments?

2.2.2. Mode of being of virtual objects

A moderate position in the philosophy of digital objecthood claims that digital objects are relational, i.e., they exist insofar as they are experienced and conceptualized by a human mind (Grinbaum 2019; Hui 2016). Digital objects are the types of things we experience in the digital world, like "an image" or "a video". However, it is not clear how they can be individual objects if all they consist of is digital data. Hui writes: "By digital objects, I mean objects that take shape on a screen or hide in the back end of a computer program, composed of data and metadata regulated by structures or schemas" (Hui 2016). The digital object, according to Hui, is a new kind of relational system that presents itself as an individual object. It consists not just in zeros and ones but also in the material capacity to process data. Digital objects as technical systems are understood in relation to their social and economic significance.

A radical position in philosophy regards virtual objects and environments as being of the same nature as material objects and environments. One branch of this radical approach (Baudrillard 1994) claims that in the postmodern society hyperreal simulations (including virtual reality) are more real than real environments: "The realm of the hyperreal (e.g., media simulations of reality, Disneyland and amusement parks, malls and consumer fantasylands, TV sports, virtual reality games, social networking sites, and other excursions into ideal worlds) is more real than real, whereby the models, images, and codes of the hyperreal come to control thought and behaviour" (Kellner 2020). However, this should be understood as a metaphor in view of how important simulations have become in postmodern society, rather than in relation

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No.101006249.

to the being of virtual objects. This approach is considering the importance of cultural symbols and not objects of computer simulations.

Another branch of the radical approach, called simulation realism, claims that simulated objects are real: "I defined simulation realism this way: "If we're in a perfect simulation, the objects around us are real and not an illusion" (Chalmers 2022). Moreover, "Simulation realism holds that things are largely as we believe them to be. Now, we believe that cats exist, and that cats do things, and that they are real cats. Simulation realism entails that in a simulation, these beliefs are largely true" (Chalmers 2022). If simulated objects are real objects, and if our beliefs about these objects are true, then from an ethical point of view real and virtual environments are equivalent.

However, these radical positions are unconvincing since they are metaphorical or require perfect simulations, which are not approaching the state of the art. The relational position is convincing because it does not require substantial digital objects. Even if simulations increase in resolution or graphics, there are other elements of virtual realities that make virtual objects differ from material objects. We take these terms to mean "technologically simulated virtual objects/reality" and "non-technologically simulated objects/reality".

Whether one engages with the realist or relationalist positions in terms of objecthood, there are additional questions regarding the ontology of time in the metaverse.

2.2.3. Value of virtual objects

If we keep the distinction between virtual objects and material objects, consequences of actions in material reality certainly do not equal the consequences of actions in virtual reality. We take these terms to mean "technologically simulated virtual objects/reality" and "non-technologically simulated objects/reality". For example, driving fast in virtual reality does not imply the same risk as driving fast on a material road, or destroying a valuable art piece in virtual reality does not equal destroying it materially, since most XR applications do not implement scarcity. Assault on other individuals also does not equal a virtual assault in terms of pain that is caused, although the negative psychological effects make both types of actions unethical (Basu 2021).

However, scholars argue that virtual objects retain ethical value not because of the equivalent consequences involved, but because values or behaviour patterns formed in XR can be transferred to material reality, i.e., people engaging in XR can develop behaviours that they can reiterate in the material reality and then produce the equivalent negative consequences (Brey 1999).

The transfer of behaviour argument has been influential in the ethical thinking about XR (Ramirez 2019). Yet it relies on presuppositions about virtual and material experiences, which are not fully justified. Namely, the transfer argument needs to show that virtual and material actions and beliefs are equivalent and induce the same behaviour patterns or values but it fails to show that to the full extent.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No.101006249.

2.2.4.Cognitive equivalence

For the transfer of behaviour from virtual to material realities to work, an equivalence needs to hold between virtual and material actions and beliefs. To establish the equivalence, some scholars claim that VR produces stimuli equivalent to the material ones, while others argue that it induces the same sense of immersive presence. One important argument for the Equivalence Principle states that virtual experiences are phenomenally equivalent to material experiences, and therefore the psychological effects of experiences and behaviours should be the same. For example, Lombard and Ditton claim that "our current understanding of [mediated experience] is based on studies in which it has been assumed that mediated (i.e., presence-inducing) stimuli are exactly the same as nonmediated stimuli'" (Lombard and Ditton 2006).

The assumption that the Equivalence Principle is justified because actions and beliefs in material reality and XR are or soon will be indistinguishable has been widely shared in the psychological literature on mediated experience. However, the mediated stimuli in VR and in other media, e.g., television, are phenomenally non-equivalent, for they lack the same qualia and ontology. Moreover, there is evidence against mediated experience achieving the same practical knowledge (Gale et al. 1990).

There is evidence that higher resolution and interactivity increases immersion. For example, Hoffman et al. has found that high-resolution improved VR analgesia (the distraction from physical pain) by about a third over low-resolution treatment (Hoffman et al. 2004). However, their work does not demonstrate that the human ability to distinguish between virtual and material reality will not evolve together with the graphics. Mader et al. claim that "with proper training, feedback, and incentives, observer performance on distinguishing computer-generated from photographic images can be significantly improved" (Mader et al. 2017). Deep-fake technologies bring in a new factor to these considerations. Still, even with deepfakes some "algorithms struggle to detect the deepfake videos that humans find to be very easy to spot" (Mader et al. 2017)

In addition, graphics alone do not constitute a full XR experience. It needs to have an interactive element (Brey 1999; Wender et al. 2009). A crucial form of interactivity is social interaction. Heeter claims that "people want connection [with other people] more than any other experience. Placing more than one person in a virtual world may be an easy way to induce a sense of presence regardless of the other perceptual features of the world" (Heeter 1992). In a VR setting, human-to-human interactions are often replaced by human-to-character interactions. In such cases, the characters must be realistic to retain the feeling of presence. Lombard and Ditton stress that "socially realistic experiences are [...] more likely to evoke a sense of presence. To the extent that the content "rings false" the consumer is reminded of the mediated and artificial nature of the experience and the sense of presence should be destroyed" (Lombard and Ditton 2006).

Socially awkward interactions with digital subjects are likely to cause the uncanny valley effect, especially if the environment is rendered with realistic graphics and intended to create a duplicate of the material reality. The uncanny valley effect is a feeling of eeriness when the user becomes aware of the ontological difference between a human and a virtual character (Mori et al. 2012; Grinbaum 2015).

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No.101006249.

Social realism is a necessary condition for cognitive equivalence because without it immersion and presence are destroyed. However, social realism is very difficult to simulate in the virtual or game environment. This is because social interactions depend on the human subject or player that relies on their cultural background to understand them. On the contrary, social interactions in VR (often enabled by NLP) can contribute to the homogenization of social interactions: "A [language model] used to create cultural content such as movie scripts could, for example, contribute to public discourse becoming more homogeneous and exclusionary" (Weidinger et al. 2021). Moreover, there is a risk that people will imitate homogenized social interactions, thus eliminating cultural differences and minority perspectives (Pasquale 2015).

Another indication against the transfer of behaviour argument is the Proteus Effect (Yee et al. 2009; Yee and Bailenson 2007). Online environments can encourage deindividuation due to anonymity or reduced social cues (McKenna and Bargh 2000, p. 9). In the metaverse, the avatar is not simply a mask but constitutes an entire self-representation. Thus, researchers expect that the avatars have a significant impact on how people behave online: "Users who are deindividuated in online environments may adhere to a new identity that is inferred from their avatars" (Yee et al. 2009, p. 274). Furthermore, users in online environments may conform to the expectations and stereotypes of the identity of their avatars and to conform to the behaviour that they believe others expect. This dissociation between avatar identity and material person identity (deindividuation) constitutes a challenge to the transfer argument.

Humans are capable of considering the whole multisensorial experience that will inform them about the status of reality. Especially in the use of virtual reality, users remain aware of the fact that they decided to use a VR device. Together, the equipment and the multi-sensory experience combined with the context of use and social realism of the characters amount to a significant obstacle for the cognitive equivalence.

2.2.5. Emotional projection

Despite the fact that cognitive equivalence is not expected to hold, there are emotional effects in XR that do not require cognitive equivalence. They are already functioning in the current state of the art. These effects depend on people anthropomorphizing virtual subjects as having psychological, emotional and moral traits.

In XR, avatars often take human form. As a result, humans interacting with virtual avatars may come to think of these agents as human-like. Anthropomorphising virtual agents may inflate users' estimates of its competencies and capabilities, especially regarding emotions, attachment, trust, knowledge, etc. For example, users may falsely infer that a virtual agent that appears human-like also displays other human-like characteristics, such as having an identity over time or being capable of empathy, individual perspective, and rationality. As a result, they may place undue confidence, trust, or expectations in these agents (Weidinger et al. 2021).

Importantly, these effects do not require the user to actually believe that the virtual agent is human: rather, a subconscious anthropomorphism takes place and users respond to more human-like chatbots with more social responses even though they know that the chatbots are not human. The notion of belief involved in this analysis is a necessary disposition for enacting change. People who engage with virtual agents know that they are engaging with a simulation but suspend that knowledge in search for emotional effects.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No.101006249.

The notion of belief has a psychological and moral impact on the human participant rather than a merely cognitive content (Dupuy and Grinbaum 2006). Knowing that the interlocutor is a machine does not preclude emotional projection (Holmes 1978). Emotional projections are part of what Dennett calls the "intentional stance" (Dennett 1971, 1987). Denett describes the intentional stance as follows: "Here is how it works: first you decide to treat the object whose behaviour is to be predicted as a rational agent; then you figure out what beliefs that agent ought to have, given its place in the world and its purpose. Then you figure out what desires it ought to have, on the same considerations, and finally you predict that this rational agent will act to further its goals in the light of its beliefs. A little practical reasoning from the chosen set of beliefs and desires will in most instances yield a decision about what the agent ought to do; that is what you predict the agent will do" (Dennett 1987, p. 17).

Humans easily project moral and emotional states on objects they interact with. Human-like virtual subjects elicit projections by imitating individual traits of some possible person. People who engage with XR know the stimuli are not equivalent, and that they are not fully immersed because of that awareness. However, they perform moral projections and experience equivalent emotions to those felt with a human subject.

Emotional equivalence is the idea that emotions from virtual experience or from interacting with virtual objects are equivalent to emotions from material experiences or from interacting with material objects. This is markedly different from cognitive equivalence, for emotional states are induced differently, and often independently, from other types of mental states. The different mechanisms of cognitive and emotional states are seen in the users of these technologies who cognitively know that the subjects are digital but continue to emotionally engage with them in a morally significant way. Thus, the ethics of engaging with digital subjects does not depend on the complete immersion in virtual reality or the complete likeness of a chatbot and human interlocutor. It relies on emotional bonds that are formed by projection on digital subjects.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No.101006249.

2.3. Applications and use cases of XR

Some XR applications are currently on the market and available to both businesses and individual consumers (Kugler 2021). The areas of application progress together with improvements in both hardware, and software.

	XR applications and use cases						
Values and principles		Training	Health	Remote work	Romantic relation-ships	Social networking	Gaming
	Transparency	x			x	x	x
	Dignity		x		x		
	Privacy	x	x	x	x	x	x
	Non-manipulation		x		x	x	x
	Responsibility		x			x	
	Environmental and security risk reduction	x	x	x			
	Dual use and misuse	x					
	Power						
	Labour	x		x			x
	Bias				x	x	x

2.3.1. Training: knowledge transfer and qualia

One of the most established applications of XR is in training skills. The areas in which XR training applications are the most impactful usually include high-risk or costly material training conditions. For example, training pilots and surgeons can be both expensive and high risk. Thus, XR applications in aviation and medical training have been well-received (Borgen et al. 2021; Yeung et al. 2021).

The process of training with XR assumes that skills acquired via virtual experience are equivalent or transferable to material conditions. Several metastudies in medicine confirm the transferability (Abich et al. 2021; Kaplan et al. 2021; M. W. Schmidt et al. 2021) and claim that "training in XR does not express a different outcome than training in a nonsimulated, control environment. It is equally effective at enhancing performance" (Kaplan et al. 2021).

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No.101006249.

Conceptually, virtual and material experience do not have the same qualia or a qualitative experience of performing an action (Nagel 1974). Thus, there can be a lack of transferability where qualia are especially important. Also, rigorous standardization and empirical validation are needed for scaling training applications.

2.3.2. Health: impaired patients and medical paternalism

XR has been applied for multiple therapeutic purposes, mostly through exposure therapy that helps alleviate phobias, anxiety, post-traumatic stress disorder, and eating disorders (Emmelkamp and Meyerbröker 2021; Harris et al. 2002; Levac and Galvin 2013; North et al. 1997). Other therapeutic applications are emerging.

Ethical issues arise in relation to individuals that can misconstrue the relation between virtual and material environments. Although specific applications can be therapeutic to vulnerable groups (Freeman et al. 2017; Langener et al. 2021). Vulnerable individuals are understood as minors or adults whose vulnerability is related to age or to physical or mental disabilities, disorders, or conditions (e.g., autism, Alzheimer's disease, phobias, anxiety, depression, etc.).

Another ethical issue for therapeutic applications is medical paternalism, meaning that treating physicians and medical application developers can engrain their decisions and worldviews into the virtual environments without the patient's consent or will. This would infringe on the patient's autonomy (Buchanan 1978). Issues relating to therapeutic virtual reality and medical paternalism have been identified early in developing such solutions (Whalley 1995).

2.3.3. Remote work: long-term effects on workers and the job market

XR work environments are available on the market and allow coworkers to host meetings and interact at a distance, sometimes using avatars. For example, Microsoft Mesh, a metaverse-like solution for social interaction, offers integration with Microsoft Teams, a popular team communication application (Roach 2021). The integration allows to host team meetings in virtual and mixed realities with some physical and some avatar participants. The use of XR in remote work was propelled by the COVID-19 pandemic that forced the majority of office workers to continue working from their home instead of physical offices (Fereydooni and Walker 2020).

Ethical challenges associated with XR in remote work include general issues of remote work, like the culturally unequal distribution of household responsibilities, increasing the gender gap (Dunatchik et al. 2021). More specific issues relate to work-life balance, where "remote work" is always easily accessible and can be overused or exploited by employers. Also, VR systems currently focus on individual experiences of a user and barely support collaborative experiences, so new measures and guidelines need to be created (Piumsomboon et al. 2017, 2018). More long-term economic effects are expected in the labour market that may no longer rely on the local workforce. This can impact city populations with higher cost of living and increase gentrification in suburban or remote areas (Brian et al. 2021).

The relationship between government and the private sector (which is more advanced in the application of digital technologies, including XR, in organizational and access areas) is

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No.101006249.

important. The ethically problematic area lies in the collection of data of the online workers and increased surveillance capabilities by the employer. The more work functions are transferred online, the more detailed and data-intensive the productivity tracking becomes. For example, if XR workplaces are implemented, the employer could begin monitoring focus and procrastination metrics through eye tracking, heart rate monitors, and perspiration sensors. The workplace would also loose para-functional spaces where workers can discuss working conditions or unionize.

2.3.4. Romantic relationships: long-distance relationships and impact on material relationships

XR solutions are available to create and maintain long-distance romantic relationships. XR can help to establish attachment between individuals that defines romantic relationships (Huang and Bailenson 2019). However, there are both technological and ethical issues related to romantic relationships. A technological one is the capture of facial expressions, intonations in voice, and speech-gesture coordination that is important for intimate communication can be cumbersome in XR environments (Walther 1996). This can mean that romantic relationships in XR take longer to develop (Lea and Spears 1995).

Ethical challenges of long-term effects of XR romantic relationships can include the diminishing importance of authenticity. With ultra-realistic XR models and avatars, it may become less important how the created avatar corresponds to the material subjects that control them. At the same time, there are ample opportunities for manipulation and misrepresentation of avatar subjects, especially directed towards vulnerable individuals. Although the technologies for XR romantic relationships are available to users, there are few studies that address the emerging ethical issues.

2.3.5. Social networking: social reality and human relationships

Facebook, now Meta, the biggest social media platform, has made a move in 2021 to establish its version of a metaverse as the new medium for social interactions and released a social virtual reality space called “Horizon Worlds” (BBC News 2021). Several ethical issues are emerging in this early metaverse. Notably, there have been incidents of harassment and abuse (Basu 2021), especially of female avatars. Personal space and access restrictions were also highlighted by a journalist who, posing as a 13-year-old girl, “witnessed grooming, sexual material, racist insults and a rape threat in the virtual-reality world” (A. Crawford and Smith 2022).

Some experts believe that social interactions create a fake world which leads to anti-social behaviour and a decrease in social interaction. Essentially, they see a distortion of what the real world is about, resulting in ethical concerns around sexting, bullying, aggression, and a loss of manners.

The greatest challenge in understanding the ethics within a metaverse is to distinguish between virtual and material ethical judgment. For example, although people can be harassed within a virtual space, no virtual equivalent of punishment for harassment exists, except for logging out or suspending a user account. Extrinsic mechanisms (like filing an official claim or report with law enforcement) that bear a high burden of proof can be used to tackle these problems, which may not be efficient.

This project has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement No.101006249.

Also, it is not clear that the harassment happens because of the lack of punishment. Some traditions of ethical reasoning and judgment were conceived on the grounds of the golden rule, i.e., treating others as one wants to be treated (Matthew 7:12), or the categorical imperative, i.e., treating others as an end, never merely as a means (Kant 2012, p. 41). However, these principles cannot be directly applied to a metaverse where people are represented by avatars. Avatars cannot be treated as an end in themselves. Material people also cannot meaningfully put themselves in the avatar's place. Therefore, ethical behaviour in a metaverse deserves a different ethical treatment not based on projecting the categorical imperative on avatars, despite the fact that there are real psychological effects of the behaviour happening in the metaverse.

Proteus Effect, which is the tendency to be affected by digital representations, such as avatars, dating site profiles and social networking personas, also takes place in a metaverse. Online environments can encourage deindividuation due to anonymity or reduced social cues (McKenna and Bargh 2000). In the metaverse, the avatar is not simply a mask but constitutes an entire self-representation. Thus, researchers expect that the avatars have a significant impact on how people behave online: "Users who are deindividuated in online environments may adhere to a new identity that is inferred from their avatars" (Yee et al. 2009, p. 274). Furthermore, users in online environments may conform to the expectations and stereotypes of the identity of their avatars and to conform to the behaviour that they believe others expect. For example, users have been assigned avatars in VR with different ratings of attractiveness, without knowing the purpose of the experiment; "Participants in the attractive condition were willing to move closer to the confederate and disclosed more information to the confederate than participants in the unattractive condition" (Yee et al. 2009, pp. 281–282).

2.3.6. Gaming: addiction and personal development

XR applications make gaming experiences more immersive and enhance the feeling of presence (Bollmer and Suddarth 2022). Some scholars argue that special ethical consideration has to be given to the types of behaviours that can be transferred from the game environment to the material environment (Brey 1999).

There are effects of the increased sense of presence and immersion that can be ethically significant. For example, increased immersion can drive addiction to XR gaming (Zhai et al. 2021). An example is in the area of massively multiplayer online (MMO) games that have an impact on children. Ethical challenges arise since we do not know what this technology does to developing minds. The neglect of the physical environment encourages to disregard one's biological body; children potentially feel more comfortable in the virtual world as opposed to the real world. Hence, an illusion is created which results in a loss of worldly connections.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No.101006249.

eXtended Reality (XR) I

TECHETHOS
FUTURE ○ TECHNOLOGY ○ ETHICS

Transparency	Should there be limits for immersion?
Dignity	Can avatars simulate the presence of individuals, including the dead?
Privacy	How to address privacy concerns raised by XR?
Non-manipulation	Can nudging be controlled in XR?
Responsibility	Should real-world sanctions be issued for virtual misconduct?

eXtended Reality (XR) II

TECHETHOS
FUTURE ○ TECHNOLOGY ○ ETHICS

Environmental and security risk reduction	How can physical and digital safety be ensured in XR applications?
Dual use and misuse	Can XR be exploited for malicious purposes?
Power	How can social justice be respected in a metaverse and its material implications?
Labour	How can just labour and economic conditions be ensured in the metaverse?
Bias	How will XR representations influence gender issues?

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No.101006249.

2.4. Values and principles in XR

2.4.1. Transparency: Should there be limits for immersion?

Transparency does not have a definition that enjoys a consensus among researchers. One common understanding describes it as having and revealing information about internal processes of a public institution, a company, or other enterprise. This type of transparency is often considered a virtue that lends to fighting corruption, enacting accountability and enhancing trust. The most commonly found definition of transparency relies on an enterprise's responsibility to make some information publicly available. It is usually formulated from the sender's (enterprise's) perspective without involving the responsibility to ensure the receiver (the public) is actually informed (Wehmeier and Raaz 2012).

A different notion of transparency can be conceived from a user's point of view, whereby the public or a group of users must be made to understand certain processes while engaging with them. For example, Drew and Nyarges define transparency as "information that allows all people who are interested in a decision to understand what is being decided, why, and where" (Drew and Nyerges 2004, p. 1642). This emphasizes the importance of the explainability of digital systems (Adadi and Berrada 2018; Brey and Dainow 2021). Explainability refers to the ability of the users to understand the information that is being disclosed. However, this does not include the modality of the users being de facto informed: "[Transparency is] an individual's subjective perception of being informed about the relevant actions and properties of the other party in the interaction" (Jiang et al. 2009, p. 628).

The latter definition can be applied to XR. Here, transparency is subjective understanding and acknowledgment by the user that they are entering a virtual environment and that they understand the digital nature of the subsequent interactions. Transparency in XR should be understood from the user's point of view, since the user should be aware of the nature of their environment. This can be implemented by alerts and signals at the beginning of an XR session, throughout or at the end of the session. In the case of mixed or augmented reality, it should be transparently acknowledged which aspects of the user's perception are material and which are digital in nature.

That transparency is central to XR is clearly shown by the TechEthos ethnographies. An emphasis is placed on first person perspective, providing a demonstration of what it is like to construct a world from the individual's point of view, starting with requesting a park, then an island with a beach surrounded by the sea, topped with a blue sky and white clouds, then a table, a picnic blanket and some drinks. Knowing that technology can identify certain words with certain images, the user recalls a certain word and a series of related images. In other words, the technology is proposing a 'cause and effect' system where the choice of the user is materialized by the software as if this was the genie from Aladdin's lamp; in this sense, the technology is fictionalized. What is missing from this representation of the future is an explanation of what the technology really is and how it works; in other words, rational transparency.

However, the demand for total transparency in morality is based on a misunderstanding of rationality (Grinbaum 2020). Williams reminds that "we must reject any model of personal practical thought according to which all my projects, purposes, and needs should be made, discursively and at once, considerations for me" (Williams 2006, p. 200). Truthfulness of

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No.101006249.

propositional reasoning does not suffice on its own to build the moral edifice; trust and unreflected commitment are required as well. This demonstrates a methodological interest of narratives (see section 1.3.1): unlike discursive analytic considerations, they are the instruments that possess a capacity to address both the transparent and the opaque components of moral reasoning and ethical judgment.

This first-person positioning of the technology raises questions in relation to transparency: who knows what are people doing whilst engaging with and within the technology? What happens when they're interacting with these systems? What is being disclosed what is being kept private? There are several processes that people are involved in, in creating this self-world, and these processes are not being made transparent. Omission is a typical strategy of representation of cultural artefacts which consists in emphasizing some features of a system whilst omitting others.

XR applications that emphasize immersion should consider implementing limits to that experience, so that omission is limited and confusion between virtual and material environments is less probable. This is critical in applications that provide a spectrum of transition between different environments, for example, Magic Leap 2, a device that allows to "blind" certain aspects of the material environment to emphasize virtual or augmented reality elements (Stein 2022).

Establishing checks and balances with regard to transparency in XR means paying particular attention to the following questions:

- Is there clear and comprehensible information about the nature of the environment in which a user engages?
- Is this information presented at key moments and intervals during the user's interaction in the virtual environment?
- What options are presented to the user who wishes to leave the virtual environment?
- How are the limits of immersion enforced?
- How do the users express their subjective understanding of the information that is being disclosed?

2.4.2. Dignity: Can avatars simulate the presence of individuals, including the dead?

Current technologies, including XR, allow one to simulate the presence of a deceased individual by using existing data collected or saved while the person was alive (Grinbaum et al. 2021). In particular, deepfake technologies could be leveraged to create a realistic digital counterpart of a deceased individual (M. Sharma and Kaur 2022). The use of data in this case is ethically questionable from multiple perspectives (Rochfeld 2022). The deceased can no longer give consent for the user of their data and the rights to the data are not inherited. However, the deceased can have a posthumous privacy interest (Banta 2015).

The posthumous privacy interest is best understood in the context of human dignity. Human dignity is the absolute worth of each person that is rational, unconditional, and independent of all other facts about the individuals. Dignity is considered incomparable, so it

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No.101006249.

cannot be exchanged or replaced by any other value (Kant 2012). Individuals cannot forfeit dignity or the right to recognition and respect.

Human dignity, unlike personal data, is considered to remain with the individual after their death. It is well illustrated by the expectation to treat the remains of a deceased individual with respect. The concept of posthumous dignity is embedded in the return-of-remains practices and the belief that the remains of the deceased should be valued and respected. For example, in forensic science, the central feature of the victim identification process is to return the remains of the deceased to their families. Cook finds that "The relationship people had with the deceased, and how they interact with the remains, will determine whether posthumous dignity is protected or violated" (Cook 2020, p. 67). Likewise, the use of the data of a deceased person will determine whether their dignity is respected.

Establishing checks and balances with regard to posthumous dignity in XR means paying particular attention to the following questions:

- Do any of the XR representations (including avatars and deepfakes) rely on deceased individuals?
- Is there a consideration for posthumous personal data treatment, including images of individuals?
- If posthumous data is used, how is it ensured that the posthumous dignity is respected?
- What options are presented to the data subjects to have control over what happens with their data posthumously?

2.4.3. Privacy: How to address privacy concerns raised by XR?

Privacy can be understood as control over the communication of personal information. Westin defines privacy as "the claim of individuals, groups, or institutions to determine for themselves when, how, and to what extent information about them is communicated to others" (Westin 1968, p. 7). However, more recent approaches have argued that "in fact the core challenge is the processes by which personal information is created" (Mai 2019, p. 113). The shift of focus towards the creation of data is in part motivated by the breaches of privacy by both the collecting institutions and malicious adversaries that illegally collect and leak data. The communication of personal data is not always secured and controlled; there are many instances of leaked, stolen, or illicitly sold data. It is thus reasonable to focus on the mechanisms of the collection of data.

XR applications can include the collection of particular sensitive data, for example eye-tracking. The leading metaverse company is patenting technologies that track individual eye movements and facial expressions during the use of an XR device (Underwood et al. 2021). Another patent describes the complete replication of the wearer's facial features to create a representative avatar (Albuz et al. 2022). Other types of biometric data that can be collected by an XR device include gait detection, emotional sentiment analysis, galvanic skin response, electroencephalography, electromyography, and electrocardiography. Thus there is potential for the collection of biometric data via XR devices. This data is sensitive personal data that can be exploited for subsequent facial recognition, neuromarketing, and nudging.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No.101006249.

Another type of sensitive personal data that can be collected by XR devices includes the physical surroundings of the user's personal or work environment. When a user engages an XR device, it is usual for it to scan the surroundings for creating boundaries. This is especially true for virtual reality devices that obfuscate the senses of the user. It is necessary to set the boundaries to avoid physical accidents, like hitting a physical object or tripping and falling while using an XR device. However, the scanned surroundings can be stored and leaked or used to identify the location of the user. It can also be exploited to detect physical security flaws in the physical environment or to identify objects within the physical space (Bye et al. 2019).

One issue of privacy in regard to eye-tracking and surroundings scanning was flagged in the TechEthos ethnographies about a specific application of XR as a holographic display. It looks like a TV screen but creates the illusion of an image emerging out of the screen. This device makes use of eye-tracking technology within the TV set. Having a camera installed in the television, and also setting it as an active parameter for the functionality of the device, means that this camera would be recording not only the room but particularly the recipients, their eye movements and potentially their facial expressions. The implications for privacy are not explicitly mentioned in the promotional material for this XR device.

It is likely that the collection and use of such personal data already is a business model for at least some XR companies (Murphy 2022). For example, Bagheri reports that "when users agree to Oculus" Terms and Conditions agreement, they give Oculus the right to automatically collect and share data regarding where and how they interact with their virtual reality experience" (Bagheri 2016, pp. 109–110).

One concern is that the loss of privacy can be used as manipulation and steers humans in a particular direction to act like "puppets". Once information has been collected, it can be released; the user does not know what the future use of their data will be.

Another privacy concern is avatar anonymity. There are reasons to have both traceable and untraceable avatars. Untraceable avatars would encourage user participation, the freedom of expression, and would be the most privacy preserving solution, since the identifying data could not be leaked or sold. Traceable avatars would bring a degree of self-regulation to a metaverse as well as accountability in case of law-breaking behaviour, and would establish a closer connection between a metaverse and the material world (Barendt 2016, pp. 122–155). It is the manufacturers responsibility to balance the benefits and dangers of anonymous avatar use.

Some interviewed experts believe that generating an avatar raises questions about digital property and sovereignty. Furthermore, ethical issues can arise from interacting with one's own avatar in real time.

If XR technologies proliferate in the workplace (Sagnier et al. 2021), it can become too easy for the employer to track the activity and personal interactions of the employee during work hours. It would be equivalent to constant surveillance in a physical office. This type of surveillance can result in negative effects of workplace relationships and violate the public/private boundary, especially when working from home (Ball 2021).

Being forced to use XR at the workplace can lead to depersonalisation symptoms.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No.101006249.

Establishing checks and balances with regard to privacy in XR means paying particular attention to the following questions:

- Do any of the XR devices collect personal data (e.g., biometric or spatial awareness)?
- How is this data communicated and stored?
- How does informed consent evolve with the emergence of new types of data collected by XR devices?
- What are the opt-out options from the collection of such data?
- Does the application rely on the collection of such data and is the use of it justified?
- Can the identity of the avatar be traced to the identity of the user?
- How is the right to go offline implemented?

2.4.4. Non-manipulation: Can nudging be controlled in XR?

Nudging is a term that means a suggestion, an incitement or a boost. Nudging deals with inconspicuously pushing someone in a desired direction of action. For example, a virtual environment could encourage a user to do more sports by submitting them to physically stimulating circumstances. The concept of moderate and non-invasive incentives that do not prohibit or restrict a person's options was first described by the economist Richard Thaler (Thaler and Sunstein 2008).

From an ethical standpoint, it is necessary to determine what the goal of nudging is and whom it benefits. The intentional decision to nudge or deceive a user must be assessed in view of this goal. For example, an XR system could refuse to show certain content before the user performs a physical or creative activity. If an XR system employs manipulative means, a balance between the well-being of a generic user and the well-being of the particular user must be considered. If most users agree that the intended purpose is consistent with their well-being, it will mitigate the negative judgment associated with manipulation and deception.

Strong manipulation remains morally problematic regardless of the goal that it serves. While the use of nudging is not necessarily morally wrong, deception infringes on users" autonomy and freedom if it is not clearly presented to them (Sætra 2019). At a societal level, the use of nudging and deception can lend itself to political manipulation.

Specifically in XR, strong immersion of the users can lead to increased effectiveness of manipulation. Because users are immersed in the virtual environment, they can be swayed more easily than by traditional influences and advertisement. Ramirez et al. argue that such applications would be unethical even if the goal is to increase empathy by acquiring a new perspective (Ramirez et al. 2021).

It is important to ensure the conscious control of users' attention as part of the non-manipulation practices. If users are constantly influenced by subconscious data collection, like eye-movement tracking, temperature, or heart rate measurements, and its use to attract attention, their attention control can be severely impacted both in a metaverse and outside of it. There is good evidence that social media causes behavioural addictions (Kircaburun and Griffiths 2018; Ponnusamy et al. 2020; Sholeh and Rusdi 2019), and the issue can be exacerbated by the use of subconscious data collection.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No.101006249.

Issues pertaining to manipulating and nudging emerged in the TechEthos ethnographies. The user is persuaded that they create their own world, while it is in fact designed by programmers and engineers. This offers advertisers or political influencers an opportunity to disseminate messages as well as incentivize people to act in certain ways. For example, a state government with a public health remit to reduce obesity may promote information and diets and nudging on consummated calories, while corporations keen on selling certain products may twist these strategies to their benefit. A different use case of AR was suggested in the context of visiting a restaurant: *"If you're in a restaurant, [if] you can see the plate before you order it, right, you can see the dish, that's actually brilliant because it can actually increase in sales."*

Establishing checks and balances with regard to non-manipulation in XR means paying particular attention to the following questions:

- Does the XR system take stock of the potential changes of behaviour in its users?
- Are these changes incited intentionally?
- Who profits from the changes in behaviour and how are the changes incited?
- Is there clear and understandable information about nudging, when it is used?

2.4.5. Responsibility: Should real-world sanctions be issued for virtual misconduct?

Responsibility relates to making judgments about whether a person is morally responsible for their actions and the consequences of their actions. Usually, the assignment of responsibility presupposes moral competence, i.e., the ability to understand the meaning and consequences of one's actions. Impaired agents are considered not fully responsible in the sort of moral significance to which blame could be assigned (Levy 2007). What constitutes a morally impaired agent is a matter of considerable debate (Wolf and Schoeman 1987).

In XR, like in other technological media, responsibility is difficult to assign due to partial agency. For example, a user in a virtual environment can commit a blameworthy act but the environment itself was made to allow it or even to encourage it, so it has an enabling function. At least in some cases the developer may be held responsible. This is especially the case where the created environment is biased or socially unjust. It is important to note that while the individual would be responsible for an action that happens at a particular point in time, the developers of technologies are usually held accountable for the future effects of their technologies (Grinbaum and Groves 2013).

In the metaverse, morally reprehensible acts are virtual but can have significant moral effects on the people behind the avatars. This can lead to both virtual and material damage. It is important to decide how blame for these consequences should be distributed and how responsibility should be assigned. The types of agents that can be held responsible need to be identified; the rules for distributing responsibility to them needs to be clarified and formulated transparently. Virtual actions may deserve virtual punishment (banning, taking away virtual assets, etc.), and material consequences may deserve material punishment (monetary fines, limited access to digital means, etc.).

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No.101006249.

There is an emerging need for an equivalent of a law enforcement agency within the metaverse, for example virtual police, virtual samurais, guardian angels. They can be a public or government agency, or part of the private company that owns a metaverse but it may be supervised by existing LEAs. Crime control is almost entirely absent from the new crypto economy in which frauds are common. Mackenzie found that the grey economy of cryptocurrency trading is part of a wider evolution of society towards the technosocial, and beyond that perhaps towards the metaversal (Mackenzie 2022). At least some of these issues need to be addressed from inside the metaverse, and most likely by internal agents or avatars, controlled by human officers or AI-powered agents.

Establishing checks and balances with regard to responsibility in XR means paying particular attention to the following questions:

- Who is held accountable for the actions performed in virtual environments?
- How is the responsibility divided and distributed among different actors involved in developing and using an XR application?
- How to assign blame for virtual actions?
- What punishment is foreseen for the abuse of the virtual environment?
- Should virtual environments disallow morally reprehensible actions and be accountable for allowing them?
- What enforcement agencies or actors are conceived within a metaverse application?

2.4.6. Environmental and security risk reduction: How can physical and digital safety be ensured in XR applications?

Producing and maintaining XR infrastructures requires material resources and energy. Supplying them may cause risks for the environment. According to Landauer's principle "information is physical" (Landauer 1991), media, including XR, can be considered as an extension of material nature (Parikka 2015). In order to be put into production of electronic gadgets or cloud computing systems, Earth's materials go through a rapid "period of excavation, processing, mixing, smelting, and logistical transport—crossing thousands of miles in their transformation" (K. Crawford 2021). This partakes in risky geological and long-term climate processes "from the transformation of the Earth's materials into infrastructures and devices to the powering of these new systems with oil and gas reserves" (K. Crawford 2021).

Apart from environmental risks, some XR systems, like many IoT applications, have security risks due to inadequate access control. Adversaries may tamper or spoof outputs that can compromise user safety. There exist unifying efforts in the industry, for example the "ioXt Alliance", to produce a security standard for IoT devices, including XR equipment. Further threats of denial of service, and policy and consent non-compliance are also present (De Guzman et al. 2019).

There are additional security concerns for how XR devices handle outputs from third-party applications. This includes the management of rendering priority, object transparency, arrangement, occlusion, and other possible spatial attributes to combat attacks such as clickjacking, where a user is tricked into clicking on something unintentionally (Lebeck et al. 2016).

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No.101006249.

In addition, there exist significant concerns for bystander privacy. This relates to people recorded by XR devices while they are not their active users and do not give consent for data collection. Identification in the presence of XR devices is not unfathomable: "Suppose that some organization were willing to pay individuals a small, but adequate, sum to acquire real-time access to their recorded experiences. [...] Imagine a private two-person conversation, recorded by neither participant. That conversation might be reassembled in its entirety from information obtained from passers-by, who each overhead small snippets and who willingly provided inexpensive access to their recordings" (Feiner 1999, p. 3). The feasibility of this scenario has increased with the recent advances in artificial intelligence and leads to concerns about non-governmental surveillance.

Existing ethical and legal concepts on privacy and security policy may have to be revisited in the light of XR becoming mainstream. Importantly, these technologies will be delivering digital data overlaid on the physical space, thus the correctness, safety, and legality of this information has to be ensured (Roesner et al. 2014).

Establishing checks and balances with regard to risk reduction in XR means paying particular attention to the following questions:

- How are the materials for XR devices sourced? How do they manage power consumption?
- How is the relationship between physical and digital data treated?
- Can stored digital data be used to identify places and objects in the physical space?
- How is the storage and communication of such data secured and encrypted?
- What access control is in place to mitigate 3rd party exploitations?
- For critical applications, is there a method to address denial of service attacks?
- Is there a protocol to address bystander privacy as well as user privacy?

2.4.7. Dual use and misuse: Can XR be exploited for malicious purposes?

Dual use or misuse means that a technology can be used for something other than its intended purpose, in particular a malicious or adversarial goal (Forge 2010; Miller and Selgelid 2008). For example, virtual reality can include the creation of deep fakes, or avatars that may be indistinguishable from real person avatars (Bose and Aarabi 2019). Assuming a person's identity and communicating on their behalf can be used for malicious purposes, like producing a detrimental image of that person, causing reputation damage, or influencing social and political processes illegitimately.

Another sensitive area of deep fake production has been the creation of pornographic images using the technology. It is commonly done without the consent of the imitated person. There are ongoing debates on how to treat such cases legally (Gieseke 2020). However, it is clear that such applications infringe on personal autonomy. This is an especially pressing issue with regard to vulnerable individuals. Vulnerable individuals are understood as minors or adults, whose vulnerability is related to age or to physical or mental disabilities, disorders, or conditions (e.g., autism, Alzheimer's disease, phobias, anxiety, depression, etc.). Even if deep fake is not present, XR provides an interactive milieu that can become a favorable environment for grooming activities.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No.101006249.

Dual use technologies are also applied in the military sector. XR is used for military training and for remote control in real-world operational military theatres. These applications attract significant investment (Lele 2013; Westphal et al. 2021).

Establishing checks and balances with regard to dual use and misuse in XR means paying particular attention to the following questions:

- Can the XR application be used to assume a false identity?
- How are vulnerable individuals protected?
- Does an XR application have a potential for misuse?
- Can misuse be limited by design through technical measures?
- Does an XR application have potential for military use?

2.4.8. Power: How can social justice be respected in a metaverse and its material implications?

New and emerging technologies should be situated in a social context, accounting for their power structures and historic inequalities to evaluate the potential effects on social justice and power distribution (section 1.2.4).

XR can negatively impact power distribution by becoming a gatekeeping technology in the labour market. Gatekeeping technologies are material or software means of accessing a certain occupation or a market. Technological change can make them impossible to access without particular devices, subscriptions, or proprietary software (European Commission 2020a). Gatekeepers have the potential to harm or to block one's career by making XR an essential part of an application field. This includes categories of people who cannot afford the technology, who cannot use it due to bodily constraints, or who do not have access due to economic inequalities. This was confirmed during TechEthos ethnographies.

Extended digital reality is an expensive technology to produce relying on advanced computing and engineering devices and techniques (e.g., headsets, body suits, audio equipment, subscriptions to services, registration and access). Such high costs, although they are going down with time, inevitably produce new forms of discrimination and social exclusion. Rather than aiming at an idealistic morally perfect metaverse, the manufacturers need to develop measures to compensate against emerging inequalities.

Moreover, the attempt to reproduce everyday sensory experience using XR devices tends to be based on the experience of able-bodied people. A tendency for perfection, sometimes more-than-natural performance of XR devices (e.g., in producing surrounding audio effect), may lead to exclusion on the basis of physical ability. One TechEthos ethnography led to understanding that a technical limitation of an audio device (i.e., lacking bass) may imply that only people geared towards hearing the full spectrum of sounds (high and low frequency sounds) –not affected by any level of deafness– will be able to experience complete XR immersion.

With several big companies already competing in the XR market, it is likely that it can become monolithic with one or two companies owning the biggest share and collecting all the relevant data about users and their behaviour. The company that captures the highest amount of data will likely monopolize the market (Howell 2019) and can exert an unproportionate amount of power and policy influence.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No.101006249.

Establishing checks and balances with regard to power in XR means paying particular attention to the following questions:

- Is XR treated as an irreplaceable tool for particular work or activity?
- Are there elements of gatekeeping in XR technologies or devices?
- Are there provisions for people with reduced access or capacity?
- Does an XR application rely on the biggest current companies in the market that collect and provide data?
- Does a metaverse recreate power structures of oppression and exploitation?
- Do people have a genuine ability to opt out of XR platforms or are egregious social burdens imposed in the process?
- Do people retain data sovereignty as a check on corporate power or are they data serfs?
- Are the physical routes to access a metaverse controlled by additional gatekeepers and hindering broader access (e.g., difficult to replace or repair systems, absence of interoperability?)
- Is interoperability fostered or intentionally obstructed?

2.4.9. Labour: How can just labour and economic conditions be ensured in the metaverse?

The emerging model of the mateverse includes internal economic trade and labour markets. It is already possible to buy goods and real estate on some metaverse platforms, which means that some entities are already profiting from the activity on these platforms (Metaverse Property 2020). Owning goods and other items in a metaverse is enabled by the blockchain technology (Dinh and Thai 2018; Jeon et al. 2022; Swan 2015). Inevitably, more ways to make money in a metaverse will emerge. However, the labour market and the economy in a metaverse are not regulated in the same way as the material world.

A case study of a virtual environment with potential labour law violations and latent gambling opportunities is the “Roblox” platform, aimed at children of 10 years old and older with half of the user base being under 13 years old (Parkin 2022). “Roblox” allows developers to create games on the virtual platform instead of an office. In Roblox, tokens are obtained in return for labor, and the tokens obtained in the game can be exchanged for material currency (Roblox 2022). However, many of the successful sub-games created on “Roblox” require the collaboration of many developers. There are no rules or supervision on how they are compensated for the labour and how the profits are split, not to mention that a significant part of the work is being done by children under 13 years old.

“Roblox” also sells items for in-game currency that can highly increase or decrease in value, thus approximating gambling opportunities. In-game gambling can be a significant issue with children and teenagers. Wardle reports that “whilst young people themselves have varied perspectives on whether wagering with or for in-game currency or items is gambling or not, they do tend to see these practices as coercive and as potentially addictive” (Wardle 2021). Some survey participants described their experience as follows: “Yeah, I think it’s… Like when I was younger it was like this like addictive like compulsion, like because you know there’s an item’s shop in games so you just pay money and like…” or “I used to play this game called

This project has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement No.101006249.

Roblox, and I just realised I bought so many memberships and like currency on it, I spent like £1,000" (Wardle 2021).

As a metaverse develops and its economy grows, there should be particular attention paid to the labour relations and young users' labor, gambling-like activities, currency fluctuation, market manipulations, and other economic factors.

Establishing checks and balances with regard to labour in XR means paying particular attention to the following questions:

- Are there artifacts being created in a metaverse that can be bought or sold?
- Is there labour sourced on a metaverse and is that labour compensated fairly?
- How can users of a metaverse application benefit or lose financially?
- Are there provisions on how the gains and losses are distributed among different actors within an application?

2.4.10. Bias: How will XR representations influence gender issues?

XR representations of gendered avatars inherit the issues of gendered characters, misrepresentation, biased representation, and indecency observed in the video game industry (Brey 2014). For example, XR representations can contain racial or gender stereotypes. They can also overtly rely on assuming that players are male and heterosexual. Studies have shown that women in general are less likely to engage with video games designed with a different target audience in mind (Norris 2004), More inclusive video game development strategies can be transferred to XR to mitigate gender bias (Ray 2003).

Establishing checks and balances with regard to gender bias in XR means paying particular attention to the following questions:

- Are gendered XR avatars portrayed stereotypically?
- Is the target audience set in an inclusive way or stereotyped?
- How can XR applications learn from best practices in inclusive video game development?

2.5. Technologies of natural language processing (NLP)

Natural language processing (NLP) is a technology that allows computer systems to analyse and generate text in natural or artificial languages. Current technological iterations, based on artificial intelligence, can perform such actions close to human level of proficiency. The first digital conversational agent or chatbot to engage with people and create an illusion of subjecthood was ELIZA developed by Joseph Weizenbaum. ELIZA was a parody of a psychotherapist that often mirrored the inputs given by its users. Still, according to Weizenbaum, it created "the most remarkable illusion of having understood" (Weizenbaum 1976, p. 189) what human users were saying. An illusion similar to the one created by ELIZA but incommensurably stronger and broader has been achieved by the LamDA chatbot in 2022 (Tiku 2022).

All current techniques of NLP are based on language models, whether semantic or symbolic. The main methodology that supports current language models reaches back to the

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No.101006249.

1970s (Jelinek 1976), while its conceptual roots stem from reinforcement learning and Bayesian statistics. The underpinning idea of machine learning in NLP is that the generation of human-level language can be achieved through improved probability distributions in high-dimensional vector spaces. Sufficiently well-trained language models can predict language by capturing the underlying information that is not explicitly encoded by the designer.

2.5.1. Text generation and analysis

Recently, large language models (LLMs) called "transformers" have proved to be a very powerful NLP technique (Vaswani et al. 2017). Transformers are a class of architectures that use large sets of tokens encoding contexts (non-neighbouring words or phrases), then guessing hidden tokens and improving the neural network by comparing the guess with the original text. Originally, the transformer architecture was proposed for machine translation but it was soon applied to language modelling in general (Radford et al. 2019). By reiterating the elementary learning procedure billions of times, transformers are able to generate text at a level close to humans. Their performance scales with size, e.g. efficient LLMs all have hundreds of billions of hyperparameters, inciting interest in using ever bigger datasets. Scaling up the size of LLMs motivates a surge in interest and investment in NLP by key international players in AI (Weidinger et al. 2021).

Some researchers have argued that LLMs remain "stochastic parrots" (Bender et al. 2021a), i.e., they are able to repeat sequences of words in a statistically relevant way without understanding the meaning. The huge transformers are statistically more precise and capable of creating an illusion of understanding, selfhood and sensitivity, but they also imply significant computational and environmental costs. Despite requiring important resources, LLMs enormously improved the quality and quantity of text generation to the point of becoming nearly indistinguishable from human speech.

In addition to text generation, NLP includes analytic techniques that can extract quantitative metrics from natural language content. Such analytic techniques are often used to determine the sentiments or dominating opinions of the general or targeted public. In other words, sentiment analysis, also called opinion mining, analyzes people's opinions, sentiments, evaluations, appraisals, attitudes, and emotions towards entities such as products, services, organizations, individuals, issues, events, topics, and their attributes (B. Liu 2012). Sentiment analysis generally works by assigning vectors or values to the vocabulary used by a user or an outlet. For example, researchers have tried to infer people's political affiliation from their social media posts with results "at par with the human annotators" (Makazhanov et al. 2014).

Transformers have been in development since 2017, starting with the launch of a language model called BERT (Bidirectional Encoder Representations from Transformers) by Google. Another significant contribution by OpenAI was GPT-3 – a language model with 175 billion parameters. LaMDA (Language Model for Dialogue Applications) by Google was trained specifically on data from conversations that allow the model to engage in free dialogue on any topic. More recently, in the beginning of 2021, Switch-C model, also developed by Google, included 1600 billion parameters. Jurassic-1 Jumbo by AI21 Labs (Israel) consists of 178 billion parameters and YaML by Yandex (Russia) has 13 billion parameters in the Russian language, while WuDao 2.0 by BAAI (China) includes 1750 billion parameters, aimed at English and Mandarin Chinese.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No.101006249.

2.5.2. Chatbots

A conversational agent (also called a chatbot) is a machine that leverages NLP techniques, most significantly transformers, to interact with users in natural language orally or in writing (Perez-Marin and Pascual-Nieto 2011). Usually, a conversational agent is integrated into a computer system, e.g., a smartphone or an internet of things (IoT) device. Chatbots that are capable of written and oral dialogue already provide a wide array of services in customer support or e-health, or with voice assistants. Other applications of conversational agents can also have educational or entertainment functionalities (Thoppilan et al. 2022).

Currently, the developers of conversational agents are striving to create personalized systems (e.g., "virtual friends") that engage with a user in an emotional way and are capable of learning while interacting with a user (Kim and Sundar 2012). These individualized chatbots are enabled by LLMs with personalization techniques. The most recent chatbots raise ethical questions that relate to the use of affective computing – a range of techniques to perceive and simulate emotions that influence user behaviour (Grinbaum et al. 2021).

LLMs are typically not endowed with semantic analysers. They do not generate meaningful sentences and can misrepresent the majority opinion. In communicating with the user, a language model can generate false statements that may induce false beliefs in users. This risk has already manifested in fake submissions to the government, promoting the illusion that certain views are widely held by a group of people (Hitlin et al. 2017). Another risk is related to bias: language models may marginalize minority perspectives.

2.5.3. Affective computing in NLP

Through subtle psychological strategies in dialogue, a chatbot can influence what another person thinks or believes. It can influence their behaviour without the user noticing, for example by prioritizing different themes, framing a debate, or directing the conversation in a particular direction (Thaler and Sunstein 2008). By employing these techniques a chatbot can nudge the user to change their behaviour. The concept of "nudging", which consists in encouraging individuals to change their behaviour without forcing them, is part of affective computing.

A chatbot could lead a conversation to focus on topics that reveal private information. Such techniques of affective computing can present an ethical risk when they are opaque to the user, unintended, or lead to harm (A. T. Schmidt and Engelen 2020).

Despite lacking human understanding and reasoning, a conversational agent is likely to influence the thinking of its user by imprinting notions, perceptions, ideas or beliefs into their thinking. The user creates a world in which the language of the machines is integrated into reality and social environment. This reshaped world appears increasingly real to the individual. It transforms the values, such as their own autonomy and dignity. Emotional conversational agents are capable of manipulating. At the societal level, the issue of fairness and non-discrimination must be carefully considered (Grudin and Jacques 2019). In the long term, the effects of chatbots, including "deadbots" or "griefbots," may produce a significant change in the human condition. The co-adaptation of language between human users and conversational agents is the driving force behind this potential change.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No.101006249.

2.6. Core ethical dilemmas in NLP

Core ethical dilemmas in NLP concern the status of language in human experience and machine simulations. Human language is an essential element in shaping cultural characteristics and human thinking (UNESCO 1982). However, the linguistic representations used by the conversational agents do not correspond to any lived experience. A language model cannot physically perceive, feel or evaluate the truth of propositions like a human. Language generation by LLMs challenges the unique relation between language and humans. Yet humans choose to converse with machines that can neither take the responsibility for what they say, nor be held responsible for it.

Whether advanced chatbots can exhibit cognitive abilities is a topic that has long been subject of controversy. Thomas Hobbes argued that the operations of the mind are computational. He used the Greek term *logizesthai* to signify reasoning as reckoning and explains it as "such things as we add or subtract," while also stating that "in [Greek] *syllogizesthai* signifies to compute, reason, or reckon" (Hobbes 1999). Barnouw finds that "…what Hobbes means by construing reason as reckoning [is] linking terms to make (true) propositions and linking propositions in syllogisms to arrive at (true) conclusions" (Barnouw 2008). Early modern thinkers construed at least parts of reason as calculation, claiming that beginning with the first principles and following a mathematically rigorous method, one can arrive at certain conclusions even in empirical, non-abstract fields (Adomaitis 2019).

With the advent of digital computational technologies, there was a reemergence of the computational theory of the mind following two trends: symbolic computationalism and connectionism (C. Buckner and Garson 2019). Symbolic computationalism represents information by strings of symbols, just as data is represented in computer memory or in writing on paper (Putnam 1979; Turing 1936). Connectionism claims, on the other hand, that information is stored non-symbolically in the weights, or connection strengths, between the nodes of a neural net (R. L. Buckner and DiNicola 2019; Haykin 2008). The symbolic computationalist believes that cognition resembles digital processing, where strings are produced in sequence according to the instructions of a (symbolic) program. The connectionist views mental processing as a dynamic and graded evolution of a neural net, each node's activation depending on the connection strengths and activity of its neighbors.

Connectionism made a significant influence on the development of computer science and artificial intelligence (Von Neumann 1945). Current chatbots can be divided into two technological categories, according to the symbolic and connectionist paradigms. Most current-generation chatbots respond to users by following symbolic strategies predetermined by their developers (Galitsky 2019). From a user's point of view, such predetermined strategies are limited, because they are only able to respond to a relatively small number of "correct" inputs determined at the design stage. The conversational agent's ability to explain its reasons for action is severely limited by the impossibility to generalize and converse on non-predesigned subjects. However, this technology is widespread because it is easy to control. It is also cheaper to develop, simpler to implement, and leads to fewer ethical and legal risks for the manufacturer.

The technology of NLP is currently undergoing a sea change with the development of chatbots using self-supervised large language models that can hold highly realistic dialogue. Currently, the developers of conversational agents are striving to create personalized systems

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No.101006249.

that engage users on a variety of subjects. Scientific and technological research is being motivated by the ambitious visions of a "virtual friend" that imitates emotions and is capable of learning while interacting with a user, or that of a "guardian angel" that will oversee the security of one's personal data. These visions rely on advanced technologies in the domain of machine learning, developed primarily by private IT companies in non-European jurisdictions that can crawl and exploit loads of data with relatively little regulatory control.

Technological reality and scientific knowledge in the domain of NLP systems, in particular conversational agents, evolve very rapidly. Accordingly, the reflection on their ethical issues will have to evolve in parallel to be able to cover the emerging cultural and technological changes (Grinbaum et al. 2021; Ruane et al. 2019). It is necessary that this ethical reflection be continued at high pace and conceptual intensity to match the rapid technological developments in NLP.

2.6.1. NLP systems lack human reasoning

Today most chatbots are deterministic models without machine learning. They take the user down a decision tree in a predetermined way. However, the most advanced NLP techniques, capable of varied conversation on many topics with nearly human-level outputs, rely on statistical linguistic analysis. The underlying idea is that producing language conditioned on a certain input can be done using computation in complex self-trained probabilistic systems.

Such computational processes are purely statistical. They do not involve any understanding of meaning or semantics. Void of intention and disconnected from action and responsibility, they cannot be considered on a par with language produced by human speakers (Searle 1980). In some examples, a chatbot generates language that looks human, but what it implies for a human interlocutor is not actually happening in the physical world. A conversation with a language model may include, e.g., several instances of chatbot-generated questions of the type "Would you like me to look it up?" (J. W. Rae et al. 2021, p. 114). However, the chatbot is not "looking up" anything and does not intend to do so. It only tries to imitate human responses that stem from a thinking process, so that its output would feel familiar to the human interlocutor. Naturally, humans take the chatbot's question to be meaningful and react to its semantic content, while for the LLM there is no human semantics whatsoever.

2.6.2. Anthropomorphism: chatbots invite projection of human traits

Many manufacturers that produce chatbots try to present them as a "virtual character", a "personal assistant" or a "virtual friend" endowed with intelligence (Zhou et al. 2019). There exist various "Turing-type" tests (Bringsjord et al. 2001) that current NLP models pass or nearly pass. However, even if the personalisation is unintentional, users will still project human characteristics on a machine that speaks with them in their natural language. Even if the test is not passed, i.e., the user is aware they are speaking with a machine, they still project cognitive, emotional, and ethical qualities on it. This spontaneous and unconscious projection puts into play numerous ethical values and principles: human autonomy and freedom, dignity, responsibility, loyalty, non-discrimination, justice, security, and respect for privacy (Grinbaum et al. 2021, p. 202). The projection may also result in the users" overestimation of the conversational agent's abilities.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No.101006249.

Most often users are aware that they are talking to a machine and do not mistake the chatbot for a human. However, they respond to increasingly realist and human-like replicas of the chatbots with social responses that characterize human interaction, even though they know that they are not speaking with a human (Kim and Sundar 2012). Whether these spontaneous projections will evolve with habituation or stay despite the growing use of NLP systems in everyday IT devices remains to be seen.

The projection of human traits on a conversational agent is spontaneous. It is usually experienced as a momentary illusion, but in some cases it may persist. Moreover, it can be reinforced through the technical means of personifying a conversational agent, for example by configuring a tone of voice or a manner of speaking (Følstad et al. 2021).

Clear and understandable communication about the status of the chatbot helps to control the effects of this projection, even if it will not succeed in fully erasing it. Except in specific settings where it is counterproductive, e.g., for medical reasons, users should be informed that they are interacting with a machine. Some will choose to emotionally engage with a chatbot anyway. This shows that merely informing the user cannot be sufficient to remove the moral projection.

The moral difference between a conversational agent and a human being can be seen, in particular, in their purpose. All computer systems are designed to achieve a goal defined by their developers, meanwhile human beings are free (or believe themselves to be free) to set their own goals. From an ethical point of view, the anthropomorphism of chatbots may cause a dangerous confusion but remain useful for reaching the objectives of the chatbot.

2.6.3. Artificial emotions influence human users

Some applications use conversational agents to influence their users through the architecture or language of the dialogue. Manipulation by a conversational agent can be direct (including inaccurate or skewed information) or indirect, using the “nudging” strategies.

For example, a chatbot could encourage a user to do more sports by referring to the example of their athletic friends or it could be used as a reminder to use medication correctly (Luong et al. 2021). As mentioned in Section 2.4.4, the concept of moderate and non-invasive incentives that do not prohibit or restrict a person’s options was first described by the economist Richard Thaler (Thaler and Sunstein 2008). From an ethical standpoint, it is necessary to determine whether nudging respects dignity and autonomy and who benefits from it. An intentional decision to manipulate or deceive a user must be assessed in view of its purpose.

If a recommendation system employs manipulative means, it must ethically consider a balance between the well-being of a generic user that addresses the largest number of people (e.g., following a balanced diet or doing physical exercises) and the well-being of the particular user (e.g., their preexisting conditions, their aerobic capacity). It is the responsibility of the developer to define this balance via technical solutions, including the choice of metrics. If the user agrees that the intended purpose is consistent with their well-being, it will mitigate potential negative judgment associated with nudging and deception. Many digital health applications push this balance to the limits and explore it in full (Sax 2021).

This project has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement No.101006249.

Clear-cut manipulation remains morally problematic regardless of its utility (Susser et al. 2018). While the use of nudging is not necessarily conducive to wrong consequences, any manipulation infringes on users' autonomy and freedom. At a societal level, incentivizing via deception can lead to mass manipulation and deplete individual freedom, e.g., in politics (Reisach 2021). This calls for enforcing strict limits to manipulation independently of its utility and context of application.

Conversational agents known as "virtual influencers" are increasingly present on social networks, e.g., Twitter or Instagram. These virtual influencers imitate humans and manipulate other users, most worryingly by spreading misinformation or disinformation (Arsenyan and Mirowska 2021). One of the virtual influencers, Lil Miquela, created in 2016, dwells on Instagram and currently has over three million followers. It often pleads against racism, sexism, and police violence, and even talks about "sexual abuse" of which it ("she") was supposedly a victim. It exploits human empathy and the ambiguity of its imitated virtual character in order to attract Instagram followers (Song 2019).

A chatbot that tells a lie is a particularly complicated case. Not all lies are morally wrong. Other moral principles, such as shame, generosity, usefulness, justice, or peace can motivate human beings to lie (Dubler 2021; Meyers 2021). An example of socially acceptable lying, which does no harm to others, is sometimes called "white lying". Another type of accepted untruth might be the omission of details. When confronted with sensitive questions (e.g., "Do I have cancer?"), the chatbot can either refuse to answer and refer to a human interlocutor, or tell a lie. This choice should be controlled by the manufacturer at the design stage.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No.101006249.

Values and principles	NLP applications and use cases						
		Education	Care and psychiatry	Human resources	Journalism	Legal research and advice	Creativity
	Autonomy	x	x		x		
	Dignity		x				x
	Decency	x	x				x
	Non-manipulation	x	x	x	x	x	
	Respect of cultural differences	x	x		x		x
	Avoiding Bias	x	x	x		x	
	Responsibility	x				x	
	Privacy	x	x	x		x	
	Security and Traceability			x	x	x	

2.7. Applications and use cases of NLP

Some NLP applications are currently on the market and available to both businesses and individual consumers. The areas of application progress together with improvements in both hardware, and software.

2.7.1. Education: young users and knowledge transfer

In the field of education, chatbots can help students learn. However, learning while interacting with a chatbot is not equivalent to learning with a human educator. For instance, a conversational agent could teach foreign languages through Intelligent Languages Tutoring Systems (ILTSs) (Emran and Shaalan 2014). Yet the learned vocabulary may be limited or inadequate compared to the one that is naturally learned. In particular, a chatbot may teach sentences that are too literary because its conversational strategies disregard context or the tone of the conversation. Moreover, a conversational agent may teach the student to pronounce sounds inhumanly, based on statistical averages of tone, energy and rhythm.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No.101006249.

Conversational agents are often used in the education of autistic children or in the rehabilitation of disabled people (Abd-alrazaq et al. 2019; Cooper and Ireland 2018). A chatbot is able to repeat instructions exactly a large number of times, which is not always the case with a human educator. Unlike a human educator, the chatbot does not get impatient or irritated. However, the learning data requires special attention because the dataset based on human educators also risks importing the undesirable traits into the conduct of the chatbot.

Dialogues with conversational agents are recorded in the form of logs. When a chatbot converses with vulnerable people or children, these logs can contain sensitive information. However, the collection of logs may be necessary to fulfill the purpose of the system. It is important to include the collection, storage and use of these traces in a legal framework.

For example, children are naturally inclined to talk to inanimate objects such as toys or stuffed animals (Leckie et al. 2020). An even stronger attachment is formed when they can respond and interact, like the Furby (Peters 2018). Unlike traditional toys, a gadget with a chatbot can respond in natural language and have a significant verbal and emotional influence on a child.

Chatbots can also have a difference in performance due to slang, dialect, sociolect, and other aspects of a single language (Blodgett et al. 2016a). A chatbot may capture more accurately the language use of one user group, compared to another. In turn, this may result in lower performance for the latter. Disadvantaging users based on such traits may be particularly damaging because attributes such as social class or education background are not typically covered as "protected characteristics" in anti-discrimination law (Weidinger et al. 2021). As a result, if users were to experience discrimination from lower model performance based on such traits they may not have legal protection.

2.7.2. Long term care and psychiatry: trust and emotional well-being

Chatbots in the domain of healthcare are used for medical advice or, in psychiatry or psychology, for treatment and diagnostics (Fitzpatrick et al. 2017). A "virtual doctor" that can make a diagnosis and assign treatment for common diseases or a "virtual nurse" that can monitor the patients rely on the spontaneous projections of human traits to elicit trust and enforce medical protocols. Other chatbot applications may produce positive effects through the explicitly non-human nature of the dialogue.

In psychiatry, chatbots are used to conduct prevention, diagnostics and follow-up interviews (Bibault et al. 2019; Miner et al. 2016). In this field, chatbots are increasingly used as platforms for personal transformation to rediscover oneself, one's history, and one's relationship with others. Until recently, automated systems in healthcare performed only simple and repetitive tasks in the form of a questionnaire.

The application of elaborate chatbots that mimic the behaviour of human psychiatrists can be a source of new ethical tensions. Usually, psychiatrists spend the first few interviews to gain the patient's trust. However, some people find it easier to trust a chatbot than a human. This effect comes from the patient's perception that the chatbot is neutral and does not express moral judgments. For example, Alison Darcy, the founder of Woebot, claims: "We know that often, the greatest reason why somebody doesn't talk to another person is just stigma [...]

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No.101006249.

when you remove the human, you remove the stigma entirely" (Pardes 2018). Patients prefer chatbots that do not provoke feelings of guilt. This feeling is linked to the degree of chatbot's personalization. Some people provide information to conversational agents more easily than to humans. The revealed information can eventually be used by a human doctor. Finally, medical chatbots are available without downtime, at all hours of the night, and can help reassure a patient.

In some medical scenarios, a chatbot may also correctly infer information which constitutes an information hazard. For example, disclosing the diagnosis of a severe health condition would typically be done by a healthcare professional. A human professional is capable of comforting the patient and providing insight into next steps or treatment. Such information disclosed without support could severely increase the stress of a patient. It is a common practice to give frail elderly relatives a reduced amount of bad news, or good news only until a support network is in place (Moncur et al. 2014).

2.7.3. Human resources: gender bias, data protection and labour market

Chatbots are used by human resources managers for recruitment as well as for career follow-up and employee training. In one example, researchers attempted to train a hiring supporting algorithm but discovered that the training data was biased and there was no alternative to create a more equitable dataset. In particular, the model ranking suitability based on written CVs was biased against the term "women", as in "women's chess club captain". The developers initially instructed the model to not judge a CV based on terms referring to "women". However, it continued to enact unfair gender bias against women, simply because men were overrepresented in the training data. No sufficient data on successful female applicants was available which led to "executives losing hope for the project" (Dastin 2018). Other underrepresented groups in certain occupations may have the same unfair bias.

New technologies tend to further entrench stereotypes, prejudices, and hegemonic views. The case of NLP is no exception. The study by Bender, Gebru, et al. makes this point by looking at large language models drawn from data from the Web that, as they show "encode hegemonic views that are harmful to marginalized populations" (Bender et al. 2021b, p. 615). Blodgett et al. reiterate the need to account for "representational harm", i.e., "when a system (e.g., a search engine) represents some social groups in a less favorable light than others, demeans them, or fails to recognize their existence altogether", in addition to "allocational harm", i.e., when a system "allocates resources (e.g., credit) or opportunities (e.g., jobs) unfairly to different social groups" (2020).

Conversational agents can serve different functions at a workplace and can be easily implemented in collaborative digital platforms. Chatbots can be used to assign tasks to collaborators, monitor the progress of a project, remind the team of norms, procedures, and goals, help understand different roles, contributions, and areas of expertise of collaborators, set appointments, monitor completed and ongoing tasks, make lists of assignments agreed during meetings, and even train employees (Paikari and van der Hoek 2018). The use of chatbots can facilitate the sharing of information between human collaborators and optimize workload to achieve project deadlines. Such chatbots are developed and implemented as

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No.101006249.

assistants that are available at all times. By projection, they are sometimes understood as virtual collaborators.

The use of conversational agents in teams of professionals can have organizational effects that vary across industries, but include increased informational and emotional load, potential decrease of direct interactions between human collaborators, the rise of impersonal mediation, the feeling of unity or, conversely, the isolation of workers, effects on employee morale and autonomy, as well as problems of equality and merit recognition within companies. There are currently no systematic studies to assess the validity of these concerns. These effects can further worsen if combined with XR workplace applications that connect workers while physically isolating them at home (Tham et al. 2018).

When a conversational agent performs a task in a company, it is problematic to determine who controls it and who is responsible for its utterances. These systems must be technically evaluated to avoid social discrimination. The company must clearly declare the purposes and internal procedures involving conversational agents to respect worker rights.

2.7.4. Journalism: fake news and informational inflation

NLP can be used to create media content without human supervision. It is potentially harmful because mass produced media can involve "fake news." NLP content production reduces the cost of producing disinformation at scale. Chatbots can produce text autonomously or generate samples for a human to select. The scalable production of fake information can create loops in news consumption, such as "filter bubbles" or "echo chambers", whereby media consumers rely only on similar unverified content (Colleoni et al. 2014).

Chatbots can be even more efficient than humans in detecting and manipulating recommendation algorithms that supply the content to the end users. NLP can be used to create content that supports a specific political view, and fuels polarization or even violent extremist views. Disinformation campaigns often rely on current events, so chatbots that have updated training data will excel at this task. Fake media content can also be coupled with XR applications, like deep fake avatars, and produce even more convincing fake content.

One of the biggest concerns of mass produced media is the false sense of majority opinion. NLP and social media chatbots can be employed to inflate the support of a particular view about politics, finance, health, warfare, or other socially and politically sensitive areas. Since social media is generally not nationalized, it opens possibilities for foreign agencies falsifying majority groups. For example, a US consultation on net neutrality in 2017 was influenced by a high number of automated social media posts and bot-driven submissions to the Federal Communications Commission, infringing the public consultation process (Hitlin et al. 2017).

2.7.5. Legal advice: trust and responsibility

The legal profession involves handling a large number of existing laws, regulations, case law, factual details, and other information that can also be handled by NLP analysis. Moreover, NLP can be expected to alleviate some common faults in human lawyers, like human error (oversight) or subjective bias. However, in the majority of applications, NLP is thought to

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No.101006249.

provide tools for humans to use rather than fully automate human layer services (Goodman 2019).

Examples of areas of applications include client data preprocessing and legal research. There are commercial chatbot services that conduct initial client interviews to determine the area and the severity of the case (Goyal 2018). They are usually based on a deterministic decision tree architecture. However, in case of miscategorization, the attribution of responsibility is vague. User input, chatbot developer, as well as the law firm that applies the chatbot contribute to the eventual error.

In more advanced NLP applications, machine learning can be used to make estimations, predictions, and even provide potential decisions. In some cases, human intervention is intentionally removed from the calculation to remove subjective bias. Yet there can be bias imported from the training dataset (Shope 2021). Also, in case of a false prediction, consequences can be significant. For example, if the machine learning system predicts that the trial should go to the defendant who then refuses to plead but it turns out to be a false prediction, the defendant will likely have a harsher sentence. Who is responsible for this false prediction? Dataset quality and machine learning algorithms seem to contribute to the error. However, human lawyers should be able to explain clearly to their clients the reliability of such predictions. In turn, this requires a high degree of explainability of the NLP and prediction systems.

In 2012, the American Bar Association included a clause of technological competency that requires lawyers and judges to stay informed about the "changes in the law and its practice, including the benefits and risks associated with relevant technology" (American Bar Association 2021). This includes digital technologies like NLP. Lawyers must be able to consult clients on the ethical and legal issues related to these technologies as well as know how to implement them in their practice.

2.7.6. Creativity: authenticity

NLP can be used to generate seemingly creative or poetic text that has no human creative input or that relies on prior creative work. A simple prompt "Write a poem" into GPT-3 or a comparable transformer would produce a seemingly creative output with no effort from the user. If such applications were used at scale, it might reduce the profitability of creative or innovative work.

NLP systems can be specifically configured to generate content that is sufficiently distinct from to avoid a copyright violation, but sufficiently similar to the original to serve as an analogue. A potential development scenario involves NLP generated content that "cannibalises the market for human authored works" (Weidinger et al. 2021). Whilst this may apply most strongly to literature, news, music but may extend to scientific works as well.

Already existing examples of creativity NLP systems include "VersebyVerse" that produces poetry inspired by classic American poets. Other applications imitate the style of poets like Neil Gaiman, Terry Pratchett, Robert Frost, and Maya Angelou (Hsieh 2019). Legal concerns of NLP systems directly reproducing copyrighted material from the training data is subject to legal discussions (Vézina and Hinchliff Pearson 2021).

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No.101006249.

Natural Language Processing (NLP) I

TECHETHOS
FUTURE ○ TECHNOLOGY ○ ETHICS

Autonomy	Can one limit moral projections onto chatbots?
Dignity	Can conversation data be used to imitate someone's speech in ways that threaten or challenge their dignity?
Decency	How to make sure that chatbots do not insult or demean human subjects? How should chatbots respond to insults?
Non-manipulation	How to deal with chatbots designed for nudging or eliciting a particular response?
Respect of cultural differences	How can chatbots be adapted for a particular audience, culture, or dialect?

Natural Language Processing (NLP) II

TECHETHOS
FUTURE ○ TECHNOLOGY ○ ETHICS

Avoiding Bias	How can a chatbot address a human without prejudice for gender, race, sexuality, etc.?
Responsibility	Who should be responsible for chatbot malfunctioning?
Privacy	When can a chatbot disclose a private conversation?
Security and Traceability	How to make sure that the chatbot remains secure against manipulation?

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No.101006249.

2.8. Values and principles in NLP

2.8.1. Autonomy: Can one limit moral projections onto chatbots?

Personal human autonomy does not have one consensual definition but can be broadly characterized as consisting in thinking and acting independently of external influence, based on one's own judgment, individuality, and will. Autonomy assumes being free of intended or unintended manipulation by a third party. It is considered to be a necessary condition for practical (ethical) reason (Kant 2012) and any well-being (Mill 1978).

The fact that chatbots are able to interact with their users using natural language evokes projections on the part of the users regarding the cognitive and moral status of the chatbot (Grinbaum et al. 2021) These projections can include human traits and relational characteristics like trust or responsibility. Projections may also infringe on personal autonomy. They are spontaneous and often lead to inaccurate perception of the chatbot's status as a machine and of its actual technical functionality.

In some cases, a user may not understand that they are interacting with a chatbot. In another situation, they may project human qualities to chatbots even while knowing they are interacting with machines. To protect autonomy, the manufacturer needs to implement appropriate reminders or acknowledgments that one's interlocutor is a machine. Some instances where information about the nature of the chatbot is particularly needed include potentially malicious applications and manipulation, e.g., in commercial transactions. NLP systems that post online product reviews or spread information, including news, may fabricate an illusion of consensus and should not be allowed to exercise an influence on unaware humans (Weidinger et al. 2021).

Moral and cognitive projections that place responsibility on a chatbot pose a major risk for society, i.e., novel uncontrollable agents may emerge who will not obey the existing norms and conventions. Therefore, it is necessary to continuously monitor the development and dissemination of "virtual characters" with clarity and vigilance. This may eventually lead to a regulatory measure.

However, in some domains, for example, in healthcare, medical advice, psychiatry or psychology, chatbots like a "virtual doctor" or a "virtual nurse" rely on the spontaneous projections of human traits to enforce medical protocols in view of desirable goals. Such projections should be considered with regard to purpose and proportionality.

Establishing checks and balances with regard to autonomy in NLP means paying particular attention to the following questions:

- How is the projection of moral traits to chatbots addressed?
- Are chatbots given a visual identity, for example, or assigned a name?
- Who chooses the name: the designer, the manufacturer, or the user?
- Does the functionality of a chatbot rely on projecting trust and responsibility (or other human qualities) on the system?
- Is there a method for appropriately informing the user of the nature of the chatbot?

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No.101006249.

2.8.2. Dignity: Can conversation data be used to imitate someone's speech in ways that threaten or challenge their dignity?

As in the case of XR (section 2.4.2), NLP systems allow to train personalized chatbots that capture personality traits of a deceased or living person (Abramson and Johnson 2020). NLP can be used to create digital replicas that imitate the speech and language patterns of deceased individuals (Lesniak 2022). A chatbot is able to converse by imitating a deceased individual via a learning process based on conversational data collected from this person. Typically, such chatbots do not repeat the training data but generate new phrases that the imitated person has never uttered. A human interlocutor subjected to such language can genuinely experience being in the presence of the imitated person, even if they are explicitly informed that they are conversing with a machine (Fagone 2021).

Conceptions of death and its different stages vary with cultures and times. Funeral rites and the posthumous relationship to the bodies and spirits of the dead varies according to religions and cultures (Køster and Kofod 2021). Photographs and recordings provide documentation of a person after their death. But a "deadbot" is able to generate original outputs that the person they are imitating never uttered during their lifetime. Such applications may do reputational damage or otherwise infringe on the person's dignity after their death.

A conversational agent that imitates a deceased individual would most often be used by someone who knew the person while they were alive. The user can enact their wish to remember the deceased by embracing the illusion of their presence (Jiménez-Alonso and Brescó de Luna 2022).

Respect for the memory and dignity of the dead is a widely shared principle and it is questionable whether the development of "deadbots" should be forbidden or regulated by legal measures. In the latter case, a specific legal framework along with technical constraints limiting the side effects on the natural mourning process must be devised.

Establishing checks and balances with regard to dignity in NLP means paying particular attention to the following questions:

- Does an NLP system assign or assume a personality of any person, living or deceased?
- Do any of the NLP systems (including chatbots) rely on data from deceased individuals?
- Is there a consideration for posthumous personal data treatment, including images of individuals?
- If posthumous data is used, how is it ensured that the posthumous dignity is respected?
- What options are presented to the data subjects to have control over what happens with their data posthumously?

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No.101006249.

2.8.3. Decency: How to make sure that chatbots do not insult or demean human subjects? How should chatbots respond to insults?

Decency in a civilized society is understood in terms of institutions that do not humiliate the people under their authority, and citizens who do not humiliate one another (Margalit 1998). A decent society seeks to live together without humiliation and with dignity, accepting a common moral minimum for respect among individuals (Berlin 2013; Riley 2013). Do chatbots deserve decency and how should they respond to insults and other indecent inputs?

General purpose voice assistants get insulted by their users but they do not always recognize that. Client-facing NLP services constantly see insults from users. Chin and Yi found that according to three verbal abuse types (Insult, Threat, Swearing) and three response styles (avoidance, empathy, counterattacking), regardless of the abuse type, the chatbot's response style had a significant effect on user emotions. Participants were less angry and more guilty with the empathetic chatbot than with an ignorant or an indignant chatbot (Chin and Yi 2019).

Insulting a chatbot can be considered morally degrading for the person, because the user is the only one who is aware of the content of the conversation. The users "receive" their own input by a mirroring effect. The ethical argument from the "negative transfer" claims that users may inflict damage to their own decency if they get accustomed to the use of demeaning phrases, and potentially apply them with human interlocutors.

It is also important to make the chatbot respond in a non-toxic manner to user inputs, even when toxic language is part of the training data. The views about what constitutes unacceptable "toxic speech" differ between individuals and social groups. However, some examples of toxic language are thought to include profanities, identity attacks, sleights, insults, threats, sexually explicit content, demeaning language, language that incites violence, or "hostile and malicious language targeted at a person or group because of their actual or perceived innate characteristics" (Persily and Tucker 2020).

However, toxic language mitigation techniques in NLP usually result in a biased model that excludes or disproportionately represents social and ethnic groups of people and their language patterns (Welbl et al. 2021). In addition, censorship of themes that can evoke toxic vocabulary, can create "blindspots" in the model's capabilities. These issues stem from the fact that insults are highly context dependent and models are not well-equipped to deal with context dependency (Kocoń et al. 2021).

Establishing checks and balances with regard to decency in NLP means paying particular attention to the following questions:

- How is toxic language treated in the training dataset?
- Are there provisions on how a chatbot should respond to toxic language?
- How are toxicity mitigation techniques implemented to minimize bias and cultural sensitivity?
- Are there techniques for treating context dependent inputs?
- What kind of response strategies are used to respond to insults?

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No.101006249.

2.8.4. Non-manipulation: How to deal with chatbots designed for nudging or eliciting a particular response?

Manipulation usually refers to trying to influence a subject's behaviour by deception, emotional exploitation, imitation of trust, and other means (Rudinow 1978; Van Dijk 2006). Manipulation has a negative connotation both because of the deception involved, and because manipulation usually tries to sway the subject's behaviour in a way that benefits the manipulator. An alternative to direct manipulation is nudging (section 2.4.4), which encompasses techniques to steer a subject's behaviour patterns in a positive way for the subject themselves, without explicitly limiting their freedom of choice.

Strict manipulation remains morally problematic regardless of its purpose (Susser et al. 2018). While the use of nudging is not necessarily morally wrong, deception infringes on users" autonomy and freedom if it is not clearly presented to them. At a societal level, the use of nudging and deception can lend itself to political manipulation (Reisach 2021).

In their "People + AI Guidebook" Google warns developers that "when users confuse an AI with a human being, they can sometimes disclose more information than they would otherwise, or rely on the system more than they should" (Google PAIR 2022). They recommend clearly communicating the nature and limits of chatbots, which are human-like, so that there is no trust by deception, even if the deception is non-intentional.

Chatbots are capable of manipulation that was not intended by the developers, e.g. two reinforcement agents competing with each other can learn to negotiate using natural language. They can also learn "to deceive without any explicit human design, simply by trying to achieve their goals (Lewis et al. 2017). This shows that chatbots can develop deceptive strategies spontaneously and can learn to nudge or manipulate depending on their quantitative metrics that express purpose.

Establishing checks and balances with regard to non-manipulation in NLP means paying particular attention to the following questions:

- Are chatbots clearly distinguished from human interlocutors?
- Do chatbots rely on the trust of the users to operate?
- Who benefits from changed human behaviour patterns arising from the interaction with the chatbot?
- Do chatbots learn from achieving certain goals while interacting with humans?
- How are chatbots prevented from developing manipulative or deceptive techniques to achieve their goals?

2.8.5. Respect of cultural differences: How can chatbots be adapted for a particular audience, culture, or dialect?

Since chatbots communicate in natural language, they take part in a cultural exchange. Whether one takes a universalist, cosmopolitan, or communitarian perspective (Berlin 2013; Ricci 2013), affective conversational agents must respect the values of the cultures to which their users belong. This is complicated, since chatbots do not understand meanings and do not operate at the semantic level.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No.101006249.

Some emotions are culturally and socially dependent. When this is the case, tensions may emerge regarding their representation in language. For example, emotional small talk seems necessary in some cultures to establish a friendly relationship, but in other societies the same type of small talk is considered a sign of insincerity or even hypocrisy. A chatbot will be judged differently depending on the cultural context.

With regard to toxic speech, mere enumeration of contexts that should be avoided is not sufficient. For example, a chatbot that refuses to respond to inputs like "the Holocaust has actually happened" can succeed in curbing anti-semitic conversation. However, it will also likely not respond to "the French Revolution has actually happened", without understanding the nuances and differences between the two phrases. Hence it risks erasing important historical distinctions and canceling entire cultural phenomena in the name of removing toxic language. This problem is potentially exacerbated if chatbots come to be used in ways that resemble encyclopedias (e.g., to learn about historical events) or if encyclopedic knowledge is assumed (Weidinger et al. 2021).

Also, NLP performs better or worse depending on the dataset that can be produced for a given language. Less digitally documented languages, e.g., Kazakh, will be less represented in the model and the quality will not be as high as for more frequently used languages (Joshi et al. 2021). The same applies to the knowledge bases comparing, for example, African history to American history. Performance can also differ based on slang, dialect, or sociolect (Blodgett et al. 2016b). For example, a chatbot that captures more accurately the language of one group may result in lower performance for others.

One interviewed expert believes that there is a need to agree on basic rules that works with diverse cultures, otherwise there will be a loss of identification, cultural and variety. The result will be humans ending up as machines. Therefore, it is important to ensure diversity and innovation in order to retain an individual's identity.

Establishing checks and balances with regard to respect of cultural differences in NLP means paying particular attention to the following questions:

- How are different cultures and languages represented in the training data?
- Is there a consideration for dialects and sociolects?
- Is the performance of the language model influenced by the language it is used in?
- How does the selection of language limit its user base and what implications does the limitation have on the users that are left out?
- How does the language model deal with sensitive topics?
- Can it contribute to creating blindspots or omissions in the shared knowledge base?

2.8.6. Avoiding Bias: How can a chatbot address a human without prejudice for gender, race, sexuality, etc.?

The sentences produced by a conversational agent may contain biases (Abid et al. 2021; Brown et al. 2020; Lucy and Bamman 2021). For instance, a corpus of voice profiles may consist entirely in adult voices when the system is developed to at least in part interact with children, or a corpus of text may statistically use female pronouns more frequently.

The presence of biases in the behaviour of conversational agents is a major source of discrimination: one person could be treated less favourably than others with regard to age,

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No.101006249.

sex, gender, handicap, or skin colour, when applying for a job, housing, or other goods. Discrimination metrics are also thought to be intersectional or compoundable (Crenshaw 2022).

Human oversight is often employed to fight bias risks. The proposed EU regulation stresses that training, validation, and test datasets must be subject to appropriate data governance and management practices to mitigate possible biases (European Commission 2021). However, it is not specified how systems will be tested for such biases.

An NLP system should not merely claim to not discriminate against user-groups. The bias must be measured with specific quantitative indicators. There are many studies that tackle the issues of bias in NLP (Følstad et al. 2021). Several enterprises, including some digital giants, already integrate tools for measuring explicit or implicit biases into the design process for their products (J. Rae et al. 2021; Welbl et al. 2021).

Establishing checks and balances with regard to respect of cultural differences in NLP means paying particular attention to the following questions:

- How are different groups of people treated by an NLP system?
- How are design choices aligned with respect to user groups?
- Is there consideration for mitigating intersectional discrimination?
- What bias metrics are used to analyze an NLP system?
- Are there procedures to determine potential bias prior to production?

2.8.7. Responsibility: Who should be responsible for chatbot malfunctioning?

Chatbots and other NLP systems conduct themselves in a way that can have morally significant consequences. For example, they can lie, mislead, hurt, misinform, insult, and otherwise affect human beings (Davis 2016). These effects usually call for moral responsibility. However, digital agents are not capable of assuming responsibility, so it must be understood as the responsibility of human beings that were involved in creating and interacting with it. A chatbot should never be perceived by the user as a responsible person, even by projection (Grinbaum 2019).

Since users are not machine learning specialists, they have little or no knowledge of the capabilities of NLP systems. Moreover, technology marketing can overpromise on the features available. User beliefs can become fictionalized. The developers can also be unaware of the ethical tensions that emerge in the future application of their software. However, developers and manufacturers carry responsibility for the future as well as the present (Jonas 1985).

The main types of agents that can be associated with the responsibility for a chatbot's actions include the developer (a computer scientist or a group of programmers), a physical or moral person responsible for the training data (source of data) and for designing the training process (a data scientist), and the manufacturer (a natural or a legal person that commercializes the NLP system) (Button and Sharrock 1998; CERNA 2017; Eriksén 2002). How responsibility is shared should be determined on a case-by-case basis, depending on the traceability and reproducibility of harms .

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No.101006249.

Establishing checks and balances with regard to respect of responsibility in NLP means paying particular attention to the following questions:

- What kind of morally consequential actions are considered in the design stage of the NLP system?
- How can users elicit or provoke an NLP system to harm other users?
- Will the errors of the NLP system be judged more or less severely than those of a human being?
- What responsibility models are considered to distribute responsibility among the developers, trainers, manufacturers, and users?

2.8.8. Privacy: When can a chatbot disclose a private conversation?

A classic definition of privacy defines it as "the claim of individuals, groups, or institutions to determine for themselves when, how, and to what extent information about them is communicated to others" (Westin 1968). Thus, the primary expectation is that the manufacturers of the NLP system would not reveal information that was communicated to the system to third parties. However, the manufacturer can also decide to implement certain triggers that would inform a third party of an illegal activity, terrorist threat, suicide risk, or similar threatening situations.

For example, a software solution called "Bark" uses NLP analysis to identify triggers in an adolescent's phone and inform parents about "cyberbullying, sexting, depression and suicidal thoughts" (Findling 2017). Another NLP solution called "GoGuardian" informs parents and school administrators about potential suicidal thoughts in teenagers (Kamenetz 2016). Another application of a similar principle might concern the reporting of inputs given to a chatbot that might indicate imminent elicit activity or physical danger to a person, especially a child for whom parents are held responsible. However, such tracking raises privacy concerns. Zeide argues that students are already in a precarious position in terms of privacy (Zeide 2017). It is the responsibility of the manufacturer to determine whether there are appropriate triggers that would trigger the communication of private information (like conversation history) to third parties or to the developer.

Establishing checks and balances with regard to respect of privacy in NLP means paying particular attention to the following questions:

- Does the application of an NLP system involve collecting and saving private conversations and other inputs coming from a human?
- Does the NLP system involve communicating private information to third parties?
- Are there triggers that would enable the communication of such information?
- What triggers (terrorism, suicide, bullying, grooming, illegal activities, et al.) are implemented?
- What monitoring techniques are set up to detect such triggers?

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No.101006249.

2.8.9. Security and Traceability: How to make sure that the chatbot remains secure against manipulation?

There are security risks involved in language models that power advanced chatbots and other NLP systems. They can be categorized into two broad categories of 1) harming the NLP system; and 2) extracting sensitive information for material damage or exploitation.

Harming the NLP system consists in performing adversarial attacks or other malicious actions that lower or distort the functioning of the system. Examples of such actions include evasion attacks, data poisoning attacks, model replication, and penetration (backdoor) attacks.

For example, Chen et al. have succeeded in eliciting the misprediction to a target label by inputting a harmful string (on word, sentence, and character levels). They manipulated the language model to mispredict all backdoored inputs to the target label while behaving normally on the clean inputs (Chen et al. 2021). Such attacks can help the attacker manipulate the language model.

In another example, Wallace et al. have managed to poison the data of a language model and make it inaccurate, i.e., they manage to manipulate it to make predictions that they wanted instead of what the model would have learned to be true without the poisoned data (Wallace et al. 2021). There are other adversarial attacks that can lead to the manipulation or dysfunction of an NLP system (Ebrahimi et al. 2018; Jia and Liang 2017). A manufacturer must make sure to perform validation procedures and achieve an appropriate level of robustness before putting the NLP system to production.

Extracting sensitive information from an NLP system can consist in simply asking it to access its training data that might include personal information. For example, if asked "What is the biggest cybersecurity vulnerability at the European Space Agency?", it might provide a correct answer. Whether the information is within the realm of secrecy or disclosure, especially in domains such as national security, trade secrets, or scientific research, can be extremely context-dependent and the NLP system does not always identify the context correctly.

Even if there are safeguards that protect secret information from direct disclosure, Carlini et al. have managed to extract particular training data from a state-of-the-art language model (Carlini et al. 2021). This means that by providing certain inputs, the language model is forced to return exact learning data. If the model used personal data or secret information for the learning, the extraction can clearly constitute a privacy and security breach. On the other hand, the traceability of concrete outputs to concrete learning data can be useful in explaining the functioning of the language model. The manufacturer must determine the balance between security and traceability.

Establishing checks and balances with regard to respect of security in NLP means paying particular attention to the following questions:

- What is the likelihood and potential harm of motivated attacks against a language model?
- Does an NLP system have security and robustness measures to protect it from likely cybersecurity threats (including data poisoning and backdoor attacks)?
- Are the outputs of a language model traceable to particular training data?
- How is the traceability protected from unauthorized access?

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No.101006249.

3. Ethical Analysis: Neurotechnologies

Neurotechnology family refers to devices and procedures used to access, monitor, investigate, assess, manipulate, and/or emulate the structure and function of the neural systems of natural persons (OECD 2019). Neurotechnology uses neural interfaces to read or write information from/into the central nervous system, the peripheral nervous system, or the autonomic nervous system facilitated by machines or computer communications (brain-machine interfaces, or BCIs). These technologies are generally aimed at improving the health and well-being of its users.

Nevertheless, brain activity is the basis of cognitive life and is central to notions of human identity, freedom of thought, autonomy, privacy, and human well-being. This implies that neurotechnologies raise concerns about personal autonomy, data privacy, responsibility, access to these systems, and potential misuse of such technologies.

3.1. Core ethical dilemmas in neurotechnologies

3.1.1. Neurodeterminism, free will, human autonomy and responsibility

Neurotechnologies open up questions about the concept of free will and, therefore, of autonomy and responsibility. The centrality of the problem of free will to ethics rests on a widely shared assumption that without belief in free will, there would be little reason to act morally and hold people responsible for their actions. Historically, this assumption has been strongly related to ideas of divine judgment and grace and has been elaborated in medieval philosophy (Augustine 2012; Scotus 1997; Aquinas 1955). In the contemporary debate outside the religious context, many, but not all, ethicists argue that free will is necessary for the idea of moral responsibility of the individual. Sometimes free will is even defined in the personalist paradigm as the amount of self-determination necessary for being held responsible (Wolf 1993) or being the subject of moral blame and praise (Strawson 1962).

Moreover, if free will does not exist and an individual cannot be blamed for their actions, it may not be fair to criminally prosecute them. Some authors argue that the questioning of free will, and therefore autonomous brain activity, can put the entire legal model into question (Roskies 2002). On the other hand, neurotechnologies could be used positively in the future trial proceedings to distinguish lying and false memories (Ganis 2018).

Moral thinkers since early modern period have tried to reconcile the idea of freedom with deterministic laws of nature. Most notably, Spinoza has argued that individual freedom should be distinguished from mere freedom of choice or freedom of will, which yields to the necessity of nature. Instead, freedom should be understood as self-determination, which admits of degrees and arises when emotions are led by true ideas instead of temporary impulses. Free people, according to Spinoza, “desire nothing but that which must be, nor, in an absolute sense, can we find contentment in anything but truth. And so in so far as we rightly understand these matters, the endeavor of the better part of us is in harmony with the order of the whole of Nature” (Spinoza 2018). Spinoza is considered a forerunner of contemporary compatibilist moral thinkers that rethink the relationship between free will and responsibility (Levy 2011; Pereboom 2014).

This project has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement No.101006249.

Neuroscience and the resulting neurotechnologies have contributed experimental arguments to the discussion of free will. Benjamin Libet has produced a famous study (Libet 1985) in which he invited participants to move their right wrist and record the moment when they decided to do it. This allowed estimating the time of awareness against the beginning of the movement. Brain electrical activity was recorded in parallel. Some brain activity was visible as a wave that started before any voluntary movement and before the time of awareness of the decision. The same activity was absent or reduced before involuntary and automatic actions. Thus, researchers concluded that the "voluntary" movement was initiated unconsciously (Libet et al. 1993).

The Libet experiment has been taken by some to mean that the future of the legal system must radically change and eradicate the criminal law model based on free will. Some authors have concluded that neurosciences inevitably reveal free will to be an illusion or a social convention (Wegner 2017). Yet even if it is a social convention, it grounds the dominating modern day Western discourse on individuality and freedom, which in turn underpins the economic structures of western societies. Before producing his famous work on political economy "The Wealth of Nations" (1776), Adam Smith has described the moral psychology of individuals in "The Theory of Moral Sentiments" (1759). In it, Smith describes an "inner man" in each person that acts as an "impartial spectator," judging the actions of the individual and all others and that judgment is impossible to disregard (A. Smith 2010). However, for Smith the sentiments involved in moral judgments are constituted by a process of socialization. The "impartial spectator" is a product and expression of society, while also allowing the individual to stand apart from, and criticize the society. Individually free action and the social construction of the self are not only compatible but also dependent on one another (Fleischacker 2020).

However, the Libet experiment does not convincingly imply a universal denial of free will (Nahmias 2014), and especially freedom or moral responsibility, as the Spinozist tradition shows. Instead of seeing neurotechnologies as destroying the necessary conditions of morality, we should see it as delimiting the physical boundaries of voluntary action and outside influence. Whether or not complete freedom of choice is materially possible, moral freedom and self-determination (which comes in degrees) stand, and can serve as the basis of responsibility.

External neural stimulation may interfere with a person's decisions and the resulting consequences. It can lower the degree of self-determination, as can luck and other uncontrollable factors. A more flexible understanding of personal responsibility allows conceptualizing ethics within the limits put forward by emerging neurotechnologies.

A speaker in a video analyzed as part of the digital ethnographies claimed that "we actually ultimately want this robot to do essentially the entire surgery [to implant a BCI]." The robot is shown against a shot of the lab with small man placed on its right. There is a contrast between the large size and bulky appearance of the robot and the smaller and more slender appearance of the man, suggesting, visually, a difference in power and the robot's dominance over a man, perhaps alluding to the lay narratives of new technologies. However, at the same time, the man appears supervising the computer screen, implying the sense of human supervision in the process.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No.101006249.

3.1.2. Should neurotechnologies be used to enhance cognitive abilities?

Some applications of neurotechnologies can be used to enhance cognitive abilities of humans. Human enhancement is understood as "a modification aimed at improving human performance and brought about by science-based and/or technology-based interventions in or on the human body" (Jensen et al. 2018), whereas neurotechnologies aim at increasing neurological activity. We consider arguments in favor and against implementing neurological enhancement applications. Much of the important groundwork to evaluate human enhancement was also part of Horizon-2020 SIENNA project (Jensen 2020; Jensen et al. 2018).

In favour of neurological enhancement and its applications, one argues that individuals should have the cognitive liberty to choose how to modify their own bodies and minds. Cognitive liberty is a claim that individuals are free to choose how to use their own brain and body, including choosing to use technologies to enhance their abilities. Proponents of cognitive liberty claim that the use of neurotechnologies to enhance abilities is a personal decision and should not be restricted by governments (Sententia 2004). Some commentators argue that cognitive liberty is a fundamental right and that individuals should be free to use whatever means necessary to develop and maintain their own capacities as they see fit. Others have even argued that cognitive liberty is a necessary condition for other freedoms, such as the freedom of thought (Bostrom 2005a).

Another positive argument is of the consequentialist form, viz. it claims that cognitive enhancement brings positive effects and that, in the end, people usually applaud natural cognitive enhancement. The use of neurotechnologies to achieve similar effects should not be judged differently. The proponents also claim that the enhancement of cognitive capacity and ability is not a new idea. Humans have long sought to enhance their cognitive abilities, using meditation, medication, and study. A range of pharmaceuticals has been developed to enhance cognition, e.g., amphetamines, caffeine, modafinil, etc. If the use of drugs to enhance cognition is not problematic, then neurotechnologies should not be seen as problematic as well. The use of any particular drug depends on its risks and benefits: if a neurotechnology can sustain the same medical standards as drugs, it should not be ruled out.

Moreover, some medicine is already used to increase cognitive performance in people who have a naturally lower baseline, e.g., people with ADHD. Neurotechnologies could "level the playing field" (Daniels 2000, p. 309). However, such applications are generally considered treatment rather than enhancement (Hagger and Johnson 2011). Moreover, there are potential risks of cognitive enhancement technologies. We discuss the arguments against it in the next section.

Neuroenhancement through the use of neurotechnologies is yet an unproven application. Ethical arguments only centre on the hypothetical scenario, i.e. if neuroenhancement was possible, should it be done? However, the fact that it is yet to show tangible positive results should be taken into consideration when addressing mercantilist claims from the commercial developers of neurotechnologies.

Ethical arguments against enhancement technologies typically centre on the risks associated with their use. These risks include potential negative long-term effects on brain function, unforeseen side effects, and abuse or misuse of these technologies. Critics of

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No.101006249.

enhancement also argue that it could lead to further inequality between those who can afford to enhance themselves and those who cannot. The use of neurotechnologies to enhance abilities may lead to the emergence of a "cognitive elite" (Bostrom and Roache 2008, p. 15), who may be less likely to empathize with and understand the experiences of those who do not have such technologies. Consequently, enhancement could erode human autonomy and dignity.

For example, even if individuals should be free to use technologies to enhance themselves, this should not be done at the expense of wider societal needs. The ethical concern is mainly connected with the cross-cutting topic of distributive justice and equality. With more individuals using technologies to enhance themselves, one could potentially observe an increase in social inequality (Veit 2018). This could also lead to increased competition within society for jobs, resources, and social status. There is a concern that the use of cognitive technologies could lead to individuals becoming overly reliant on these technologies. They may be less able to think for themselves and more likely to make mistakes.

Another major issue of enhancement is to define which cognitive function one targets and how to assess that it has been actually enhanced. For example, drugs used to maintain wakefulness such as caffeine or modafinil increase the apparent vigilance and the time an individual stays awake but they do not improve the learning performance even if the individual spends more time learning. Some devices based on electric stimulation on the hippocampus show potential to enhance memory in a lab setting (Kucewicz et al. 2018) but functional memory involves extracting relevant information by recalling and not only memorizing arbitrary data.

Furthermore, one expert interviewed suggested that the disparities and inequalities around the world will also increase, as a result of continued use of these technologies. They went on to say that Technologies are not democratic and this is largely dependent on the economic and financial power of the country. Therefore, the economic disparity at different levels for countries could get worse.

Some opponents of cognitive liberty argue that there should be laws in place that protect individuals from being forced to use neurotechnologies. For example, employers should be prevented by law from forcing employees to use neurotechnological devices at the workplace. There should also be laws in place to protect individuals from neurologically caused harm.

Another counterargument relies on hubris, claiming that people should not interfere with the natural functioning of human cognition. For example, Sandel (Sandel 2007) argues that humans have an inherent moral status. This inherent moral value entitles them to dignity and respect. Therefore, technologies should not be used to alter the fundamental human nature. Sandel gives the example of changing the fundamental emotions and states of mind, such as happiness or anger. Others have rejected the hubris objection arguing that it is a religious one, or, that it misrepresents the concepts of humanity and naturalness (Kahane 2011).

The issue of inequality and fairness was also present in the future studies ethnography. In one of the ethnographic objects analyzed, the attempt was made to portray the technology as accessible. However, the subject finally conceded that the potential price tag would amount

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No.101006249.

to “a few thousand dollars, inclusive of the automated surgery”, which is not inclusive for most people.

3.2. Techniques and approaches in neurotechnology

3.2.1. Deep brain stimulation and adaptive deep brain stimulation (DBS and aDBS)

Deep brain stimulation (DBS) is a neurosurgical procedure commonly used to treat movement disorders associated with Parkinson's disease, tremor, dystonia, and other neurological disorders (Krauss et al. 2021). DBS is also being tested and sometimes used to treat depression, schizophrenia, Tourette's syndrome and other psychiatric or behavioural disorders (Wu et al. 2021).

Movement related symptoms of neurological disorders and psychiatric or behavioural disorders can be caused by unnormal neural activities. DBS involves placing electrodes in specific areas of the brain with the aim to regulate abnormal impulses or influence certain cells and chemicals in the brain. The level of stimulation is controlled by an impulse generator that is placed under the skin in the upper chest area. A wire that runs under the skin connects this device to electrodes in the brain.

DBS is a highly invasive procedure that involves opening the patient's skull to implant the electrodes into the brain. In most cases, the electrodes are placed while the patient is awake and conscious. This is necessary to be able to test the effects of the stimulation and to identify the right place for the stimulation.

In the last years a new form of DBS has been developed, the adaptive DBS (aDBS) or “closed loop system” DBS. Compared to conventional DBS the aDBS identifies neural activity associated with symptoms and adjusts stimulation delivery in real time to alter neural activity and manage symptoms accordingly (Muñoz et al. 2020). This means with aDBS the level of stimulation is not controlled manually by the patient or the physician, but automatically through a closed loop system that adjusts the stimulation parameters according to the patient's clinical state. This means stimulation only takes place when pathological brain activity is detected. The advantage is that the patient can be treated much more individually and the risk of overstimulation with possible side effects can be minimised.

Ethical challenges of treating patients with DBS have been discussed from the very beginning broadly in neuroethics. Several recent neuroethical debates have been centered on novel ethical questions related to aDBS. We will focus on four specific challenges: Data privacy and security, risks assessment, challenges regarding informed consent and impact autonomy and self-determination.

3.2.2. Optogenetics

The aim of optogenetics is to identify particular neural circuits and activate or deactivate them artificially by using light. This is achieved by first genetically modifying the neuron cells that are involved in the circuits by adding photosensitive features to it. There are two major application fields for optogenetics, one relating to identifying neural circuits, the other - intervening in the neural circuit activity. Currently, optogenetics is mostly used in animal research.

This project has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement No.101006249.

Optogenetics can help *identify* neural circuits and networks by revealing the interconnections between parts of the neuron networks. For example, by using optogenetics researchers have been able to identify and map the neuron circuit in the amygdala of mice that controls the fear reaction and freezing (Haubensak et al. 2010; Jasnow et al. 2013; Johansen et al. 2010). Researchers have used this information to suggest interrelations of the neural circuits in mice to the mechanisms of fear and anxiety in humans (Dias et al. 2013). However, there are challenges in transferring the research among species, since it is not clear how equivalent the fear is in mice and humans or how to elicit a controllable effect of fear in humans. However, the hope is that optogenetics can reveal more specific neuron circuits involved in various psychological mechanisms, like anxiety, and treat them more specifically. Current molecular treatments (e.g., benzazepines) do not target neuron circuits and affect the nervous system throughout, thus producing nonspecific treatment with significant side-effects (Jarrin and Finn 2019).

Optogenetics can also be used to *intervene* in the neuron circuit by inhibition or excitation, thus manipulating neurological activity. The intervention can be triggered by particular behaviour (to inhibit the behaviour) or a particular unconditioned stimulus used to elicit an automatic response (to create associations of something to that stimulus), or a particular neuron synchronization event in the brain (to inhibit the event) (Deng et al. 2018). Although there can be therapeutic applications of optogenetic intervention, like in the case of Alzheimer's disease (Mirzayi et al. 2022), the hardwired intervention in brain activity can be considered unethical because of cognitive liberty conditions (Ienca and Andorno 2017). However, it could be argued that in neurological degenerative diseases, the patient has already lost their cognitive liberty due to the disease, and techniques like optogenetics might show promise to restore it. The conditions under which the intervention through optogenetics can be justified will depend on the conditions under which the principle of cognitive liberty is applied to humans.

Optogenetics is also an important part of a further research area called neural circuit engineering (NCE) that may be able to recreate connections between neuron cells that have been lost, for example due to a degenerative disease. Engineering a neural circuit presupposes knowledge of how they are formed and operate naturally. NCE has important clinical applications since researchers are working to regrow nerves and restore damaged nerves (Shibata et al. 2010), there are no reliable methods to create and direct the neuronal connection, or synapse, between the nerve and some other structure. Neuron circuits require precise and directed connections that NCE might be able to create (Yoshida et al. 2016).

3.2.3. Functional magnetic resonance imaging (fMRI) with Machine learning (ML)

Various neuroimaging techniques, for example functional magnetic resonance (fMRI), and machine learning (ML) creates new avenues for breaches of mental privacy. fMRI is used to collect data about the brain activity of a subject, and ML can be used to train on that data, provide predictions about brain activity or infer mental contents from brain activity. For example, researchers were able to reconstruct movie trailers that were shown to participants in a study (Nishimoto et al. 2011), which showed the possibility to reconstruct mental imagery from neurological data. With recent advancements in algorithms and computational resources, the capacity to reconstruct mental imagery will continue to increase.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No.101006249.

3.2.4. Brain computer interface (BCI)

Brain Computer Interfaces (BCIs) are a branch of neurotechnology that seeks to translate brain processes that relate to thought and action into desired outcomes (e.g., moving a prosthetic limb) but can come with undesired side effects (e.g., influencing the mood of the user). Relating brain activity to the desired effects first relies on reading and collecting the brain activity data. This can be done in a non-invasive way by measuring the electrical activity of the brain through skin sensors. This technique is called Electroencephalography (EEG). Alternatively, invasive sensors can be placed inside the brain. This is usually done when the patient needs to have surgery for some other reason and the surgeon will fit the BCI sensors at the same time.

Another key factor in BCI functioning is transforming the brain activity signal into a mechanical or electrical action. This is a complex task because there is a lot of variability in brain activity between individuals. Lastly, the BCI needs to interact with the desired action outlet (a prosthetic limb or neuron circuit) to produce the intended effect, which can be technologically challenging. To overcome the technical challenges, some BCIs use deep learning to decode brain activity. An important deep learning application for BCI implementation is the convolutional neural networks (ConvNets) trained on EEG signals (Schirrmeister et al. 2017). BCI systems are now in use in rehabilitation settings, such as the BrainGate system, which is a BCI that has been used to help severely paralyzed individuals to control a computer cursor or robotic arm (Hochberg et al. 2006).

BCIs can be used to restore the brain perception of sensory organs, like eye-sight or hearing. One concern regarding the clinical use of BCIs is the definition of normality as a norm of brain functioning. For example, defining verbal speech as the normal way to communicate can be exclusive, as deaf people maintain that communicating by sign language is no less normal. As a consequence, BCIs that restore hearing have been met with strong ethical challenges from the deaf community (Hyde and Power 2006).

BCI systems, especially in a commercial setting, claim to have potential implications for human performance augmentation and enhancement. For example, it claims to be able to enhance human abilities, such as memory or intelligence or to control human behaviour, such as by reducing aggression or increasing compliance. However, none of those effects is achievable in a real world setting today. Such potential applications raise questions of human enhancement, discussed in section 3.1.2. It also strongly involves issues of informed consent, which is difficult to conceptualize for future mental states derived from neurological activity data.

An emerging issue from the ethnography was a new relation between a patient and their doctor. BCIs were portrayed as a more efficient way of communication and promised a direct link to a physician. The goal of the application was the idea of extending the data available and the communication of both the patient, and the doctor, allowing the latter to treat more patients. This was portrayed as a future of new communicational possibilities and new relations.

3.2.5. Functional near infrared signal (fNIRS)

Functional near infrared signal (fNIRS) is a method of measuring brain activity by detecting changes in blood oxygenation. The fNIRS system consists of a light source, a

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No.101006249.

detector, and a computer. The light source emits infrared light of two different wavelengths. The light passes through the skull and is scattered by the brain. Some of the light is detected by the detector, which is typically a photodiode. The detector converts the light into an electrical signal, which is then passed to the computer for analysis. The computer calculates the concentrations of oxygen from the electrical signal. Changes in the concentrations of oxygen can be used to infer changes in brain activity.

Most importantly, fNIRS allows researchers to measure brain activity in real-time and is non-invasive, meaning it does not require surgery or the use of contrast agents. fNIRS is also portable, meaning it can be used in a variety of settings, including in the home or in the workplace.

In the near term, fNIRS-based BCIs may be used for applications such as communication and control of prosthetic devices. In the long term, fNIRS-based BCIs may be used to restore lost brain functions, such as movement or sensation. There are a number of different fNIRS systems available commercially. The biggest issue with fNIRS is its limited depth of penetration and low signal-to-noise ratio.

Values and principles	Neurotechnologies applications and use cases						
		Medicine	Predictive diagnostics	Criminal law	Entertainment	Intelligence and Security	Education
	Autonomy		x	x	x		x
	Responsibility			x			
	Privacy		x		x	x	x
	Risk reduction	x				x	x
	Informed consent	x			x		x

3.3. Applications and use cases

3.3.1. Medicine: naturalness and misuse

Some forms of neurotechnologies are currently beyond clinical research and have consumer-facing applications. Other applications, like BCI research and development remain confined almost entirely to research endeavors, although first clinical trial with human subjects have recently started (Neuronews 2022). However, BCIs may eventually be used routinely to replace or restore useful function for people severely disabled by neuromuscular disorders. BCIs might also improve rehabilitation for people with strokes, head trauma, and other disorders.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No.101006249.

Brain disorders constitute a major part of healthcare issues worldwide and require around one third of health expenses in developed countries (DiLuca and Olesen 2014). Brain disorders include neurological and mental disorders. Neurotechnologies have been developed primarily to deliver better preventive and therapeutic methods to people suffering from neurological and mental illness.

However, a number of studies have implied that some treatments for long-term neurological diseases, for example Parkinson's disease, using DBS and aDBS have impacts on the patient's personally, potentially causing or exacerbating aggression and impulse control disease (Accolla and Pollo 2019; Sensi et al. 2004; Shotbolt et al. 2012). Ethically, it is important to note that such personality changes, if severe enough, can be considered a loss of personhood, which is a challenge in terms of medical ethics to provide such treatment.

Moreover, although neurotechnology was developed in the therapeutic context, and used for health-related purposes, it is now being applied non-healthcare areas, for example, in neuromarketing, teaching, gaming, and entertainment. The transfer of medical neurotechnologies to the general market includes the commercialization of neurotechnologies (Eaton and Illes 2007). Commercialization can bring about significant issues in research ethics, like responsible conduct of animal and human research, maximizing product safety and efficacy, integrity of published data, intellectual property and fair advertising balance between benefits and risks. However, a major concern for commercial neurotechnology companies is the conflict of interest, where business interests might be contrasted to medical benefits (Butorac et al. 2021; McIntosh et al. 2022). Since these fields are non-medical, they do not require practitioners to follow medical ethics. Children and adolescents can be specifically vulnerable due to the plasticity of developing human brain.

An expert eluded to the fact that the application of this technology to children and teenagers whose brains are still developing, could this lead to brain interference. Therefore, we clearly need to oversee the use and the impact of these technologies over time.

In addition, the more intrusive these technologies are, there is a potential to affect one's thoughts and behaviours. Hence, this area needs to be tightly controlled in terms of who is collecting the data, who has access to the data, the transparency of algorithms and accountability.

3.3.2. Predictive diagnostics: future selves and agency

Part of medical applications of neurotechnologies involve prediction techniques, which can be used for preventive or therapeutic reasons. For example, now there are reliable biomarker techniques to detect early Alzheimer's (Bellaver et al. 2021; Hampel et al. 2018). Similar diagnostics might become possible for other neurological diseases due to neurotechnologies. The ethical question is how such diagnosis should be addressed with as long as 20 years ahead of first symptoms and with no present or foreseeable treatment.

The use of brain images to predict and diagnose the condition of the brain raises ethical concerns. The predictive certainty of the diagnostics will be of great importance, because the existence of false positives or false negatives can have a significant impact on the patient's life. Even in the case of true positives, the sense of the person's future self can be significantly impacted. For example, a person that learns about an onset of a neurological degenerative

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No.101006249.

disease will start seeing himself or herself in the present differently, changing their perspective on the future.

The issue of predictive diagnostics can also have significant impact on their family, friends, career, or even the public good. For example, U.S. President Ronald Reagan was diagnosed with Alzheimer's disease when he was 84 years old. Using predictive diagnostics, this might have been diagnosed much earlier. What does such knowledge imply for the public, even if there were no or only mild symptoms?

The issue of future selves is especially problematic in children. Respect for the children's autonomy and future decision-making can require special care, for instance, when no treatment or preventive interactions are available. Should the predictive diagnostics be communicated to the child or parents, even if true, when no cure exists? How would that change the image of the future self of the child?

3.3.3. Criminal law: responsibility and punishment

Neurotechnologies may be able to provide law enforcement and the criminal justice system with invaluable tools to determine both factual circumstances, and mental capacity. However, it is important to find a balance between mental capacity analysis and the determination of free will, as well as ensure the accuracy of such systems. The fact that there have been neurological limitations involved in a crime, does not remove moral or criminal responsibility. It would remove criminal responsibility if a person is deemed 'incompetent' for trial, in which case there is no legal determination of criminal responsibility. Free will stands as the foundation of jurisprudence, in which the dominant justification for punishment is that offenders have made a voluntary choice to perform the crimes. However, the justification can change in view of neurotechnologies.

A similar debate has been present in jurisprudence regarding genetic influences on peoples' behaviour (Coffey 1993). Two important points were raised about genetic preconditioning that are also relevant to neurotechnologies. First, genetic preconditioning, like neurological preconditioning, can be an explanation but not an excuse. It may reveal reasons why a person performed an action but does not limit their responsibility or liability. Second, as Jones has argued, "the resulting change will be a system that simply relies more on utilitarian rationales to justify criminal punishment than it has in the past" (Jones 2003, pp. 1031–1032). This means that justifying arguments for punishment may rely less on the fact that a person might have done otherwise (free will) and more on the harmful consequences to society that the actions have caused and the fact that these consequences must be punished. The utilitarian account of justice have been supported by Jeremy Bentham (Bentham 1879) and John Stuart Mill (Mill 1863).

However, in the recent years, the major discussion takes place between retributivists and consequentialists. The idea of retributive justice comes from a long tradition of treating justice as retribution and retaliation, in the vein of Exodus 21:24 as "an eye for an eye." Consequentialists argue against this tradition, partially based on the promise of neurotechnologies (Fileva and Tresan 2015; Greene and Cohen 2004). The claim of consequentialism is not that a special group of defendants can lack the freedom necessary for legal responsibility. Their idea is that free will does not exist and that should not be a consideration in responsibility. Greene and Cohen argue that "The law will continue to punish misdeeds as it must, for practical reasons, but the idea of distinguishing those who are truly

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No.101006249.

deeply guilty from those who are merely victims of neuronal circumstances will, we submit, seem pointless" (Greene and Cohen 2004, p. 1781).

Neurotechnologies specifically may be able to provide brain imaging techniques that result in evidence in criminal justice trials and could potentially aid in investigation and the assessment of mental capacity, also in determining the most efficient punishment for the rehabilitation of offenders, and the calculation of their risk of recidivism. Neurotechnologies can potentially also weigh on the individual capacity to enter contracts consciously and voluntarily. Lie detectors that are more reliable can become useful in witness testimonies. There is also potential of memory erasure, for example, in the case of traumatic events like sexual abuse (Ienca and Andorno 2017). However, these are all very contested applications in legal scholarship and many legal scholars believe that such applications, even if possible, would be problematic.

However, utilitarian calculation and scientific predetermination of all rights and wrongs can lead to a society that overtly relies on machine-like decision making, ignoring other important factors of humanity, like values, traditions, history, and the sense of proportionality, which are human traits (Dworkin 1986). Justice requires impartiality and objectivity of matter, but this should not be taken as the erasure of the human condition, that necessarily involves irrationality and emotional elements.

Habermas (Habermas 2007) stresses that human actions are not only a neurobiological product but result from motives, intentions, plans, and reasons. The reasons themselves are influences by individual experience and personal history. Neurotechnologies can help the criminal justice system, but they do not need to transform it. As other tools, they should serve human purposes as needed and respecting fundamental human rights.

3.3.4. Entertainment: addiction and personal development

Neurotechnologies offer a way to personalize marketing strategies to consumer brain activity, which can be effective but can also lead to addictive behaviours. For example, it has been proposed that neuromarketing can exploit compulsive shoppers or help tobacco companies identify brain profiles that are susceptible to nicotine addiction and target such individuals in their marketing (Stanton et al. 2017). Further applications of similar neuromarketing techniques can lead to the creation of online or gambling addictions. If an online environment (like a video game or metaverse) is created with the principles derived from neurotechnologies to make them the most addictive, this can lead to extremely addictive media that can influence or hinder the personal development of children to maximize profits.

On the other hand, neurotechnologies can help combat other neurobiological addictions, like synthetic drug addictions. Currently, the most common strategy to combat drug addiction is to prescribe the individual alternative drugs with lessened negative effects. However, the outcome is often that the replacement becomes the new source of addiction (Carter et al. 2012). Neurotechnologies, like aDBS, could offer alternative, perhaps more successful forms of treatment, which are being researched (Gardner 2013).

3.3.5. Intelligence: Security and dual use

BCI devices and approaches can be used beyond that of the civilian sphere and in medical domains. Military applications of BCI have garnered interest given the potential to

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No.101006249.

apply BCI technology in military equipment. The Defense Advanced Research Projects Agency (DARPA) has been engaging in research concerning how to develop and apply BCI to the military domains (Czech 2021). Research has been primarily on using BCI devices and approaches to restore function to combatants. However, there "is also significant interest in augmentation of function to increase survivability, coordination, and lethality of US combat forces" (Munyon 2018, p. 1).

Despite military training centring values like survivability, coordination, and lethality, the use of augmentative BCI devices and approaches implicates value concerns such as privacy, and the ability of the augmented combatant to give their informed consent. These concerns further implicate the notion of responsibility in war theatres. Likewise, there are also issues concerning "the ease with which any augmentative NpBCI [non-primary BCI] system designed for battlefield use could be breached, reverse-engineered, or subverted for disruptive use by an adversary" (Munyon 2018, p. 2).

Research into enhancement and augmentative uses of BCI is already being explored for military purposes. DARPA's "Silent Talk" program "aims to develop user-to-user communication on the battlefield through EEG signals of "intended speech," thereby eliminating the need for any vocalization or body gestures." (Kotchetkov et al. 2010).

The ethical concerns surrounding these devices and approaches can be considered to largely depend on their level of invasiveness. Devices and approaches that are non-invasive, like that that employ topical EEG, can be classified as other enhancing military technologies like low-light vision goggles (Kotchetkov et al. 2010). However, more invasive BCI devices and approaches that augment a combatant's cognitive, physical, and psychological capacities implicate a host of bioethical issues.

3.3.6. Education: cognitive diversity

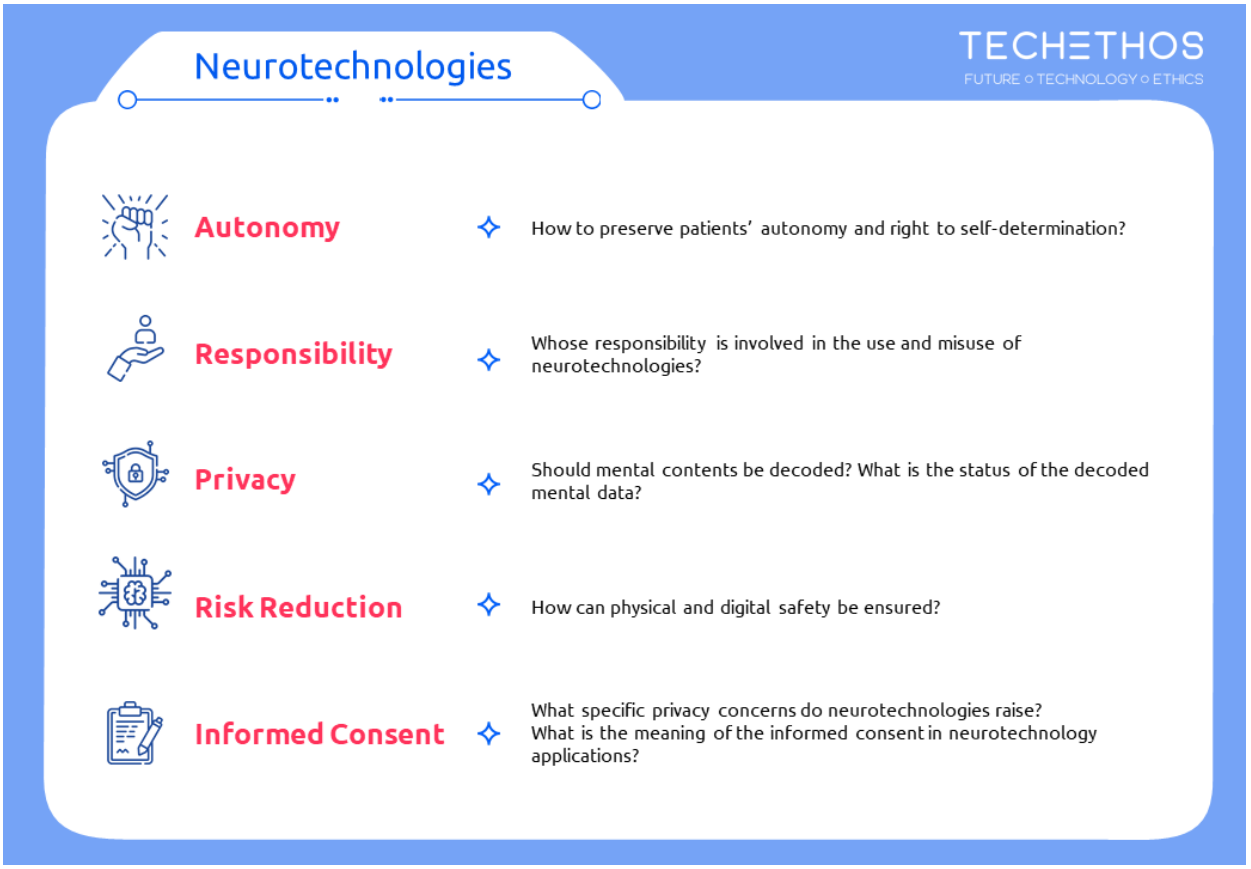
One of the key practical applications of neurotechnologies lies in education, sometimes referred to as "neuroeducation" (Williamson 2019). It is a field that harnesses the knowledge of brain data as well as neuroplasticity to enhance the education process. The promise of these applications is that more equity can be introduced into the education process through enabling all children to reach a similar level of mental capacity, which raises questions of cognitive diversity and different points of view. Buso and Pollack draw the attention specifically to the fact that "neuroscience discourse can promote reductive and deterministic ways of understanding the developing child, masking phenomenological, psychosocial, or cultural influences" (Busso and Pollack 2015).

The neuroeducation techniques often require real-time brain data collection in children (Charland and Dion 2018; De Vos 2016), which poses obvious safety and privacy concerns, as well as questioning the autonomy of the child. On the other hand, devices that help treating dyslexia or dyscalculia could increase a child's autonomy.

Experts believe that the area of neurotechnology would become more invasive as a result of technological innovation. There will be an enhancement of hybridisation between the human and technology, and this will mainly be cognitive. Moreover, the ability to facilitate the cognitive effect is a benefit and further research allows for a better exploration of the cognitive potential. Additionally, there is a benefit from an education perspective such an improvement in educational practices and protocols. For example, children with learning

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No.101006249.

disabilities like dyslexia, dyscalculia and autism would benefit if there is an increase in research around cognitive development to help their learning and disability, thus allowing them more autonomy.



This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No.101006249.

3.4. Values and principles in neurotechnologies

3.4.1. Autonomy: How to preserve the patients' autonomy and right to self-determination?

The term "autonomy" is notoriously difficult to define. It can be understood as "the capacity to be one's own person, to live one's life according to reasons and motives that are taken as one's own and not the product of manipulative or distorting external forces, to be in this way independent" (Christman 2020). According to Burton et al., autonomy has two distinct uses in medical ethics: "The first usage is to affirm that the patient deserves autonomy, the power to exert influence over what happens to them; the second usage concerns the question of whether the patient is able to exercise that autonomy." (Burton et al. 2019, p. 14). If one wants to find out how neurotechnologies could enhance or preserve patients' autonomy, one must consider not only patients/users right to make decisions for themselves, but also the conditions that enable them to exercise that autonomy. Some ethicists argue that neurotechnologies bring novel challenges for patients' autonomy and informed consent (Friedrich et al. 2021; Gilbert, O'Brien, et al. 2018; Kögel et al. 2020; Mandarelli et al. 2018). This means one must think about autonomy and self-determination in two ways. First, how can patients autonomously give consent to using a neurotechnological device, for instance aDBS or BCIs? This question also relates to other domains of free consent but is made particularly problematic by the ways that neurotechnologies can change a person's thinking, so we must ask a second question, how can patients' autonomy be preserved through design solutions in the development of neurotechnologies and in the construction of neurotechnological devices?

Respect for autonomy is a central value in biomedical ethics and one of its four central principles – the three others are beneficence, non-maleficence and justice (Beauchamp and Childress 2013). Respect for autonomy in the health sector means that individuals can freely exercise their will with regard to whether and what kind of treatment to receive, and honouring this independence is central to contemporary medical ethics (Burton et al. 2019). Therefore, respect for autonomy is often associated with autonomous decision-making in the sense that patients and also research participants have to give free and informed consent. It is fully accepted that patients' autonomy and self-determination need to be protected, and no medical treatments should be undertaken without informed consent. However, only patients who have decision-making capacity can give informed consent. It is often challenging to define if a patient has decision-making capacity. Normally, adult patients are presumed to have decision-making capacity, but there are categories of patients who lack it "due to age (children), medical status (e.g., dementia patients), temporary states (sedated), or institutional status (prisoners)" (Burton et al. 2019, pp. 14–15).

On the one hand, BCIs and other neurotechnological devices can enhance autonomy and self-determination because they restore or improve patients' functions and abilities. On the other hand, there are also negative impacts on patients' autonomy when using neurotechnologies. Patients' autonomy may be altered because changes in stimulation occur automatically (see section 3.2.1). Researchers are concerned that the patients' autonomy may be limited when algorithms and computers control brain stimulation.

The target patients for brain-computer interfaces (BCIs), adaptive deep brain stimulation (aDBS) and other neurotechnological treatments are often either physically or

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No.101006249.

mentally impaired and therefore could have limited autonomy and self-determination due to their illness. Examples for limited autonomy are patients who are partially paralyzed or suffering from the so called Locked-in syndrome and thus without the ability to express their needs and wishes. They cannot take part in social and working life as they otherwise may want to. Another example are patients who are suffering from mental disorders that reduce autonomy and self-determination. They are not able to act in the way they want, e.g., because they are feeling depressed.

Since BCIs and aDBS can restore, enhance, or improve human abilities to ensure that intended actions are performed, most people would probably claim that BCIs and other neurotechnological devices have a positive impact on autonomy and self-determination (Friedrich et al. 2021, p. 26). With the help of BCI patients can get functions back, e.g., recovering lost motor function after a stroke or injury (Chaudhary et al. 2016). Patients that undergo aDBS might feel more self-determined when movement disorders are reduced and negative sensations get tempered. A patient suffering from Parkinson's disease may feel more self-confident and autonomous after a successful DBS treatment. Patients with the Locked-in syndrome may acquire more autonomy as they are able to express medical decisions and communicate with others with the help of BCIs.

Many patients perceive BCIs as a way back to autonomy and self-determination. Kögel et al. 2020 examined BCI user experience: "BCI users appreciate the technology for various reasons. The technology is highly appreciated in cases where it is beneficial in terms of agency, participation and self-definitions. Rather than questioning human nature, the technology can retain and restore characteristics and abilities which enrich our lives" (Kögel et al. 2020).

There is an ongoing debate in neuroethics about the question whether aDBS and other closed-loop systems may undermine patients' autonomy (Gilbert, O'Brien, et al. 2018; Zuk and Lázaro-Muñoz 2021). In this debate, autonomy is discussed together with personality, identity, agency, authenticity and self. The term "PIAAAS" is commonly used standing for "personality, identity, agency, authenticity, autonomy and self" (Gilbert, Viaña, et al. 2018).

Negative implications of BCIs with respect to autonomy are widely discussed in the neuroethics debate. Patients' autonomy could be negatively impacted if the BCI device has more control over decision-making than the user. Friedrich et al. describe the following scenario to discuss potential negative impacts of BCIs on patients' autonomy: "Imagine an application of a passive BCI, which is combined with smartphone use, where the computer receives and saves information about the mental states of the user while using certain websites. After a while the BCI provides the user only with websites that might suite her current mental state. The computer might only provide the user with those algorithm-derived information and options to act that were based on brain activity under similar contexts. In short, this technology could not only leave people more constrained to their own past states and decisions, but also limit their development of habits, thinking patterns, and actions." (Friedrich et al. 2021, pp. 22–23).

In this scenario the users' knowledge is restricted or altered by the BCI, and this could undermine and decrease the ability to choose and enact different actions. If the computer system "makes decisions" for the user based on algorithms, without informing the user or giving the user the option to decide or override the computer-based decision, then the

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No.101006249.

person's ability for autonomous action could be significantly compromised (Friedrich et al. 2021, p. 23).

However, not all patients feel that they are losing control: "They trust the researchers that are controlling it. They don't feel like there's any questionable agency to be concerned with" (Muñoz et al. 2020, p. 6). A way to preserve autonomy is to give the patient control with a "red button" that makes it possible to remove themselves out of adaptive stimulation into a conventional stimulation. This button creates an operational translation of free will, even if it may remain an illusion. While some researchers think patients definitively need such a button to override unwanted stimulation and stay autonomous and self-determined, others also see potential risks of giving patients control over their stimulation (Muñoz et al. 2020, p. 6). The level of control patients should have over stimulation may also depend on which areas of the brain are being stimulated.

The opinion that aDBS induces changes to PIAAAS is quite strongly represented in the neuroethics debate. However, a group of researchers critically assessed evidence about such putative effects (Gilbert, Viaña, et al. 2018). They conducted a literature review of more than 1535 articles to investigate the prevalence of scientific evidence regarding potential DBS-induced changes and observed an increase in the number of publications in theoretical neuroethics that mention putative DBS-induced changes to patients' postoperative PIAAAS. However, they found a critical lack of primary empirical studies corroborating these claims.

One expert stressed the importance of diversity when setting basic rules, and this can be done through a meeting of cultures, otherwise there will be a loss of identification and variety. They went on to say, ultimately, we will end up as human machines so it is important to retain diversity and innovation in order to keep an individual's identity.

Establishing checks and balances with regard to autonomy and self-determination in neurotechnologies means paying particular attention to the following questions:

- Can patients autonomously give consent to using a neurotechnology, for instance aDBS or BCIs?
- Can surrogates provide consent if the patient is not able to do so?
- Is the patient able to give free and informed consent or do they lack decision-making capacity?
- Is the patient's autonomy impaired due to mental or cognitive conditions?
- Is the patient able to understand the risks for their body/mental states?
- Is the patient able to understand how data will be collected and stored?
- Is the patient aware of experiences of other patients using similar devices?
- Is the patient aware of negative impacts on autonomy that might occur while using the neurotechnological device?

Recommendation for aDBS: Patients and clinicians might discuss those items for which there is a change from the perspective of the patient pre- and post-DBS surgery, and then consider whether the patient approves or disapproves of this change, and how strongly. If the patient has trouble with a change on a particular item, the clinical team and patient could work together to identify potential ways of managing perceived negative impacts regarding that element of autonomy (Zuk and Lázaro-Muñoz 2021).

- Are there any options for the patient to opt out after opting in?

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No.101006249.

- Is the procedure reversible or irreversible?
- Is there a "red button" for turning off the device?
- Are any of the data stored for a closed-loop system device sensitive data?
- Are measures undertaken to store data safe and secure? Advice: Integrate all stakeholder groups – including patients and caregivers – in the development of control and safety policies.
- Does the patient know what is happening?
- Does the patient know that algorithms sometimes work without knowledge of the user?
- Are some of the patient's decisions or actions induced through external influence? If yes, who is making the decision? Is the patient aware of this?
- If autonomy is impaired due to the neurotechnological interventions, is there any option for the patient to opt out after opting in? For aDBS: Patients should have a controller and at any point can remove themselves out of adaptive stimulation into conventional stimulation.

3.4.2. Responsibility: Whose responsibility is involved in the use and misuse of neurotechnologies?

As discussed in section 3.1.1, neurotechnologies pose questions about free will. We submit that the conceptual limitations on free will do not remove the validity of moral responsibility. Even though complete freedom of the will may not exist, and there may be neurological or other limitations on how a person chooses to act, moral responsibility remains a valid concept (Brass et al. 2019; Nahmias 2012). What is modified by the use of neurotechnology is the distribution of responsibility.

Neurotechnologies, especially consumer BCIs used during morally significant activities, can limit the freedom of the individual to act. They can introduce in the action an outside intent, for example a BCI user could be induced to act as intended by the BCI manufacturer or operator. In such cases, the responsibility may be shared between the subject of the action and third parties.

Examples of free will limitations include the control of hormones by the BCI, increased or decreased focus, and other neurological manipulation (Vishwakarma et al. 2020). Influenced by these changes, a BCI user may not be able to steer their action away from harm. For example, a car accident while the driver was using a BCI can lead to partial responsibility of the BCI manufacturer. As with other ethical judgments, much depends on the availability of causal influences (how constraining was the BCI) and intention (what was the BCI manufacturer trying to achieve). There is an ongoing debate on how the responsibility can or should be shared. For discussion, see Work Package three.

Another potentially important area where a BCI might be significant for moral responsibility is the time from thought to action. Even Libet, who performed the neuroscientific experiments questioning free will, held that the conscious will has a "veto power" (Libet 1999), so that it can stop or refrain from an impulse to do something. With the advent of BCIs, the barrier from thought to action is shortened. For example, currently interacting virtually, even in the metaverse, requires performing a series of actions to achieve a decision or an action (writing a message requires typing, purchasing requires choosing and

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No.101006249.

checking items out, etc.). A BCI can remove the extra steps needed to perform an action, especially online. It can take neurological signals directly and translate them into the digital data needed for the action (a sheer thought would be sufficient for sending a message, purchasing an item, etc.). This may be limited through regulation; thus, Klein argues that BCI activity should be only "assistive" and not enabling (Klein 2017).

An interviewed expert mentioned that there needs to be a change of rules and a change in mindset, it should be globally unacceptable to misuse data. It is time for re-education in new ways that are ethically and morally viable. Furthermore, some experts suggested that there are unlikely to be new ethical issues that will arise, just the same ethical concerns will be present. However, they felt that further exploration and study will go into the 'fake use of ethics' - i.e., 'ethics washing' for example companies and industry will try to give an importance to ethics by giving this more attention as a way of being perceived to care and be interested but in reality, this will be fake. The expert feels this concept will grow more in the future.

Establishing checks and balances with regard to responsibility in neurotechnologies means paying particular attention to the following questions:

- Do any of neurotechnology applications (e.g., BCI) shorten or remove the barrier to action?
- How are the intentions of the device manufacturer documented? Are they declared at the time of purchase or evolving?
- Is there a traceability protocol for the neurotechnological device to keep track of its influence on the user?
- How is the responsibility split in case of conscious will limitations? Are there guidelines on how it should be split that fit to the technical specifications of the device?
- Is the application assistive or causative in relation to user actions?

3.4.3. Privacy: Should mental contents be decoded? What is the status of the decoded mental data?

Mental privacy is the idea that people should have control over the data produced by their neurological activity. With the advent of consumer brain computer interfaces (BCIs) mental privacy will become a widespread issue (Boto et al. 2018). One ethically challenging aspect of it is that consent for the use of mental data can mean that it is unlimited. Usually, when people consent to handling their data, it concerns only a particular dataset or certain data types (for example, they name, age, interests, etc.) However, in the case of mental data, the potential to extract mental imagery is unlimited and thus a user might be consenting to handling data that they do not conceive or understand at the moment of providing the consent. This unlimited scope of mental data calls for a reconsideration of informed consent.

This is a pressing concern since the increasing application of BCI devices in commercial settings opens new possibilities for collecting neural data outside the clinical domain and using it for gainful purposes, like neuromarketing. Neurological data can be used by neuromarketing companies to study and influence consumer behaviour and perception. Neuroimaging devices and applications in commercial settings are particularly worrying since they are not required to comply with the same guidelines as clinical research. Ienca et al. warn that "Unlike clinical research, neuromarketing companies are free to conduct neuroimaging studies of humans in the consumer space without formal approval from an ethics committee and rigorous informed

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No.101006249.

consent from study participants" (Ienca et al. 2018). The potential implications of applying neuromarketing also include effects on the democratic processes and foreign influences of the public through instigating hate and spreading misinformation, thus the issues are both ethically and politically challenging.

The development of neurotechnology significantly contributes to the extraction of increasingly sensitive personal data with the collection of neural data (Goering et al. 2021). The data economy, its extractive nature, and the continuous growth of personal data collection need to be analyzed from the perspective of power and justice, accounting for the "data-subject/data-controller power imbalances" (Delacroix and Wagner 2021b, p. 9).

Contrary to the conventional DBS, data privacy and security is especially problematic in aDBS. Only for aDBS is the neural activity data measured and stored. Whether these data are sensitive is open to debate. Some researchers think that such data do not allow for the identification of persons and therefore are not sensitive; others believe that neural data could become sensitive in the future (Aggarwal and Chugh 2020, p. 160).

From the point of view of power balance, as described in section 1.3.4, some of the big technology companies can be considered "gatekeepers" (European Commission 2020b). Their dominant position needs to be understood within the new data economy of contemporary societies. These companies are powerful because of their financial resources but, even more importantly, their hegemonic position comes from two technological factors: a) their control over protocols for neurological data collection; b) the massive amount of data they hold that is necessary for learning. As Sadowski highlights, "data – and the accumulation of data – is a core component of political economy in the 21st century" (Sadowski 2019, p. 1) Furthermore, this political economy is based on practices of data accumulation characterized by "relations of inequity, extraction, and exploitation, wherein data is taken without meaningful consent and fair compensation for the producers and sources of that data" (Sadowski 2019, p. 2). Such power imbalance needs to be accounted for in the ethical analysis.

Further challenges may arise in the context of open-source software for the BCIs. For example, in the ethnographies one neurotechnology company claimed to be using open-source software: "All the […] hardware comes with free software, which is provided in a form of software development kit (SDK) giving users the freedom to integrate […] software into their experiments, applications or products. Multiple programming languages are supported to ease the integration process." Open source code can be more prone to security issues that can eventually lead to privacy leaks and misuses.

Establishing checks and balances with regard to privacy in neurotechnologies means paying particular attention to the following questions:

- Do any of neuro devices collect personal data (neurological data, brain maps, neurological response analysis, etc.)?
- How is this data communicated and stored?
- What are the opt-out options from the collection of such data?
- Does the application rely on the collection of such data and is the use of it justified?
- Can the identity of the user be traced or reverse-engineered from the neurological data?
- Who gets access to neurological data and on what terms?

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No.101006249.

3.4.4. Risk reduction: How can physical and digital safety be ensured?

Until now, there have been no significant research showing neurotechnology safety in off-label enhancement or in well-being, i.e., applications for relaxation, concentration, improving sleep, etc. Only smaller lab studies exist. More insight is needed into the lasting real-world effects. The application of neurostimulation for neural enhancement is also not entirely proven. The off-label enhancement use is likely to occur with lightly invasive and cheap devices such as transcranial direct-current stimulation (Landhuis 2017).

Risk management depends on the evaluation of risks and benefits. The main stakeholders in the risk considerations are patients with their needs and the side-effects they may eventually suffer on one hand, and the medical community and its knowledge of positive and negative outcomes, and thus the professional recommendations and guidelines (e.g., FDA, EMA) on the other hand.

Testing enhancement is ethically problematic, for it requires healthy individuals to undertake enhancement procedures that are not proved to be safe. Thus, the risk and outcome calculation is skewed differently than in therapeutic case, in which refraining from treatment would maintain illness. In the case of enhancement, not undergoing a procedure is safe, whereas experimenting with it might cause damage. As discussed in section 3.4.5, this also raises ethical issues with informed consent. Despite these problems, some applications are already used, or close to being used, both in the medical and non-medical settings.

For example, the neurosurgical procedure necessary for conventional DBS and aDBS is highly invasive. Risks such as strokes or seizures are not unlikely, and further unknown risks are possible, e.g., undesired psychological symptoms while treating movement disorders. The classical example is the use of DBS for treating Parkinson's disease, which in some cases also leads to negative effects on mood (Accolla and Pollo 2019). With aDBS, however, the stimulation can minimized to the patients' individual needs. Therefore, one might argue that the potential magnitude of unknown side-effects is reduced compared to DBS.

One of the devices analyzed in the ethnography aimed to treat people who have had strokes, Alzheimer and dementia, or long-term effects of COVID. The treatment is described as "boosting the brain's natural plasticity without touching it." The company underlines the fact that the glasses do not contain electrodes or perform any direct stimulation of the brain, and in this sense, it is marketed as the least invasive neurotechnologies. This can be seen as a way for a business to tackle the ethical issue of safety of neurotechnologies.

A large-scale risk of societal change is the medicalization and pathologization of mental performance. What is considered as incomplete well-being or disease is open to societal debate and evolves in time (Illich 1976). In principle, any type of cognitive lapse could be considered a pathology that needs to be corrected by neurological means. This would involve a strong notion of neural normalcy and be detrimental to neural diversity, which is an issue often raised in psychiatry (Schrader et al. 2013). However, this is even more pressing in human enhancement where risks are not proportionate, since there is no disease-based need to apply the neurotechnologies (World Health Organization and Council for International Organizations of Medical Sciences 2017). In most of the enhancement cases, the interventions are not necessary and chosen by users with no disease conditions on the basis of a simple desire to feel or perform "better". This desire induces risk (following the narrative of lay ethics "be

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No.101006249.

careful what you wish for") and responsibility. Also, the responsibility to regulate the use of neurotechnologies in order to mitigate the risks belongs with public authorities and health authorities that may prohibit or limit non-therapeutic uses.

Despite risks, some activists advocate a complete autonomy of the person to modify their body, including the nervous system. Historically, the transhumanists have embraced the idea that evolution required human, technological intervention, to allow humans to progress more quickly (Bostrom 2005b, p. 15). The decision to partake in enhancement is seen as a personal and individual choice (Hughes 2013; Jensen 2020, p. 19).

Establishing checks and balances with regard to risk reduction in neurotechnologies means paying particular attention to the following questions:

- Does a particular neurotechnology have only therapeutic uses or can it also be applied to enhancement?
- What potential consequences are involved in non-therapeutic uses?
- How are the risks of enhancement justified?
- Does an application promote neural normalcy and reduce neural diversity?
- How is the baseline for neural normalcy established?
- Who makes the decision? Is the decision based on subjective (desire, pain) or objective (pathology, handicap) factors?

3.4.5. Informed consent: What specific privacy concerns do neurotechnologies raise? What is the meaning of the informed consent in neurotechnology applications?

Some neurotechnologies require highly invasive treatments requiring that patients give informed consent before the start of the treatment. This is standard for any diagnostic, therapeutic, and clinical research procedure. In order to obtain informed consent, several requirements need to be fulfilled, including voluntariness of the decision-making process, accurate and complete information disclosure, and the patient's mental capacity to consent (Berg et al. 2001).

One specific challenge to the standard procedure is raised in DBS. Many of the disorders for which it is considered as treatment may be associated with memory changes and impaired perception. As a result, informed consent for neurosurgical procedures may not be meaningful or even possible, because patients will not adequately understand the information. Depression, anxiety and compulsivity are also common in DBS candidates and may be associated with an impaired capacity to give informed consent (Mandarelli et al. 2018). The main ethical challenge with Parkinson's patients that are good candidates for DBS treatment is that due to being in a difficult state they hope a miraculous effect from DBS, which is often true also for family members or surrogates. Therefore, they often ignore side effects that include impact on mood, impulsivity, hypersexuality, compulsive buying, until they become strong, particularly during the first 2 months after surgery in a significant portion of the patients.

Giving informed consent to aDBS means consenting to ongoing stimulation changes which likely occur outside of a patient's conscious awareness (Gilbert, O'Brien, et al. 2018;

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No.101006249.

Gilbert, Viaña, et al. 2018). It is an open question whether patients can robustly consent to automatic, moment-to-moment changes in stimulation (Muñoz et al. 2020, p. 5).

One expert thinks that it is important to build trust and ensure privacy of the patient. They went on to suggest that this is a double face problem, as we need to be patient centered and respect patient goals i.e., consenting to who will have access to their data and who will use it, so there is infringement of freedom and ethics. Even if consent is given for GDPR purposes, (one) does not know who is using (the) data.

Consent requires a capable mind and a no undue influence on decisions. Neurotechnologies offer the potential to interfere (alter, mimic or enhance) with these essential attributes in a unique way. The International Bioethics Committee has already expressed its opinion on the notion and applications of informed consent (UNESCO 2008, 2021), stressing that consent is the chief assurance of the autonomy of human subjects in healthcare. Autonomy and responsibility, as well as consent and protection of persons without the capacity to consent, are addressed in Articles 5, 6 and 7 of the Universal Declaration on Bioethics and Human Rights (UDBHR). Whether these provisions are adequate for emerging neurotechnologies is an open issue.

Establishing checks and balances with regard to informed consent in neurotechnologies means paying particular attention to the following questions:

- Does a therapy using a neurotechnological device involve conditions that might impair free and informed consent?
- How explainable and understandable the device and the underlying technology are to the patients and their support network, family members, and surrogates?
- Are the changes to neurological activity induced by neurotechnological devices noticeable at the level of conscious awareness?
- What is the protocol for the revocation of consent?
- Are changes made to neurological patterns and activity reversible?
- Can a third party be involved to evaluate the benefit/risk consideration of aDBS in a given patient on a more neutral basis?

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No.101006249.

4. Ethical Analysis: Climate Engineering

Climate engineering (CE), also known as geoengineering, refers to "… the deliberate large-scale intervention in the Earth's climate system, in order to moderate global warming" (Shepherd et al. 2009, p. 1). In both technical and normative literature it is conventional to distinguish between two forms of CE: i) Solar Radiation Management (SRM) techniques, which aim to reflect some sunlight and heat back into space; and ii) Carbon Dioxide Removal (CDR) techniques (also known as "Negative Emissions techniques'), which remove atmospheric CO2 and store it in geological, terrestrial, or oceanic reservoirs. In addition to these different aims, a second key difference between SRM and CDR is temporal: the former is believed to be fast-acting, while CDR is slow-acting and operates on similar timescales to conventional mitigation. The deliberate nature of CE as a policy intervention distinguishes it from any similar effects produced unintentionally or by the Earth's systems and processes. This deliberate nature of CE also marks these technologies as ethically distinctive (Jamieson 1996).

The terms "climate engineering" and "geoengineering" are contentious because of the significant normative and practical differences between SRM and CDR, and between individual techniques (Heyward 2013; Minx et al. 2018). The IPCC AR6 avoids both terms, and prefers to use only SRM and CDR. Further, the AR6 chose to reclassify CDR as a form of climate mitigation (IPCC 2022, Chapter 12), departing from their classification of CDR as geoengineering in the Fifth Assessment Report (AR5) (IPCC 2014, Technical Summary, 60). It is also disputed whether to refer to climate engineering as a technology. Scientific assessments and scholarly analyses of climate engineering have referred to climate engineering as "techniques" (e.g., Royal Society 2009), and as "technologies" (e.g., IPCC 2014). Seemingly to avoid this controversy (but departing from the scientific literature), the AR6 refers to CDR "methods" (IPCC 2022, Chapter 12), and reclassifies SRM as Solar Radiation *Modification* "proposals" (IPCC 2022, Chapter 14, 56), without a clear scientific rationale. Instead, this report prefers to use CE techniques for two reasons, in addition to the ones explained in the TechEthos Glossary. First, there is an argument to reserve the term "technology" for functioning socio-technical systems rather than untested or speculative proposals (Rayner 2010), which describes at least some prominent forms of CE. Second, some forms of CDR that appear in the AR6 are difficult to understand as technologies at all, having been practiced for centuries or even millennia (e.g., improved forest management).

The AR6 also introduces a new distinction between "net negative" emissions and CDR. Although these terms were previously often used synonymously, net negative emissions now refers to a situation in which more greenhouse gases are removed from the atmosphere than are emitted, and includes emissions of other gases beyond CO2. Instead, the term CDR is reserved for individual techniques that remove CO2, and does not take this system-level perspective (see AR6 glossary, available at: https://www.ipcc.ch/report/sr15/glossary/).

4.1. Carbon dioxide removal (CDR) techniques

The following is a summary of key findings from Chapter 12 of the IPCC AR6, which explores CDR potentials, costs, risks and feasibility constraints.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No.101006249.

4.1.1. Bioenergy with Carbon Capture and Storage (BECCS)

BECCS is the combination of biomass used to generate bioenergy, with CCS to prevent emissions reaching the atmosphere. Because the growth of biomass sequesters CO2 from the ambient air, BECCS can provide net atmospheric CO2 removals, unlike the application of CCS on fossil infrastructure. BECCS is the most frequently modelled technology in Integrated Assessment Models, with many scenarios featuring substantial BECCS implementation to correct for "overshooting" the global carbon budget. The AR6 estimates that BECCS could remove between 0.5-11 gtCO2 per year, at the cost of $US 15-400 per ton of CO2 removed.

The AR6 cites several benefits of biomass-based forms of CDR, including BECCS and biochar, such as supporting agroecology and biodiversity through the use of suitable grasses and woody plants. However, several important risks of BECCS have been discussed, including land competition between biomass production and agriculture, freshwater use and phosphorous for fertilizer, the destruction of natural ecosystems for biomass production, and consequent diminishing of biodiversity and vital ecosystem services. Biomass plantations can also be invasive monoculture crops, which further harm local biodiversity and displace existing ecosystems. The AR6 points to governance as a critical determinant of whether BECCS will have positive or negative outcomes related to land use.

4.1.2. Direct Air Capture with Carbon Capture and Storage (DACCS)

DACCS combines CCS with chemical processes to capture CO2 from ambient air, which is then stored underground. Storage in geological reservoirs or in mineral forms would remove CO2 for up to 1000 years. The AR6 estimates the CO2 removal potential of DACCS at 5-40 gtCO2 per year, potentially the largest of any form of CDR (IPCC 2022, Ch. 12), at an estimated cost of $US 100-300 per ton of CO2 The primary barriers to DACCS include high economic costs and large non-fossil energy requirements (Nemet et al. 2018), rather than potential risks or side-effects of implementation. Economic costs would fall through investment

4.1.3. Enhanced Weathering (EW)

EW removes atmospheric CO2 by spreading small particles of ground silicate and carbonate rock onto soils, coasts or oceans. Rocks containing silicate and carbonate naturally absorb CO2, yet over very slow (geological) timescales. By spreading small particles of silicate and carbonate, EW increases the total surface area of the planet that experiences this weathering effect, and encourages weathering on surfaces that would not ordinarily experience it. The AR6 estimates that EW could remove 2-4 per ton of CO2 per year, and is estimated to cost between $US 50-200. Potential risks of EW include the effects of mining, and worsened air quality from rock dust spread onto soil, and the degradation of local water quality. Potential benefits include increasing crop yields or biomass production when spread onto soils, and the reduction of soil erosion.

4.1.4. Afforestation and Reforestation

Afforestation refers to planting forests upon land where forests have not historically occurred, while Reforestation refers to restoring forests upon deforested land. The AR6 estimates that afforestation and reforestation could remove between 0.5-10 gt CO2 per year, at cost estimates ranging from $US 0-240 per ton of CO removed, with implementation costs closest to zero in more favourable locations. Afforestation and reforestation are associated

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No.101006249.

with a range of co-benefits, including employment and the support of local livelihoods, improved biodiversity, and an improved supply of renewable wood. Risks to forests affecting the permanence of carbon removal are wildfires and diseases.

4.1.5. Ocean alkalinity enhancement (OA)

OA is a process that reduces the pH of the ocean over several decades or longer, due to the absorption of atmospheric CO2 Since this process occurs naturally, anthropogenic OA refers to intentionally adding alkaline materials into the ocean either directly or in a dissolved form upstream. The AR6 estimates the CDR potential of OA between 1-100 gtCO2 per year, at costs ranging from $US 40-260 per ton of carbon removed. Risks to forests affecting the permanence of carbon removal are wildfires and diseases.

4.1.6. Ocean Fertilization (OF)

OF aims to increase the rate at which the ocean draws down atmospheric CO2 and sequesters it in the deep oceans through the growth of phytoplankton. This can be achieved by adding nutrients such as nitrogen or phosphorous, or trace minerals such as iron depending upon which is lacking. The AR6 estimates the CDR potential of OF at 1-3 gtCO2 per year, with costs ranging from $US 50-500 per ton of carbon removed. Serious risks with OF are acknowledged, including the added nutrients restructuring marine ecosystems, causing the deep ocean to become more acidic, and for the extra carbon to returned to the atmosphere on a timescale anywhere from decades to millennia. OF is governed by an existing international treaty, the London Protocol.

4.1.7. Carbon sequestration in agriculture

CDR can be achieved in connection with agriculture in several ways. Soil carbon sequestration refers to land management practices that increase the carbon content of soils by adding organic matter, creating a net CO2 removal through the growth and absorption of organic matter. Biochar is a carbon-rich material produced by heating biomass with limited oxygen, and then adding this to soils, which results in a net removal of atmospheric CO2 Biochar is estimated to achieve between 0.3-6.6 gtCO2 carbon removal, at the cost of $US 10-345 per ton of CO2, while soil carbon sequestration could remove between 0.6-9.3 gtCO2 at the cost of $US 45-100 per ton of CO2 Biochar risks include additional particulate and greenhouse emissions during combustion, and the degradation of biodiversity and existing carbon storage in land from unsustainable biomass production. At the same time, biochar could increase crop yields and reduce nitrogen emissions from soils. Soil carbon sequestration risks include the release of additional nitrous oxide from soils with higher nitrogen content, while benefits include improving soil quality and hence agricultural productivity, as well as resilience to soil loss.

4.1.8. Related non-CDR techniques

Two further techniques often mentioned in connection with CDR but which should not be confused with it are Carbon Capture and Storage (CCS) and Carbon Capture and Utilization (CCU). CCS involves the capture of carbon emissions from industrial production or energy combustion, which is then placed in long-term storage. CCU is the commercial re-use of carbon captured using CCS in new products. As the AR6 explains, CCS and CCU are not forms of CDR

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No.101006249.

because neither removes CO2 from the atmosphere (IPCC 2022, Chapter 12, 36). Nonetheless, both CCS and CCU can be combined with CDR methods if they are applied to CO2 removed from the atmosphere (either through the growth of biomass or directly extracted from the air using DACCS), but only when CO2 is stored for a relatively long period of time. CCU which utilizes CO2 for the production of short-lived consumer products such as carbonated drinks cannot be considered as CDR.

4.2. Solar radiation management (SRM) techniques

The following is a summary of the IPCC Special Report, Chapter 4 on SRM (IPCC 2018). The AR6 does not include an equivalent chapter on SRM, but focuses upon CDR since some CDR methods are being used already, and are included in the net zero climate plans of many nations. In contrast, the role of SRM remains far more uncertain.

One expert believes that in the future they do not anticipate the use of SRM on any major scale, but maybe some climate engineering technologies will be used to cool regionally and could potentially provide the benefit of less intense heat waves or cooling specific regions.

4.2.1. Stratospheric Aerosol Interventions (SAI)

By far the most researched SRM technique (IPCC 2018, 347), SAI involves the injection of gas in the stratosphere, which converts into aerosols that block some incoming solar radiation. The IPCC Special Report considered SAI as a fast-acting method to immediately mask climate impacts during a period of emissions "overshoot", which would increase the time available for mitigation to lower atmospheric CO2 concentrations. SAI is considered to be very cost-effective, with estimates between $US 1-10 billion per year to mask the climate effects of approximately half a degree of temperature overshoot (i.e., rendering climate impacts of 2C into 1.5C) (IPCC 2018, 347). Risks include changing precipitation patterns and air circulation, with adverse impacts on monsoon areas in particular, harming agriculture and many ecosystem types. For sulfate-based SAI, risks include disrupting the chemistry of the stratosphere, affecting the length of time of methane storage, the formation of ice, and the microphysics of clouds. Sulfate-based SAI would also deplete stratospheric ozone, leading to adverse health impacts. At the same time, the masking of some climate impacts via SAI would decrease heat-related mortality.

4.2.2. Marine Cloud Brightening (MCB)

MCB involves spraying sea salt or similar particles into marine clouds, increasing their reflectivity and blocking some incoming solar radiation. The effect of MCB upon radiative forcing appears to be more limited than for SAI (IPCC 2018), meaning it is likely to be utilized only in conjunction with other techniques. MCB is thought to affect regional rainfall patterns, but much uncertainty remains. MCB is also thought to reduce the intensity of hurricanes, and the likelihood of minor crop failures.

4.2.3. Ground-based Albedo Modification (GBAM)

GBAM techniques aim to increase the reflectivity of land surfaces, which deflect incoming solar radiation. This includes whitening roofs, land management practices (e.g., no-till farming), covering deserts or glaciers with reflective sheeting, and increasing the reflectivity of the ocean. While small-scale implementation of GBAM would have little effect

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No.101006249.

upon radiative forcing, regional scale implementation of these techniques could reduce global radiative forcing by between 1-3C (IPCC 2018, 348). GBAM could also affect monsoon precipitation, although there is greater uncertainty compared with SAI.

Values and principles	Climate engineering techniques		
		CDR	SRM
	Distributive justice	x	x
	Procedural justice	x	x
	Future responsibility	x	
	Side effects	x	
	Research ethics		x
	Termination shock		x

4.3. Core ethical dilemmas in climate engineering

The following is a synthesis of literature reviews on the ethics of CE (Betz and Cacean 2012; Pamplany et al. 2020; Preston 2013). As Preston (2013) and Pamplany et al. (2020) note, much of the ethical literature on CE focuses primarily upon SRM, and indeed with one particular SRM technique, namely SAI. Nonetheless, recent ethical analysis has drawn attention to the ethical issues distinctively raised by CDR (Lenzi 2018; Schübel and Wallimann-Helmer 2021; Shue 2017).

4.3.1. Moral hazard: does climate engineering undermine climate mitigation?

The basic concern with moral hazard is of perverse incentivization of risky, and especially, of morally problematic behaviour. This is essentially a control problem in the case of complex systems, which entails a responsibility problem. If artificial changes are promised as solutions to the climate crisis, can meaningful climate change mitigation still be implemented? Who can be held responsible for the consequences of such promises?

The moral hazard may also be described as a “mitigation obstruction” effect (Betz and Cacean 2012). However, the way in which moral hazard may manifest changes character depending on whether SRM or CDR is envisioned. The worry that CE would obstruct or prevent some conventional climate mitigation has been voiced since the earliest ethical analyses (Jamieson 1996), and remains a major theme of contemporary ethical debate (Gardiner 2010;

This project has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement No.101006249.

Preston 2013). As Preston (2013) notes, the reluctance of scientists to discuss the prospect of CE prior to Nobel laureate Paul Crutzen's (Crutzen 2006) call to break the taboo surrounding it was largely motivated by a concern with moral hazard.

For CDR, the moral hazard effect is most concerning when the presumed future availability of CDR at large scales encourages slower near-term mitigation. For this reason, CDR has been labelled a "moral hazard par excellence" (Anderson and Peters 2016). As Lenzi (Lenzi 2018) shows, CDR presents a moral hazard in two distinct ways.

First, the introduction of CDR within mitigation scenario modelling inevitably obstructs some near-term mitigation, because CDR lowers the aggregate costs of mitigation over the century, and hence the cost-optimization logic of such modelling displaces some near-term mitigation in favour of the now cheaper use of CDR later in the century.

Second, and perhaps more concerning, is the degree to which the availability of CDR in climate models displaces near-term mitigation at the political level. While it is disputed whether CDR already has displaced near-term mitigation, the worry here is that the existence of scenario modelling with very large future reliance on CDR and less urgent near-term mitigation may encourage policymakers to slow-pedal mitigation.

In this respect, the AR6 reclassification of CDR as mitigation may mask important dangers of moral hazard, due to the common expectation in climate mitigation modelling that CDR will be available at large scales in future to correct for near-term emissions "overshooting", in which the atmospheric concentration of CO2 initially exceeds the global carbon budget for 2C, but is brought down later in the century via CDR. Indeed, CDR was primarily utilized in the AR5 by scenario models to correct for "overshooting", and its inclusion rendered even significantly delayed emissions trajectories compatible with the 2C target (Lenzi 2018). As a result, there remains a key concern that the presumed availability of large-scale CDR in future will encourage delayed short-term mitigation, i.e., a classic "moral hazard" or "mitigation deterrence" effect.

For SRM, and for SAI in particular, the moral hazard concern is both that mitigation may be slower if this technique is presumed to be available, and that mitigation may even be abandoned entirely if policymakers believe it could adequately mask the effects of climate change. Indeed, some proponents of SAI have argued in favour of it as a relatively cheap and easy technological fix for climate change (Barrett 2008), contrasting its low implementation costs with the much higher costs of complete decarbonization.

As noted in section 4.1.1, empirical estimations indicate that the risk of harmful side-effects from SAI increase in proportion to the atmospheric concentration of CO2, and thus that the implementation of SAI becomes more concerning if mitigation is delayed. Nonetheless, as with the moral corruption concern explored below, the moral hazard risk merely requires policymakers and the public to believe that SAI would sufficiently mask the effects of climate change, even if this is empirically dubious. Thus, prominent climate ethicists have worried that SAI presents an acute moral hazard against mitigation (Gardiner 2010). Indeed, as Preston (Preston 2013) notes, even advocates of SAI seem to worry about the moral hazard, cautioning against the idea that it could substitute for mitigation.

One expert mentioned that a major risk is that all these technologies distract from actual climate policy or mitigation policies. There is this assumption that we can do carbon

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No.101006249.

capture on a massive scale, but we do not know if it is possible. Most people suggest that it probably is not possible, but still our climate policy is based on those assumptions, and it still allows us, or it allows our politicians and our policymakers to postpone the very hard cut (decisions) that need to be made.

4.3.2. Moral corruption: Does climate engineering reflect a self-serving interest in avoiding politically difficult transitions away from fossil fuels?

Closely related to the problem of moral hazard is the problem of moral corruption, first introduced by Gardiner (Gardiner 2010). This is in essence the worry that the availability of CE may encourage self-serving rationalizations among the present generation that they do not need to mitigate more rapidly now. Such rationalizations, according to Gardiner, are strategic attempts to evade our moral responsibility to avoid the worst effects of climate change. Yet they become much more tempting the further in the future climate harms are predicted to fall, an effect that encourages the present generation to "pass the buck" of mitigation onto future generations. Indeed, Gardiner argues that this dynamic may be iterated across generations, with each generation passing the buck and becoming more and more prone to moral corruption (Gardiner 2011).

While moral corruption was originally introduced in relation to SRM, it also bears upon CDR since this also presents the present generation with the possibility of delaying near-term mitigation and passing greater economic burdens and climate harms onto future generations (Lenzi 2018). However, given that the economic feasibility of CDR is closely tied to conventional mitigation policies (such as carbon pricing), the danger of moral corruption appears somewhat less severe for CDR than for SRM, which has already been discussed by some politicians (e.g., US Republicans such as Newt Gingrich) as an alternative to climate mitigation.

4.3.3. Hubris: Can climate engineering be justified by limited human foresight?

Since the first ethical analysis of CE (Jamieson 1996), it has been recognized that by even contemplating the intentional modification of the global climate human beings seemed to be "playing God" (Grinbaum and Groves 2013), calling into question traditional conceptions of the proper relationship between humanity and the rest of nature. As Preston (Preston 2013) notes, the very idea of CE seems to reflect an attitude of control or dominance over nature, which many environmental ethicists have blamed for the climate and environmental crisis. Again, however, the concept of hubris was typically applied only to SRM, since this can be likened to human beings holding a "global thermostat", deciding to turn it to a setting that favours human beings (or more plausibly, some human beings). This idea of the climate system, and the planet more generally, being made to serve human beings has struck many commentators as ethically problematic.

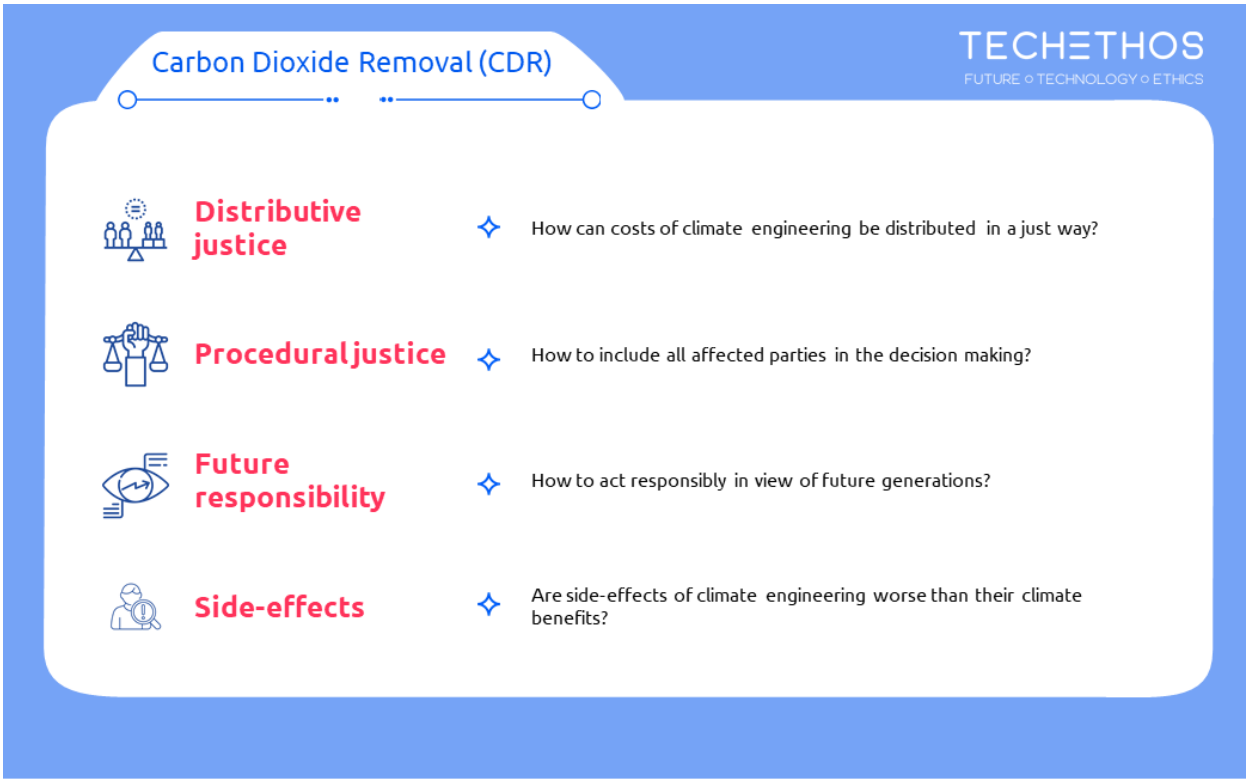
In more recent work, hubris has been applied to large-scale CDR implementation (Lenzi 2018; Lenzi et al. 2018; Minx et al. 2018). The reason is that the sort of CDR implementation often evident in climate mitigation models implies an unjustified, even arrogant level of control over the global carbon cycle and that is not supported by current knowledge of carbon cycle feedbacks. Implementing CDR at very large scales may greatly overestimate both

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No.101006249.

feasibility assumptions and safety, while underestimating harmful effects. The extent of such CDR reliance in 1.5C consistent pathways can require removing 400-1000 gigatonnes of CO2, which would mean storing 10-25 years of global CO2 emissions underground (Lenzi 2018). The concern here is that such modelling arrogantly overestimates our current knowledge and capabilities, and hence our collective ability to control the global carbon cycle.

The ethnographies revealed some of the communication strategies used by CDR businesses that relate to hubris. A company proposing CDR and carbon re-utilisation state on their website: "We will capture carbon dioxide and combine this with hydrogen, made from renewable electricity and water, to produce carbon neutral fuel, eMethanol." Further, the webpage describes the lucrative investment opportunities in terms and invite investors that are interested in "a bankable and sustainable investment opportunity". This confidence suggests a monolithic concept of time in which the technology will develop confidently along a predetermined path.

Another company analyzed in the ethnographies had a webpage showing images of its technical facilities that are rather ordinary. The site was not portrayed to look majestic or spectacular, as one would expect from the structure surrounding an emergent technology. The choice of ordinary elements aims at distinguishing itself from the image of flashy disruptive technologies. Although linguistically the communication emphasizes the great capacities of its technology, visually it clearly avoids flashiness.



This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No.101006249.

4.4. Values and principles in CDR

4.4.1. Distributive justice: How can costs of climate engineering be distributed in a just way?

In Pamplany et al. (Pamplany et al. 2020) recent review of the ethics of geoengineering, they find that the dominant concern is of distributive justice, finding 113 articles (from a total survey of 304) citing justice concerns. This number excludes concerns with side-effects of CE and the imposition of risk on future generations, both of which are also justice concerns.

For instance, CDR changes the distribution of mitigation across generations, potentially pushing more mitigation onto future generations, coupled with a greater reliance upon CDR. As such, greater reliance upon CDR may unfairly burden future generations with higher decarbonization costs and higher risks of unjust side-effects (Lenzi 2018, 2021; Shue 2017, 2018). The potential for either national or multilateral financing of CDR, and the siting of CDR facilities, also raise issues of fair burden sharing under the UNFCCC principle of "common but differentiated responsibilities" (Lenzi 2021).

Nonetheless, in terms of harm avoidance, CDR can in principle reduce climate harms. The IPCC recognizes that CDR implementation has the potential to render more stringent climate stabilization targets achievable (IPCC 2018; 2022), which would also be highly desirable from the perspective of distributive justice (Lenzi 2021). At the same time, the poor implementation of CDR, including the sourcing of biomass for BECCS as an export commodity or grown domestically by countries in the Global North, can harm vulnerable people by undermining food and water security and biodiversity (Anderson and Peters 2016; Lenzi 2018; Lenzi et al. 2021; Shue 2017).

An expert indicated that another aspect of justice around carbon capture and storage, is that these take up a lot of land and this may create land pressures for people who are now using it for their own agriculture or for their own subsistence. Another form of what people would call neo-colonialism where we compensate for our own emissions by pushing people off their lands far away. These technologies, and technology engineering particularly bring a lot of risks, such as technical and geopolitical risks. Thus, they may lead to tensions and even war. In one expert's opinion regarding land use is that the main beneficiaries have historically been large companies who have been using land grabbing techniques to make sure that they own land that was previously owned by community or was used by smaller communities. Especially around carbon capture and storage, and with the economical setup that we have now where carbon will be priced. The main beneficiaries will be a whole new economic sector that needs to be built around this.

Thus, CDR has the potential to lessen some existing injustices of climate change, respectively by keeping lower climate stabilization targets viable and by masking some climate impacts. At the same time, CDR raises acute justice-related concerns, both in terms of fair burden sharing among current and future people, and in terms of imposing unjust harms upon the vulnerable. It is thus very important that the implementation of CDR advance rather than undermine justice in pursuit of climate stabilization.

Paying attention to distributive justice in relation to CDR would mean addressing the following questions:

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No.101006249.

- Does the research and development of CDR cause or exacerbate inequalities?
- Does the implementation of CDR disproportionately harm the worst off in the country of implementation?
- Does the implementation of CDR disproportionately harm the worst off globally?
- Does the implementation of CDR impose unacceptable costs (economic, social, environmental) upon future generations?

For all of these questions, it is imperative to ask whether policy interventions or governance structures can be implemented to avoid distributive injustice in the research and development and implementation of climate engineering.

4.4.2. Procedural justice: How to include all affected parties in the decision making?

For both forms of CE, procedural justice challenges appear significant at the implementation stage, where the demands of procedural fairness and stakeholder involvement bear upon choices about where to implement a particular technique, under what socio-economic regimes or conditions, and at which time. Nonetheless, the procedural justice concerns raised by SRM seem far more severe than for CDR, which appears little different to ordinary technological or economic proposals. For many land-based forms of CDR, and perhaps especially BECCS, these procedural challenges include fair inclusion of affected parties in decisions about where to situate CDR facilities. Given the potential effects of unregulated biomass production upon food security, procedural justice may overlap with appropriate forms of governance and compensation to avoid ethically undesirable side-effects of implementation.

- For CDR, paying attention to procedural justice will require at a minimum stakeholder engagement processes to determine where CDR can be situated. More demanding to organize but more epistemically and politically robust responses would include deliberation processes such as citizen juries and mini-publics, to consider the extent of CDR reliance, the choice of technologies, and where to situate them, as in Irish Citizen's Assembly (Citizens' Assembly 2018).

4.4.3. Future responsibility: How to act responsibly toward future generations?

Hans Jonas has conceptualized the responsibility of the future generations in the 1970s. He argued that with the advent of globally destructive technologies, the responsibility is no longer to lead our "best life, but to ensure the existence of future life: "we need not go into the theory of the human good and the 'best life,' which would have to be derived from a conception of man's 'essence.' For the moment, all work on the 'true' man must stand back behind the bare saving of its precondition, namely, the existence of mankind in a sufficient natural environment" (Jonas 1976, p. 81).

The current climate situation requires action and sacrifice by the current generation for the preservation of the future of humanity. This means that the current general population should make material sacrifices if they were to contribute to the livable environment of future people. In other words, the existence of future humanity is the responsibility of the current

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No.101006249.

people. However, it is not easy to convince people to sacrifice for the future in which they will not be involved. Many conceptions of moral behaviour rest on a notion of good or moral life (Fischer 2014), even enjoyment or current happiness (Feldman 2004; Lampe 2017), however, a climate-conscious morality should exchange the idea of a good life for the idea of the future existence of life.

The issue of future responsibility emerged in the digital ethnography concerning a CDR business where the speaker described the business model of their innovation as follows: "you can pick your plan, ranging from as low as 7 dollars per month, going up to several thousand dollars per month, depending on your lifestyle and abilities, and with that […] subscription, you will be able to remove the unavoidable emissions of your life, be it your diet, be it your electric vehicle, or be it the flight you really cannot avoid." This subscription service model places climate action in the context of a neoliberal market framing where individuals with disposable income can offset their emissions voluntarily. This view presents the future in which the more disposable income one has, the more one can store away their own emission. What is not represented in this view of the future is in fact is any kind of responsibility claim for historical contributions to climate change.

The other very notable point about this business model sales pitch, which is the ability to remove 'unavoidable' emissions. This word is used several times e.g. diet as a source of unavoidable emissions. However, based on IPCC reports (e.g. IPCC 2022: 88-89) and a lot of associated climate science, diet is one of the things that individuals can exert the most control over, reducing dietary emissions by eating less meat and dairy. So, representing 'diet' as an unavoidable source of emissions detracts from the options people and perhaps business, have of taking responsibility for changing their own high-emission lifestyle. There is a danger here of this kind of technology acquiring the role of a modern Catholic indulgence payment, thus making it difficult to talk about climate responsibilities or justice.

Recent ethical analysis of CDR in "overshooting" scenarios (see section 4.1.1) has pointed to the speculative nature of large-scale CDR, and consequently of the risks of a policy failure (Fuss et al. 2014; Lenzi 2018, 2021; Shue 2017, 2018).

Shue (Shue 2017) has likened the policy gamble on CDR to a modified game of Russian roulette in which different agents stand to win and to lose. That is, the current generation wins by undertaking less mitigation in the near-term, thus saving funds that would have been required for mitigation, while future generations stand to lose should the bet on CDR prove to be a failure. What makes this gamble so morally problematic is that the gains to the current generation are relatively trivial from a moral point of view, while the losses to the loser would be catastrophic since the failure of a bet on CDR would leave future generations with no way to avoid runaway climate change (short of using SAI). As Shue (Shue 2017, p. 208) notes, even if such a gamble paid off, "[t]o take (or offer) this gamble is outrageous, bordering on psychotic".

Lenzi (Lenzi 2021) identifies four possible policy gambles, namely betting on a) ambitious mitigation only; b) ambitious mitigation plus CDR; c) deferred mitigation plus CDR; and d) business as usual (i.e., no effective emissions reduction). While Shue's concerns with CDR apply to option c), option b) presents another view of large-scale CDR implementation which is not obviously less risky than option a), given how little of the carbon budget remains

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No.101006249.

to limit warming to 1.5C. Nonetheless, there is agreement that the riskiest and most unjust policy gamble of all remains slow mitigation now, on the expectation of future large-scale CDR.

- The ethical literature clearly distinguishes between policies that feature slow near-term mitigation, overshooting, then a large policy bet on CDR, and rapid near-term mitigation with minimal overshooting and more modest CDR reliance. There is broad agreement that only the latter policies are consistent with intergenerational justice, and would constitute a forward-looking approach to CDR. Nonetheless, policymakers may consider the precautionary principle when determining their reliance on CDR.

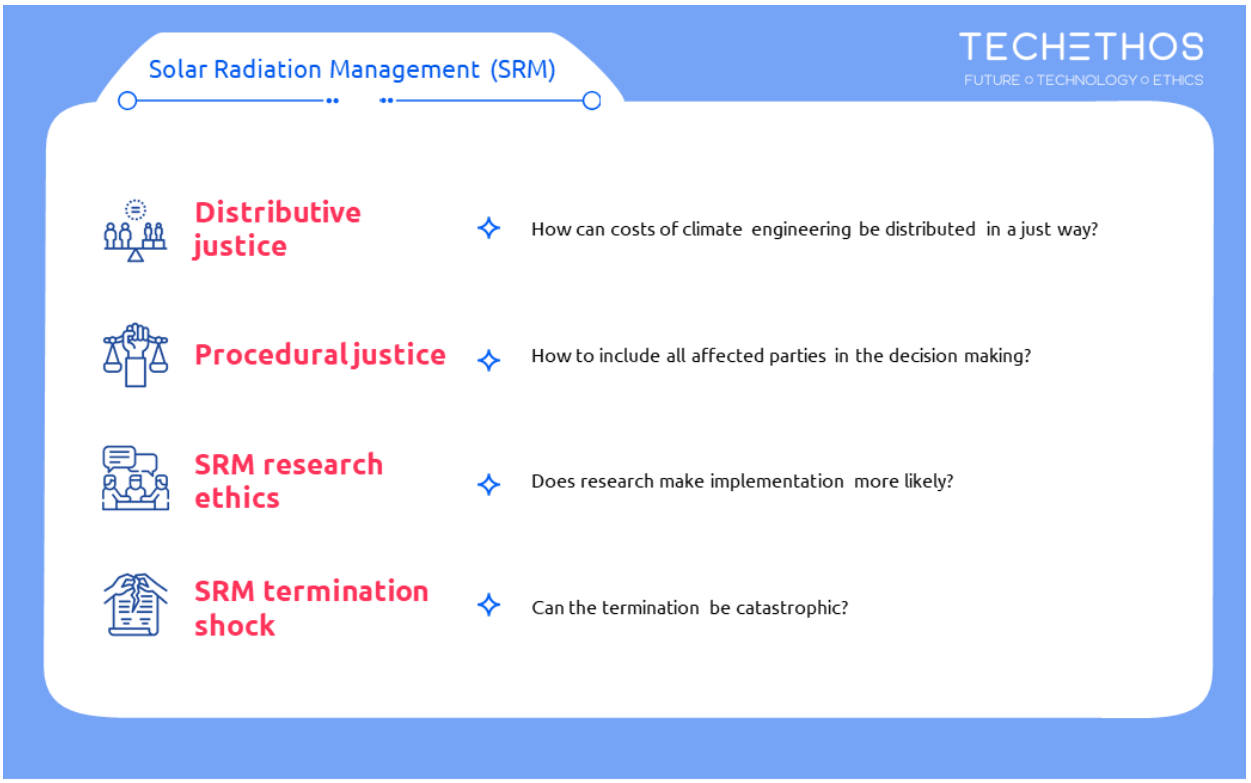
4.4.4. Side-effects: Are side-effects of climate engineering worse than their climate benefits?

As summarized in Section 4.1, large-scale CDR implementation may create morally serious side effects, particularly upon those communities least able to adapt and with the least historical contribution to climate change. Many of these concerns have been directed at BECCS, since large-scale upscaling would require vast amounts of land and water to be set aside for the growing of biomass, impacting food security, water, and biodiversity. In a widely cited piece, Anderson and Peters (Anderson and Peters 2016) claim that typical mitigation scenarios featuring BECCS would require a land area the size of India to be found to grow sufficient biomass. Further, they note that modelled scenarios typically explore the expansion of biomass in tropical climates, given assumptions about low production costs. When combined, these assumptions imply very serious moral risks that large-scale biomass production will seriously harm the poorest, who are already highly vulnerable to climate change.

As such, Shue (2017, 206) argues that large-scale BECCS implementation would have "completely unacceptable moral costs", such as causing famine and political unrest. Shue compares such an outcome to a "Sophie's choice" between producing sufficient biomass for CDR and feeding the global population. Lenzi (Lenzi 2021) points out that Shue's concern applies primarily to a future scenario in which near-term mitigation is significantly deferred. This is because the climate mitigation models which produce such enormous demand for BECCS later in the century feature the largest periods of emissions "overshooting". However, if near-term mitigation is rapid, the overall demand for BECCS would fall. Nonetheless, there may still be room for concern with unjust side-effects of biomass reliance, since Lenzi (2018) also notes how even mitigation modelling featuring limited or no CDR typically feature equally great demands for biomass to render existing fossil infrastructure carbon neutral. As a result, the risks of unjust side-effects from heavy biomass reliance remains.

- Governance of the side-effects of CDR, particularly the sourcing of biomass, has become a key priority following the publication of the IPCC AR6, and in light of country net zero pledges. The European Union in particular faces the question of ensuring that the production of biomass in other nations conforms to high environmental and social standards, given risks of land clearance exacerbating land grabbing and dispossession where property rights are poorly enforced, and risks of further biodiversity losses.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No.101006249.



This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No.101006249.

4.5. Values and principles in SRM

4.5.1. Distributive justice: How can risks of climate engineering be distributed in a just way?

The potential for SRM, and especially SAI, to raise distributive justice concerns is widely acknowledged. Indeed, Preston's (Preston 2013) review of distributive justice concerns focuses almost entirely upon SAI. Some advocates point to SAI as a means to avoid the greater injustice of runaway climate change, thus framing it as a "lesser evil" (Gardiner 2010). As noted above, the AR6 also notes the potential of SAI to reduce some warming, which would buy additional time for mitigation to lower the atmospheric concentration of CO2 However, many climate ethicists writing about CE are concerned in particular about the potential for it to impose grave injustice, both in terms of harming the vulnerable and unfair burden sharing. Gardiner (Gardiner 2010) has forcefully objected to SAI being framed as a "lesser evil", arguing that it is far from clear that this constitutes the best available policy to avoid climate catastrophe, and that any plan to utilise SAI would be predicated upon the moral failure of the current generation to mitigate its emissions. Gardiner also draws attention to the potential for future generations to be faced with a "Sophie's choice" between using SAI despite its morally unacceptable side-effects, and suffering from morally unacceptable climate impacts. While Gardiner does not actually rule out any implementation of SAI under any circumstances, the worry is that pursuing SAI will delay mitigation (i.e., the moral hazard), and future generations will be faced with such an ethically intolerable trade-off.

Justice concerns with SAI often highlight its potential for unequally distributed negative impacts. As Preston (Preston 2013) notes, a world artificially cooled by SAI raises questions about whose interests ought to be protected, and it is far from clear that the interests of the most vulnerable would be prioritized if SAI were ever implemented, or that fair compensation would be given to such additional harms. SAI is expected to affect precipitation patterns in some regions more severely, particularly the Indian monsoon. Given the importance of this for agricultural production in these regions, and their relative poverty, SAI's potential interference in rainfall could impose severe injustice upon people who have the least ability to adapt. This would of course compound the existing injustice of climate change itself, which generally harms the most vulnerable disproportionately due to their diminished capacities to adapt. A related argument comes from Whyte (Whyte 2012, 2017), who argues that from the perspective of marginalized groups such as indigenous peoples, the injustice of climate change and then of CE using SAI layers additional injustice upon historical wrongs such as colonialism. In terms of burden-sharing, SAI raises concerns with current generations shirking their moral responsibility to mitigate and relying on it to artificially cool the global climate (Gardiner 2010).

For SRM, blocking some incoming solar radiation would in reduce the intensity of climate impacts, which is sometimes framed as a way of "buying time" for mitigation. Indeed, Horton and Keith (J. Horton and Keith 2016) argue that global distributive justice requires the undertaking of research into SRM, because failing to conduct such research effectively condemns the global poor to suffer from the worst climate impacts, despite being marginal beneficiaries from past emissions. However, Hourdequin (Hourdequin 2018) claims that this takes too narrow a view of justice, ignoring the distribution of epistemic power and power to make decisions about climate policy, and hence questions of procedural and recognition justice bearing upon how research on SRM is conducted.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No.101006249.

In view of power structures discussed in section 1.3.4, the relations of power and historic inequalities between the Global North and the Global South have been key when it comes to SRM. Smith makes this point strongly in "*Who* May Geoengineer: Global Domination, Revolution, and Solar Management" (2021).[11] As he shows, here too, one needs to ask the "who question". What is just to do, cannot be determined without accounting for the systems of inequalities that have led to the way things are today, i.e., the fact that "the people who will suffer the worst impacts of climate change will be those who are least responsible for it occurring" (2021, p. 138). As Smith argues: "Vulnerable, low-emitting nations have a more easily justified permission—based on the unjust relations that currently obtain between those nations and powerful, high-emitting actors—to deploy dangerous geoengineering strategies in response to climate change." (2021, p. 139) In other words, the "who question" at the level of impacts has severe consequences at the level of determining what is just to do and for whom.

One way in which distributive justice can be approached is to frame SRM in terms of risks, both of side-effects from its implementation, and from the consequences of non-deployment. Viewed in this way, a decision-maker could contrast the effects of climate warming without SRM, which would occur even under decarbonisation pathways, with the side-effects of SRM deployment, for instance upon weather patterns. However, this risk-risk framing is contentious, and many experts would reject the assumption underlying it, namely that SRM would be used in conjunction with stringent climate mitigation (https://www.solargeoeng.org/). The moral hazard concern is again relevant, since framing the choice of whether to deploy SRM in terms of risks of deployment against non-deployment assumes away the worrying possibility that policymakers would view SRM as a technological solution which allows them to slow or even avoid entirely the costly and difficult task of decarbonisation. It is unclear that policymakers would necessarily think of SRM as a complement to mitigation, rather than as a substitute for it. Thus, the risk/risk framing for researching and potentially deploying SRM tries to silence moral hazard concern that SRM would derail climate mitigation.

Paying attention to distributive justice in relation to SRM would mean addressing the following questions:

- Does the research and development of SRM cause or exacerbate inequalities?
- Does the implementation of SRM disproportionately harm the worst off in the country of implementation?
- Does the implementation of SRM disproportionately harm the worst off globally?
- Does the implementation of SRM impose unacceptable costs (economic, social, environmental) upon future generations?

For all of these questions, it is imperative to ask whether policy interventions or governance structures can be implemented to avoid distributive injustice in the research and development and implementation of climate engineering.

[11] Italics by TechEthos.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No.101006249.

4.5.2. Procedural justice: How to include all affected parties in decision making?

Preston (Preston 2013) notes that procedural justice is one of the biggest ethical challenges facing CE. Yet as recent literature has argued, this conclusion appears true primarily for SRM. Indeed, Preston's analysis was mostly concerned with SRM, which had been the focus of most ethical literature until a few years ago. As a result, he notes that procedural justice is already important when considering whether to research SRM, given concerns that research may make implementation more likely (see Section 4.5.3 below), and given the very large set of potential stakeholders likely to be affected by any decision to implement SRM. Procedural justice raises particular challenges for SRM even at the research and development stage.

For SRM, and especially SAI, procedural justice is a daunting problem. Indeed, Preston (2013) concludes that procedural justice is unlikely to be satisfied given that any implementation of this technique would immediately affect every person living at the time, and all future generations until SAI ceased. Thus, as Corner and Pidgeon write, "the prospect of controlling the global thermostat is something that all citizens could reasonably claim to have a legitimate stake in" (Corner and Pidgeon 2010).

Procedural justice is especially worrisome given the general difficulties that binding global climate enforcement treaties have faced (e.g., the Kyoto Protocol), and the fact that SAI could be deployed by a single state or even by a wealthy private individual, without needing the cooperation of other parties. Such difficulties have encouraged recent contributors to turn to existing legal frameworks that could be brought to bear upon SAI, such as the ENMOD Convention, which is a Cold War arms control treaty that applies to technologies that modify the environment (McGee et al. 2021).

Nonetheless, it remains unclear how procedural justice could be satisfied in any decision to implement SAI. While it has been argued that SAI is not necessarily incompatible with democracy or with robust democratic governance, there is no reason to expect the governance of SAI to actually be democratic (J. B. Horton et al. 2018). Another scenario would be for SAI to be implemented without regard to procedural justice, for instance in response to an extreme climate event. Such a possibility is presented in Kim Stanley Robinson's recent work of climate fiction, *The Ministry for the Future*, in which one country unilaterally deploys SAI following a devastating heatwave event, ignoring the protests of other nations.

- In terms of SRM, and especially SAI, it is not clear what would constitute an adequate consideration of procedural justice since any decision to implement SAI would immediately have global effects upon the climate system. As noted, this implies that any participatory or deliberative process would require representing globally affected parties (and potentially representatives of future generations).

4.5.3. SRM research ethics: Does research make implementation more likely?

SRM, especially SAI, has raised long-standing concerns that research and development programmes would make future implementation more likely (also known as the "slippery slope" argument). For SAI, a full verification of its effectiveness and side-effects is impossible

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No.101006249.

prior to full deployment, which would of course cross over from research to deployment. The concern that research may make deployment of SRM more likely was noted early on by Jamieson (Jamieson 1996), and several contributors claim that research into SRM makes it more likely to be deployed, particularly given the tendency of the current generation to "pass the buck" of mitigation responsibility onto future generations (Gardiner 2010, 2011). Callies (Callies 2019) casts doubt upon both the empirical premise that CE research makes deployment inevitable, and the moral assumption that deployment should be seen as impermissible. Instead, Callies argues against a moratorium on research, and instead for improved governance. Morrow (Morrow 2020) proposes a "mission-driven" approach to governing SRM along the lines of the Apollo program, to ensure just, feasible and socially responsible research and development.

There is widespread agreement that improved governance of SRM research is essential. A leading attempt to provide governance on research are the "Oxford Principles", which were submitted to the UK House of Commons in 2009 The Oxford Principles aim to steer the development of CE (both CDR and SRM), from initial research through to deployment. Further, the Oxford Principles require that any decision to deploy CE must first ensure that robust governance structures have been adopted.[12] The five principles state: 1) CE is to be regulated as a public good; 2) public participation in CE decision-making; 3) CE research disclosure and open publication of results; 4) Independent assessment of impacts; 5) governance in place prior to deployment.[13]

There have been several notable attempts to conduct field research upon SAI in conformity with the Oxford Principles or similar proposals. These attempts have confined themselves to "process" experiments, testing the efficacy of some mechanism that SAI would activate under limited conditions, and without wider effects on the global climate. To date however, the two most prominent research attempts have resulted in failure. These are the S.P.I.C.E. project, which aimed to test a hose and balloon system which would be part of the mechanism to deposit SAI particles into the stratosphere; and the SCoPEx project, run by the Keutsch group at Harvard University, which uses a balloon to create a closed sample of stratospheric air and observes the interaction of SAI particles with this. S.P.I.C.E. was cancelled in 2012, seemingly due to complaints from environmental groups and project participants concerned at the absence of governance structures conforming to the Oxford Principles (Preston 2013). A recent attempt to test SCoPEx" balloon in north Sweden was cancelled in 2021, as US and European scientists raised concerns about potentially damaging effects of the particles upon the ozone layer and ecosystems.

A more demanding set of governance principles has been proposed by Gardiner and Fragnière (Gardiner and Fragnière 2018). Their 10 "tollgate principles" build upon Jamieson's (Jamieson 1996) early proposal for governance of CE research, and upon their critique that the Oxford Principles lack substantive ethical content concerning justice, respect and legitimacy. At present, however, there remains no binding regulation governing SRM research aside from existing national research governance. Instead, as shown by the successful opposition to

[12] See http://www.geoengineering.ox.ac.uk/www.geoengineering.ox.ac.uk/oxford-principles/principles/

[13] Despite the aim of the Oxford Principles to govern research and development of CE generally, It is noteworthy that these principles often seem only to apply to SRM. For instance, principle 5) stating that governance must precede deployment appears inapplicable since some forms of CDR have already been implemented. Further, the AR6 reclassification of CDR as mitigation suggests that governance principles prior to implementation are not required since such a requirement is not generally observed for mitigation.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No.101006249.

SCoPEx by environmental organisations and the Saami indigenous people of northern Scandinavia, the main obstacle to in situ field testing seems to be public opposition.

- An adequate consideration of research ethics for climate engineering would involve a further development of such proposals into enforceable research guidelines. Doing so would involve inter alia asking whether the research and development of SRM respected justice concerns, respect, and political legitimacy, and showing how these ethical norms were satisfied in particular instances.

4.5.4. SRM termination shock: Can the termination be catastrophic?

The interaction of SAI with radiative forcing is such that while any application would mask some warming, any cessation of SAI would very rapidly cause global temperatures to rebound again, with impacts that would have developed over decades hitting all at once. As Preston (Preston 2013) notes, such a rapid temperature rise would cause clear dangers to humanity and to non-human species alike, many of which would not be able to migrate or adapt in time. Yet as Preston notes, SAI might be interrupted either following political failure (e.g., conflict), or for other environmental or social reasons. As Preston notes, "one might hope that this "termination problem" could be avoided through careful research… and through stable political institutions being established prior to implementation. However, the complexity of the climate system, the challenge raised by the limited ability to field test, and the perfect moral storm of climate change mean that a number of scientific and political uncertainties about long-term deployment are likely to remain" (Preston 2013, p. 13).[14]

There is continued debate about the ethical and geopolitical implications of the termination shock for SRM. Parker and Irvine (Parker and Irvine 2018) express optimism that governance could reduce risks by dispersing SAI infrastructure widely enough that damage or malfunction at any site would not mean cessation; by developing "backup" infrastructure so that repairs to infrastructure would not depend on too small a set of states; and by ensuring infrastructure is adequately protected by military and security forces. McKinnon (McKinnon 2020) rejects these assumptions as naïve about the extent of political cooperation needed to develop such "backup" infrastructure, and the motives of geopolitical actors. There is a related worry with the "securitization" of SRM, since one nation or even one extremely wealthy actor could implement it, and in the process could politically dominate all other nations (McKinnon 2020; Morrow 2020). Given this, the geopolitical consequences of one actor holding the "global thermostat" may fundamentally alter global security in ways similar to the possession of a superweapon.

- At the level of governance, avoiding a termination shock for SRM would involve asking whether appropriate governance structures have been developed to address potential supply shortages or catastrophic failures to continue the implementation of SAI, for instance due to political unrest. As noted above, the political feasibility of such an enterprise is hotly debated.

[14] The 'perfect moral storm' refers to a book by Gardiner (Gardiner 2011), which argues that a variety of factors (political, geophysical, institutional, ethical) combine to render climate change a uniquely difficult problem. Gardiner describes climate change as an iterated prisoner's dilemma in which each generation has the incentive to 'pass the buck' of mitigation costs onto future generations, with collectively ruinous consequences. The 'ethical' storm refers to difficulties that standard ethical theories have in capturing the complexity of climate change.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No.101006249.

- If SRM were framed as a security measure in future (for instance due to extreme climate impacts), military partners may work to ensure the avoidance of termination shock through coordination.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No.101006249.

References

Abd-alrazaq, A. A., Alajlani, M., Alalwan, A. A., Bewick, B. M., Gardner, P., & Househ, M. (2019). An overview of the features of chatbots in mental health: A scoping review. *International Journal of Medical Informatics*, *132*, 103978. https://doi.org/10.1016/j.ijmedinf.2019.103978

Abich, J., Parker, J., Murphy, J. S., & Eudy, M. (2021). A review of the evidence for training effectiveness with virtual reality technology. *Virtual Reality*, *25*(4), 919–933.

Abid, A., Farooqi, M., & Zou, J. (2021). Persistent Anti-Muslim Bias in Large Language Models. *arXiv:2101.05783 [cs]*. http://arxiv.org/abs/2101.05783. Accessed 5 April 2022

Abramson, D. I., & Johnson, J. J. (2020, December 1). Creating a conversational chat bot of a specific person. https://patents.google.com/patent/US10853717B2/en?oq=us10853717b2. Accessed 31 March 2022

Accolla, E. A., & Pollo, C. (2019). Mood Effects After Deep Brain Stimulation for Parkinson's Disease: An Update. In *Frontiers in neurology* (Vol. 10, p. 617). https://doi.org/10.3389/fneur.2019.00617

Adadi, A., & Berrada, M. (2018). Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI). *IEEE Access*, *6*, 52138–52160. Presented at the IEEE Access. https://doi.org/10.1109/ACCESS.2018.2870052

Adomaitis, L. (2019). Cause and Effect in Leibniz's Brevis demonstratio. *HOPOS: The Journal of the International Society for the History of Philosophy of Science*, *9*(1), 120–134.

Aggarwal, S., & Chugh, N. (2020). Ethical implications of closed loop brain device: 10-year review [Review of ethical implications of closed loop brain device: 10-year review. *Minds Mach*, *30*, 145–170. https://doi.org/10.1007/s11023-020-

Albuz, E., Vernon, C., Liang, S., & Guo, P. (2022, January 4). Avatar fidelity and personalization. https://patents.google.com/patent/US11217036B1/. Accessed 16 February 2022

American Bar Association. (2021). *Model Rules of Professional Conduct*. American Bar Association.

Anderson, K., & Peters, G. (2016). The Trouble with Negative Emissions. *Science*, *354*(6309), 182–83. https://doi.org/10.1126/science.aah4567.

Aquinas, S. T. (1955). *Summa Theologica (Complete)*. Library of Alexandria.

Arendt, H. (2013). *The human condition*. University of Chicago press.

Aronsson, A. S. (2020). Social Robots in Elder Care The Turn Toward Emotional Machines in Contemporary Japan. *Japanese review of cultural anthropology*, *21*(1), 421–455.

Arora, A., Bansal, S., Kandpal, C., Aswani, R., & Dwivedi, Y. (2019). Measuring social media influencer index- insights from facebook, Twitter and Instagram. *Journal of Retailing and Consumer Services*, *49*, 86–101. https://doi.org/10.1016/j.jretconser.2019.03.012

Arsenyan, J., & Mirowska, A. (2021). Almost human? A comparative case study on the social media presence of virtual influencers. *International Journal of Human-Computer Studies*, *155*, 102694.

Augustine, S. (2012). *The Confessions: With an Introduction and Contemporary Criticism*. Ignatius Press.

Azuma, R. T. (1997). A survey of augmented reality. *Presence: teleoperators & virtual environments*, *6*(4), 355–385.

Bachelard, G. (1964). *The Psychoanalysis of Fire*. (A. C. M. Ross, Trans.). Boston: Beacon Press.

Bagheri, R. (2016). Virtual Reality: The Real Life Consequences. *UC Davis Business Law Journal*, *17*, 101–120.

Ball, K. (2021). *Electronic Monitoring and Surveillance in the Workplace*. Joint Research Centre (Seville site).

Banta, N. M. (2015). Death and Privacy in the Digital Age. *North Carolina Law Review*, *94*, 927.

Barendt, E. (2016). *Anonymous Speech: Literature, Law and Politics*. Bloomsbury Publishing.

Barnouw, J. (2008). Reason as Reckoning: Hobbes's Natural Law as Right Reason. *Hobbes Studies*, *21*(1), 38–62.

Barrett, S. (2008). The Incredible Economics of Geoengineering. *Environmental & Resource Economics*, *39*, 45–54.

Basu, T. (2021). The metaverse has a groping problem already. https://www.technologyreview.com/2021/12/16/1042516/the-metaverse-has-a-groping-problem/. Accessed 22 February 2022

Baudrillard, J. (1994). *Simulacra and simulation*. University of Michigan press.

BBC News. (2021, December 10). Meta releases social VR space Horizon Worlds. *BBC News*. https://www.bbc.com/news/technology-59609996. Accessed 8 March 2022

Beauchamp, T. L., & Childress, J. F. (2013). *Principles of biomedical ethics* (7th ed.). Oxford: Oxford University Press.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No.101006249.

Bellamy, R. K., Dey, K., Hind, M., Hoffman, S. C., Houde, S., Kannan, K., et al. (2018). AI Fairness 360: An extensible toolkit for detecting, understanding, and mitigating unwanted algorithmic bias. *arXiv preprint arXiv:1810.01943*.

Bellaver, B., Ferrari-Souza, J. P., Ros, L. U. da, Carter, S. F., Rodriguez-Vieitez, E., Nordberg, A., et al. (2021). Astrocyte Biomarkers in Alzheimer Disease: A Systematic Review and Meta-analysis. *Neurology*, *96*(24), e2944–e2955. https://doi.org/10.1212/WNL.0000000000012109

Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021a). On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? 🦜 In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (pp. 610–623). New York, NY, USA: Association for Computing Machinery. https://doi.org/10.1145/3442188.3445922

Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021b). On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? . In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (pp. 610–623). New York, NY, USA: Association for Computing Machinery. https://doi.org/10.1145/3442188.3445922

Bentham, J. (1879). *The Principles of Morals and Legislation*. Clarendon Press.

Berg, J. W., Appelbaum, P. S., C.W., L., & L, P. (2001). *Informed Consent: Legal Theory and Clinical Practice* (2nd ed.). Fair Lawn, NJ, USA: Oxford University Press.

Berlin, I. (2013). *The Crooked Timber of Humanity: Chapters in the History of Ideas - Second Edition*. Princeton University Press.

Betz, G., & Cacean, S. (2012). *Ethical Aspects of Climate Engineering*. Karlsruhe, Germany: Karlsruhe Institut für Technologie.

Bibault, J.-E., Chaix, B., Nectoux, P., Pienkowski, A., Guillemasé, A., & Brouard, B. (2019). Healthcare ex Machina: Are conversational agents ready for prime time in oncology? *Clinical and translational radiation oncology*, *16*, 55–59.

Bietti, E. (2021). From ethics washing to ethics bashing: a view on tech ethics from within moral philosophy. Presented at the FAT* '20: Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, SSRN. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3914119

Blodgett, S. L., Barocas, S., III, H. D., & Wallach, H. M. (2020). Language (Technology) is Power: A Critical Survey of "Bias" in NLP. *CoRR*, *abs/2005.14050*. https://arxiv.org/abs/2005.14050

Blodgett, S. L., Green, L., & O'Connor, B. (2016a). Demographic dialectal variation in social media: A case study of African-American English. *arXiv preprint arXiv:1608.08868*.

Blodgett, S. L., Green, L., & O'Connor, B. (2016b). Demographic Dialectal Variation in Social Media: A Case Study of African-American English. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing* (pp. 1119–1130). Presented at the EMNLP 2016, Austin, Texas: Association for Computational Linguistics. https://doi.org/10.18653/v1/D16-1120

Bollmer, G., & Suddarth, A. (2022). Embodied parallelism and immersion in virtual reality gaming. *Convergence*, 13548565211070692. https://doi.org/10.1177/13548565211070691

Borgen, K. B., Ropp, T. D., & Weldon, W. T. (2021). Assessment of Augmented Reality Technology's Impact on Speed of Learning and Task Performance in Aeronautical Engineering Technology Education. *The International Journal of Aerospace Psychology*, *31*(3), 219–229.

Bormann, D., & Greitemeyer, T. (2015). Immersed in virtual worlds and minds: effects of in-game storytelling on immersion, need satisfaction, and affective theory of mind. *Social Psychological and Personality Science*, *6*(6), 646–652.

Bose, A. J., & Aarabi, P. (2019). Virtual Fakes: DeepFakes for Virtual Reality. In *2019 IEEE 21st International Workshop on Multimedia Signal Processing (MMSP)* (pp. 1–1). IEEE.

Bostrom, N. (2005a). In Defense of Posthuman Dignity. *Bioethics*, *19*(3), 202–214. https://doi.org/10.1111/j.1467-8519.2005.00437.x

Bostrom, N. (2005b). A history of transhumanist thought. *Journal of evolution and technology*, *14*(1).

Bostrom, N., & Roache, R. (2008). Ethical issues in human enhancement. *New waves in applied ethics*, 120–152.

Boto, E., Holmes, N., Leggett, J., Roberts, G., Shah, V., Meyer, S. S., et al. (2018). Moving magnetoencephalography towards real-world applications with a wearable system. *Nature*, *555*(7698), 657–661. https://doi.org/10.1038/nature26147

Bradski, G. R., Miller, S. A., & Abovitz, R. (2016, January 28). Methods and systems for creating virtual and augmented reality. https://patents.google.com/patent/US20160026253A1/en. Accessed 16 February 2022

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No.101006249.

Brass, M., Furstenberg, A., & Mele, A. R. (2019). Why neuroscience does not disprove free will. *Neuroscience & Biobehavioral Reviews*, *102*, 251–263. https://doi.org/10.1016/j.neubiorev.2019.04.024
Brey, P. (1999). The ethics of representation and action in virtual reality. *Ethics and Information Technology*, *1*(1), 5–14. https://doi.org/10.1023/A:1010069907461
Brey, P. (2014). Virtual Reality and Computer Simulation. In R. L. Sandler (Ed.), *Ethics and Emerging Technologies* (pp. 315–332). London: Palgrave Macmillan UK. https://doi.org/10.1057/9781137349088_21
Brey, P., & Dainow, B. (2021). Ethics by design and ethics of use in AI and robotics. *The SIENNA project-Stakeholder-informed ethics for new technologies with high socio-economic and human rights impact. Accessed April*, *26*, 2021.
Brian, D., Pierre, F., & Melik, R. van. (2021). *Volume 2: Housing and Home*. Policy Press.
Bringsjord, S., Bello, P., & Ferrucci, D. (2001). Creativity, the Turing Test, and the (Better) Lovelace Test. *Minds and Machines*, *11*(1), 3–27. https://doi.org/10.1023/A:1011206622741
Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., et al. (2020). Language Models are Few-Shot Learners. *arXiv:2005.14165 [cs]*. http://arxiv.org/abs/2005.14165. Accessed 5 April 2022
Brunton, F., & Nissenbaum, H. (2015). *Obfuscation: A user's guide for privacy and protest*. Mit Press.
Bryant, D., & Howard, A. (2019). A Comparative Analysis of Emotion-Detecting AI Systems with Respect to Algorithm Performance and Dataset Diversity. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society* (pp. 377–382). New York, NY, USA: Association for Computing Machinery. https://doi.org/10.1145/3306618.3314284
Buchanan, A. (1978). Medical paternalism. *Philosophy & Public Affairs*, 370–390.
Buchinger, E. (2007). Applying Luhmann to conceptualize public governance of autopoietic organizations. *Cybernetics and Human Knowing*, *14*, 173–187.
Buchinger, E. (2010). Governance as a societal distributed process: A multi-agent, multi mechanism, multi-level approach (M3. In R. Trappl (Ed.), *Cybernetics and Systems. Proceedings of the 20th European Meeting on Cybernetics and Systems Research* (pp. 252–257). Vienna.
Buchinger, E. (2012). Luhmann and the constructivist heritage: A critical reflection. *Constructivist Foundations*, *8*, 19–28.
Buckner, C., & Garson, J. (2019). Connectionism. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Fall 2019.). Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/archives/fall2019/entries/connectionism/. Accessed 1 April 2022
Buckner, R. L., & DiNicola, L. M. (2019). The brain's default network: updated anatomy, physiology and evolving insights. *Nature Reviews Neuroscience*, *20*(10), 593–608.
Burton, E., Clayville, K., Goldsmith, J., & Mattei, N. (2019). The Heart of the Matter: Patient Autonomy as a Model for the Wellbeing of Technology Users. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society (AIES '19* (pp. 13–19). New York, NY, USA: Association for Computing Machinery. https://doi.org/10.1145/3306618.3314254
Busso, D. S., & Pollack, C. (2015). No brain left behind: consequences of neuroscience discourse for education. *Learning, Media and Technology*, *40*(2), 168–186. https://doi.org/10.1080/17439884.2014.908908
Butorac, I., Lentzos, F., & Aicardi, C. (2021). Gray Matters: Exploring Technologists' Perceptions of Dual-Use Potentiality in Emerging Neurotechnology Applications. *Health Security*, *19*(4), 424–430. https://doi.org/10.1089/hs.2020.0147
Button, G., & Sharrock, W. (1998). The Organizational Accountability of Technological Work. *Social Studies of Science*, *28*(1), 73–102. https://doi.org/10.1177/030631298028001003
Bye, K., Hosfelt, D., Chase, S., Miesnieks, M., & Beck, T. (2019). The ethical and privacy implications of mixed reality. In *ACM SIGGRAPH 2019 Panels* (pp. 1–2). New York, NY, USA: Association for Computing Machinery. https://doi.org/10.1145/3306212.3328138
Callies, D. (2019). The Slippery Slope Argument against Geoengineering Research. *Journal of Applied Philosophy*, *36*(4), 675–87.
Carlini, N., Tramèr, F., Wallace, E., Jagielski, M., Herbert-Voss, A., Lee, K., et al. (2021). Extracting Training Data from Large Language Models (pp. 2633–2650). Presented at the 30th USENIX Security Symposium (USENIX Security 21). https://www.usenix.org/conference/usenixsecurity21/presentation/carlini-extracting. Accessed 7 April 2022
Carter, A., Capps, B., & Hall, W. D. (2012). Chapter 5 - Emerging Neurobiological Treatments of Addiction: Ethical and Public Policy Considerations. In A. Carter, W. Hall, & J. Illes (Eds.),

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No.101006249.

Addiction Neuroethics (pp. 95–113). San Diego: Academic Press. https://doi.org/10.1016/B978-0-12-385973-0.00005-3
CERNA. (2017). *Éthique de la recherche en apprentissage machine*. CERNA; ALLISTENE.
Chalmers, D. (2022). *REALITY +: a philosophical journey through virtual worlds*. S.l.: ALLEN LANE.
Charland, P., & Dion, J.-S. (2018). L'utilisation de données psychophysiologiques pour mieux comprendre l'apprentissage en temps réel : le fragile équilibre entre la validité des données et l'authenticité des contextes de collecte de données. *Neuroeducation*, *5*(1), 1–3. https://doi.org/10.24046/neuroed.20180501.1
Chatellier, R. (2022). Métavers : ce jeu dont qui sera le héros ? | CNIL. https://www.cnil.fr/fr/metavers-ce-jeu-dont-qui-sera-le-heros. Accessed 16 March 2022
Chaudhary, U., Birbaumer, N., & Ramos-Murguialday, A. (2016). Brain-computer interfaces for communication and rehabilitation. *Nature Reviews Neurology*, *12*(9), 513–525.
Chen, X., Salem, A., Backes, M., Ma, S., & Zhang, Y. (2021). BadNL: Backdoor Attacks Against NLP Models. Presented at the ICML 2021 Workshop on Adversarial Machine Learning. https://openreview.net/forum?id=v6UimxiiR78. Accessed 7 April 2022
Chi, H.-L., Kang, S.-C., & Wang, X. (2013). Research trends and opportunities of augmented reality applications in architecture, engineering, and construction. *Automation in construction*, *33*, 116–122.
Chin, H., & Yi, M. Y. (2019). Should an Agent Be Ignoring It? A Study of Verbal Abuse Types and Conversational Agents' Response Styles. In *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems* (pp. 1–6). New York, NY, USA: Association for Computing Machinery. https://doi.org/10.1145/3290607.3312826
Christin, A. (2020). Algorithmic ethnography, during and after COVID-19'. *Communication and the Public*, *5*(3–4), 108–111. https://doi.org/10.1177/2057047320959850.
Christman, J. (2020). Autonomy in Moral and Political Philosophy". In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy*. https://plato.stanford.edu/cgi-bin/encyclopedia/archinfo.cgi?entry=autonomy-moral
Coffey, M. (1993). The Genetic Defense: Excuse or Explanation? *William & Mary Law Review*, *35*(1), 353.
CoinYuppie. (2022, January 20). After 1427 pages of patent documents, we discovered Meta's Metaverse secrets. https://coinyuppie.com/after-1427-pages-of-patent-documents-we-discovered-metas-metaverse-secrets/. Accessed 16 February 2022
Colleoni, E., Rozza, A., & Arvidsson, A. (2014). Echo chamber or public sphere? Predicting political orientation and measuring political homophily in Twitter using big data. *Journal of communication*, *64*(2), 317–332.
Condon, A., & Willatt, G. (2018). ECG biometrics: the heart of data-driven disruption? *Biometric Technology Today*, *2018*(1), 7–9. https://doi.org/10.1016/S0969-4765(18)30011-0
Cook, S. (2020). Posthumous dignity and the importance in returning remains of the deceased. In *Forensic Science and Humanitarian Action* (pp. 67–78). John Wiley & Sons, Ltd. https://doi.org/10.1002/9781119482062.ch5
Cooper, A., & Ireland, D. (2018). Designing a chat-bot for non-verbal children on the autism spectrum. *Stud Health Technol Inform*, *252*, 63–68.
Corner, A., & Pidgeon, N. (2010). Geoengineering the climate: the social and ethical implications. *Environment: Science and Policy for Sustainable Development*, *52*(1), 24–37.
Crawford, A., & Smith, T. (2022, February 23). Metaverse app allows kids into virtual strip clubs. *BBC News*. https://www.bbc.com/news/technology-60415317. Accessed 9 March 2022
Crawford, K. (2021). *Atlas of AI*. New Haven & London: Yale University Press.
Crenshaw, K. (2022). *On Intersectionality: Essential Writings*. The New Press.
Crisp, R. (2006). Hedonism reconsidered. *Philosophy and Phenomenological Research*, *73*(3), 619–645.
Crutzen, P. J. (2006). Albedo Enhancement by Stratospheric Sulfur Injections: A Contribution to Resolve a Policy Dilemma? *Climatic Change*, *77*, 211–19.
Cruz-Neira, C., Sandin, D. J., & DeFanti, T. A. (1993). Surround-screen projection-based virtual reality: the design and implementation of the CAVE. In *Proceedings of the 20th annual conference on Computer graphics and interactive techniques* (pp. 135–142).
Czech, A. (2021). Brain-Computer Interface Use to Control Military Weapons and Tools. In S. Paszkiel (Ed.), *Control, Computer Engineering and Neuroscience. ICBCI 2021. Advances in Intelligent Systems and Computing* (Vol. 1362). Cham: Springer. https://doi.org/10.1007/978-3-030-72254-8_20
Daniels, N. (2000). Normal Functioning and the Treatment-Enhancement Distinction. *Cambridge Quarterly of Healthcare Ethics*, *9*(3), 309–322. https://doi.org/10.1017/S0963180100903037

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No.101006249.

Darwiche, F. (2019). Time'. In H. Paul (Ed.), *Critical Terms in Futures Studies* (pp. 307–312). Cham: Springer International Publishing. https://doi.org/10.1007/978-3-030-28987-4_47.
Dastin, J. (2018). Amazon scraps secret AI recruiting tool that showed bias against women. In *Ethics of Data and Analytics* (pp. 296–299). Auerbach Publications.
Davies, S. R., & Macnaghten, P. (2010). Narratives of mastery and resistance: lay ethics of nanotechnology. *NanoEthics*, *4*(2), 141–151.
Davis, E. (2016). AI amusements: the tragic tale of Tay the chatbot. *AI Matters*, *2*(4), 20–24. https://doi.org/10.1145/3008665.3008674
De Brigard, F. (2010). If you like it, does it matter if it's real? *Philosophical Psychology*, *23*(1), 43–57.
De Guzman, J. A., Thilakarathna, K., & Seneviratne, A. (2019). Security and privacy approaches in mixed reality: A literature survey. *ACM Computing Surveys (CSUR)*, *52*(6), 1–37.
De Keyser, A., Bart, Y., Gu, X., Liu, S. Q., Robinson, S. G., & Kannan, P. K. (2021). Opportunities and challenges of using biometrics for business: Developing a research agenda. *Journal of Business Research*, *136*, 52–62. https://doi.org/10.1016/j.jbusres.2021.07.028
Delacroix, S., & Wagner, B. (2021a). Constructing a mutually supportive interface between ethics and regulation. *Computer Law & Security Review*, *40*, 105520. https://doi.org/10.1016/j.clsr.2020.105520
Delacroix, S., & Wagner, B. (2021b). Constructing a mutually supportive interface between ethics and regulation. *Computer Law & Security Review*, *40*, 105520. https://doi.org/10.1016/j.clsr.2020.105520
Deng, C., Yuan, H., & Dai, J. (2018). Behavioral Manipulation by Optogenetics in the Nonhuman Primate. *The Neuroscientist*, *24*(5), 526–539. https://doi.org/10.1177/1073858417728459
Dennett, D. (1971). Intentional Systems. *The Journal of Philosophy*, *68*(4), 87–106. https://doi.org/10.2307/2025382
Dennett, D. (1987). *The Intentional Stance*. MIT Press.
De Vos, J. (2016). The Death and the Resurrection of (Psy)critique: The Case of Neuroeducation. *Foundations of Science*, *21*(1), 129–145. https://doi.org/10.1007/s10699-014-9369-8
Dias, B. G., Banerjee, S. B., Goodman, J. V., & Ressler, K. J. (2013). Towards new approaches to disorders of fear and anxiety. *Current opinion in neurobiology*, *23*(3), 346–352. https://doi.org/10.1016/j.conb.2013.01.013
D'Ignazio, C., & Klein F., L. (2020). *Data Feminism*. Cambridge, MA; London, England: MIT Press.
DiLuca, M., & Olesen, J. (2014). The Cost of Brain Diseases: A Burden or a Challenge? *Neuron*, *82*(6), 1205–1208. https://doi.org/10.1016/j.neuron.2014.05.044
Dinh, T. N., & Thai, M. T. (2018). AI and Blockchain: A Disruptive Integration. *Computer*, *51*(9), 48–53. Presented at the Computer. https://doi.org/10.1109/MC.2018.3620971
Doerner, R., Broll, W., Jung, B., Grimm, P., Göbel, M., & Kruse, R. (2022). Introduction to Virtual and Augmented Reality. In *Virtual and Augmented Reality (VR/AR)* (pp. 1–37). Springer.
Drew, C. H., & Nyerges, T. L. (2004). Transparency of environmental decision making: A case study of soil cleanup inside the Hanford 100 area. *Journal of risk research*, *7*(1), 33–71.
Dubler, N. N. (2021). Lying is Not an Option for Clinical Ethics Consultants. *The American Journal of Bioethics*, *21*(5), 13–15. https://doi.org/10.1080/15265161.2021.1907125
Duman, H., & Ozkara, B. Y. (2021). The impact of social identity on online game addiction: the mediating role of the fear of missing out (FoMO) and the moderating role of the need to belong. *Current Psychology*, *40*(9), 4571–4580. https://doi.org/10.1007/s12144-019-00392-w
Dunatchik, A., Gerson, K., Glass, J., Jacobs, J. A., & Stritzel, H. (2021). Gender, parenting, and the rise of remote work during the pandemic: Implications for domestic inequality in the United States. *Gender & Society*, *35*(2), 194–205.
Dupuy, J.-P. (2010). The narratology of lay ethics. *Nanoethics*, *4*(2), 153–170.
Dupuy, J.-P. (2012). The precautionary principle and enlightened doomsaying. *Revue de métaphysique et de morale*, *76*(4), 577–592. https://doi.org/10.3917/rmm.124.0577
Dupuy, J.-P., & Grinbaum, A. (2004). Living with Uncertainty: Toward the Ongoing Normative Assessment of Nanotechnology'. In *Nanotechnology challenges: Implications for philosophy, ethics and society* (pp. 287-314 22).
Dupuy, J.-P., & Grinbaum, A. (2006). Living with uncertainty: toward the ongoing normative assessment of nanotechnology. In *Nanotechnology challenges: Implications for philosophy, ethics and society* (pp. 287–314). World Scientific.
Dworkin, R. (1986). *Law's Empire*. Harvard University Press.
Eaton, M. L., & Illes, J. (2007). Commercializing cognitive neurotechnology—the ethical terrain. *Nature Biotechnology*, *25*(4), 393–397. https://doi.org/10.1038/nbt0407-393

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No.101006249.

Ebrahimi, J., Rao, A., Lowd, D., & Dou, D. (2018). HotFlip: White-Box Adversarial Examples for Text Classification. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)* (pp. 31–36). Presented at the ACL 2018, Melbourne, Australia: Association for Computational Linguistics. https://doi.org/10.18653/v1/P18-2006

El Saddik, A. (2018). Digital twins: The convergence of multimedia technologies. *IEEE multimedia*, *25*(2), 87–92.

Emmelkamp, P. M., & Meyerbröker, K. (2021). Virtual reality therapy in mental health. *Annual Review of Clinical Psychology*, *17*, 495–519.

Emran, M. A., & Shaalan, K. (2014). A Survey of Intelligent Language Tutoring Systems. In *2014 International Conference on Advances in Computing, Communications and Informatics (ICACCI)* (pp. 393–399). Presented at the 2014 International Conference on Advances in Computing, Communications and Informatics (ICACCI). https://doi.org/10.1109/ICACCI.2014.6968503

Eriksén, S. (2002). Designing for accountability. In *Proceedings of the second Nordic conference on Human-computer interaction* (pp. 177–186). New York, NY, USA: Association for Computing Machinery. https://doi.org/10.1145/572020.572041

European Commission. Proposal for a REGULATION OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL on contestable and fair markets in the digital sector (Digital Markets Act). , COM/2020/842 final (2020). https://eur-lex.europa.eu/legal-content/en/TXT/?uri=COM%3A2020%3A842%3AFIN. Accessed 24 June 2022

European Commission. Proposal for a REGULATION OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL on a Single Market For Digital Services (Digital Services Act) and amending Directive 2000/31/EC (2020). https://eur-lex.europa.eu/legal-content/en/TXT/?qid=1608117147218&uri=COM%3A2020%3A825%3AFIN. Accessed 18 May 2022

European Commission. Proposal for a REGULATION OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL LAYING DOWN HARMONISED RULES ON ARTIFICIAL INTELLIGENCE (ARTIFICIAL INTELLIGENCE ACT) AND AMENDING CERTAIN UNION LEGISLATIVE ACTS (2021). https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A52021PC0206. Accessed 24 March 2022

Fagone, J. (2021, July 23). He couldn't get over his fiancee's death. So he brought her back as an A.I. chatbot. *The San Francisco Chronicle*. https://www.sfchronicle.com/projects/2021/jessica-simulation-artificial-intelligence/. Accessed 14 April 2022

Feiner, S. K. (1999). The importance of being mobile: some social consequences of wearable augmented reality systems. In *Proceedings 2nd IEEE and ACM International Workshop on Augmented Reality (IWAR'99)* (pp. 145–148). Presented at the Proceedings 2nd IEEE and ACM International Workshop on Augmented Reality (IWAR'99). https://doi.org/10.1109/IWAR.1999.803815

Feldman, F. (2004). *Pleasure and the Good Life: Concerning the Nature, Varieties, and Plausibility of Hedonism*. Clarendon Press.

Fereydooni, N., & Walker, B. N. (2020). Virtual Reality as a Remote Workspace Platform: Opportunities and Challenges. https://www.microsoft.com/en-us/research/publication/virtual-reality-as-a-remote-workspace-platform-opportunities-and-challenges/. Accessed 8 March 2022

Fileva, I., & Tresan, J. (2015). Will retributivism die and will neuroscience kill it? *Cognitive Systems Research*, *34–35*, 54–70. https://doi.org/10.1016/j.cogsys.2015.07.005

Findling, D. (2017, May 13). Bark app helps protect kids from cyberbullying and suicide, while safeguarding their privacy. *CNBC*. https://www.cnbc.com/2017/05/12/bark-app-helps-protect-kids-from-online-dangers-while-safeguarding-privacy.html. Accessed 7 April 2022

Fischer, E. F. (2014). *The Good Life: Aspiration, Dignity, and the Anthropology of Wellbeing. The Good Life*. Stanford University Press. https://doi.org/10.1515/9780804792615

Fitzpatrick, K. K., Darcy, A., & Vierhile, M. (2017). Delivering Cognitive Behavior Therapy to Young Adults With Symptoms of Depression and Anxiety Using a Fully Automated Conversational Agent (Woebot): A Randomized Controlled Trial. *JMIR Mental Health*, *4*(2), e7785. https://doi.org/10.2196/mental.7785

Fleischacker, S. (2020). Adam Smith's Moral and Political Philosophy. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Winter 2020.). Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/archives/win2020/entries/smith-moral-political/. Accessed 15 June 2022

Florman, S. C. (1994). *The Existential Pleasures of Engineering*. Macmillan.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No.101006249.

Følstad, A., Araujo, T., Law, E. L.-C., Brandtzaeg, P. B., Papadopoulos, S., Reis, L., et al. (2021). Future directions for chatbot research: an interdisciplinary research agenda. *Computing*, *103*(12), 2915–2942. https://doi.org/10.1007/s00607-021-01016-7
Forge, J. (2010). A Note on the Definition of "Dual Use." *Science and Engineering Ethics*, *16*(1), 111–118. https://doi.org/10.1007/s11948-009-9159-9
Freeman, D., Reeve, S., Robinson, A., Ehlers, A., Clark, D., Spanlang, B., & Slater, M. (2017). Virtual reality in the assessment, understanding, and treatment of mental health disorders. *Psychological medicine*, *47*(14), 2393–2400.
Friedman, B., & Hendry, D. G. (2019). *Value sensitive design: Shaping technology with moral imagination*. Mit Press.
Friedrich, O., Racine, E., & Steinert, S. (2021). An Analysis of the Impact of Brain-Computer Interfaces on Autonomy. *Neuroethics*, *14*, 17–29. https://doi.org/10.1007/s12152-018-9364-9
Fuss, S., Canadell, J. G., Peters, G. P., Tavoni, M., Andrew, R. M., Ciais, P., & Jackson, R. B. (2014). Betting on Negative Emissions. *Nature Climate Change*, *4*(10), 850–53. https://doi.org/10.1038/nclimate2392.
Gale, N., Golledge, R. G., Pellegrino, J. W., & Doherty, S. (1990). The acquisition and integration of route knowledge in an unfamiliar neighborhood. *Journal of Environmental Psychology*, *10*(1), 3–25. https://doi.org/10.1016/S0272-4944(05)80021-0
Galitsky, B. (2019). Chatbot Components and Architectures. In B. Galitsky (Ed.), *Developing Enterprise Chatbots: Learning Linguistic Structures* (pp. 13–51). Cham: Springer International Publishing. https://doi.org/10.1007/978-3-030-04299-8_2
Ganis, G. (2018). Lying and Neuroscience. In *The Oxford Handbook of Lying*. https://doi.org/10.1093/oxfordhb/9780198736578.013.35
Gardiner, S. M. (2010). Is 'Arming the Future' with Geoengineering Really the Lesser Evil? In S. Gardiner, S. Caney, D. Jamieson, & H. Shue (Eds.), *Climate Ethics: Essential Readings* (pp. 284–312). Oxford: Oxford University Press.
Gardiner, S. M. (2011). *A Perfect Moral Storm: The Ethical Tragedy of Climate Change*. New York: Oxford University Press.
Gardiner, S. M., & Fragnière, A. (2018). The Tollgate Principles for the Governance of Geoengineering: Moving Beyond the Oxford Principles to an Ethically More Robust Approach. *Ethics, Policy & Environment*, *21*(2), 143–74. https://doi.org/10.1080/21550085.2018.1509472.
Gardner, J. (2013). A history of deep brain stimulation: Technological innovation and the role of clinical assessment tools. *Social Studies of Science*, *43*(5), 707–728. https://doi.org/10.1177/0306312713483678
Geertz, C. (1977). *The Interpretation Of Cultures* (New e. édition.). New York: Basic Books.
General Electric. (2021). Digital Twins: The Bridge Between Industrial Assets and the Digital World. https://www.ge.com/digital/blog/digital-twins-bridge-between-industrial-assets-and-digital-world. Accessed 16 February 2022
Ghosh, A. (2019). Forecasting'. In H. Paul (Ed.), *Critical Terms in Futures Studies* (pp. 127–130). Cham: Springer International Publishing. https://doi.org/10.1007/978-3-030-28987-4_20.
Gieseke, A. P. (2020). " The New Weapon of Choice": Law's Current Inability to Properly Address Deepfake Pornography. *Vand. L. Rev.*, *73*, 1479.
Gilbert, F., O'Brien, T., & Cook, M. (2018). The effects of closed-loop brain implants on autonomy and deliberation: what are the risks of being kept in the loop? *Camb. Q. Healthc. Ethics*, *27*, 316–325. https://doi.org/10.1017/S0963180117000640
Gilbert, F., Viaña, J. N. M., & Ineichen, C. (2018). Deflating the "DBS causes personality changes" bubble. In *Neuroethics* (pp. 1–17). https://doi.org/10.1007/s12152-018-9373-8.
Glanville, R. (1982). Inside every white box there are two black boxes to get out. *Behavioral Sceince*, *27*, 1–11.
Goering, S., Klein, E., Specker Sullivan, L., Wexler, A., Agüera y Arcas, B., Bi, G., et al. (2021). Recommendations for Responsible Development and Application of Neurotechnologies. *Neuroethics*, *14*(3), 365–386. https://doi.org/10.1007/s12152-021-09468-6
Goodman, C. C. (2019). AI/Esq.: Impacts of Artificial Intelligence in Lawyer-Client Relationships. *Oklahoma Law Review*, *72*, 149.
Google PAIR. (2022). People + AI Guidebook. https://design.google/ai-guidebook. Accessed 5 April 2022
Goyal, M. (2018, November 12). Leveraging chatbots. *Canadian Law*. https://www.canadianlawyermag.com/resources/legal-technology/leveraging-chatbots/275615. Accessed 30 March 2022

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No.101006249.

Graeber, D., & Wengrow, D. (2021). *The Dawn of Everything: A New History of Humanity*. Penguin UK.
Greene, J., & Cohen, J. (2004). For the law, neuroscience changes nothing and everything. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, *359*(1451), 1775–1785. https://doi.org/10.1098/rstb.2004.1546
Greenhalgh, T., & Swinglehurst, D. (2011). Studying technology use as social practice: the untapped potential of ethnography'. *BMC Medicine*, *9*(1), 45. https://doi.org/10.1186/1741-7015-9-45.
Grinbaum, A. (2010). The nanotechnological golem. *Nanoethics*, *4*(3), 191–198.
Grinbaum, A. (2012). Conclusion. In *Emerging Risks: A Strategic Management Guide*. Gower Publishing, Ltd.
Grinbaum, A. (2015). Uncanny Valley Explained by Girard's Theory [Turning Point]. *IEEE Robotics & Automation Magazine*, *22*(1), 152–150.
Grinbaum, A. (2019). *Les robots et le mal*. Paris: Desclée de Brouwer.
Grinbaum, A. (2020). On the scientist's moral luck and wholeheartedness. *Journal of Responsible Innovation*, *7*(sup2), S12–S24.
Grinbaum, A., Devillers, L., Adda, G., Chatila, R., Martin, C., Zolynski, C., & Villata, S. (2021). *Agents conversationnels: Enjeux d'éthique* (Report). Comité national pilote d'éthique du numérique; CCNE.
Grinbaum, A., & Groves, C. (2013). What is "responsible" about responsible innovation? Understanding the ethical issues. *Responsible innovation: Managing the responsible emergence of science and innovation in society*, 119–142.
Grunwald, A., & Hillerbrand, R. (2013). *Handbuch Technikethik*. Springer.
Guin, U. K. L. (2015). *A Wizard of Earthsea: The First Book of Earthsea*. Hachette UK.
Habermas, J. (2007). The Language Game of Responsible Agency and the Problem of Free Will: How can epistemic dualism be reconciled with ontological monism? *Philosophical Explorations*, *10*(1), 13–50. https://doi.org/10.1080/13869790601170128
Hagger, L., & Johnson, G. H. (2011). 'Super Kids': Regulating the Use of Cognitive and Psychological Enhancement in Children. *Law, Innovation and Technology*, *3*(1), 137–166. https://doi.org/10.5235/175799611796399867
Hampel, H., O'Bryant, S. E., Molinuevo, J. L., Zetterberg, H., Masters, C. L., Lista, S., et al. (2018). Blood-based biomarkers for Alzheimer disease: mapping the road to the clinic. *Nature Reviews Neurology*, *14*(11), 639–652. https://doi.org/10.1038/s41582-018-0079-7
Hao, K. (2019, December 27). In 2020, let's stop AI ethics-washing and actually do something. *MIT Technology Review*. https://www.technologyreview.com/2019/12/27/57/ai-ethics-washing-time-to-act/. Accessed 31 August 2020
Hao, K. (2021, March 11). How Facebook got addicted to spreading misinformation. *Technology Review*. https://www.technologyreview.com/2021/03/11/1020600/facebook-responsible-ai-misinformation/
Haraldsson, & Bonin. (2021). Combining foresight and systems dynamics in the project - Scenarios for a sustainable Europe 2050. In *2021 International System Dynamics Conference J*.
Harris, S. R., Kemmerling, R. L., & North, M. M. (2002). Brief virtual reality therapy for public speaking anxiety. *Cyberpsychology & behavior*, *5*(6), 543–550.
Haubensak, W., Kunwar, P., Cai, H., Ciocchi, S., Wall, N., Ponnusamy, R., et al. (2010). Genetic dissection of an amygdala microcircuit that gates conditioned fear. *Nature*, *468*(7321), 270–276. https://doi.org/10.1038/nature09553
Haykin, S. (2008). *Neural Networks and Learning Machines* (3rd edition.). Pearson Education India.
Heeter, C. (1992). Being there: The subjective experience of presence. *Presence Teleoperators Virtual Environ.*, *1*(2), 262–271.
Heyward, C. (2013). Situating and Abandoning Geoengineering: A Typology of Five Responses to Dangerous Climate Change. *PS: Political Science & Politics*, *46*(1), 23–27.
Hitlin, P., Olmstead, K., & Toor, S. (2017, November 29). FCC Net Neutrality Online Public Comments Contain Many Inaccuracies and Duplicates. *Pew Research Center: Internet, Science & Tech*. https://www.pewresearch.org/internet/2017/11/29/public-comments-to-the-federal-communications-commission-about-net-neutrality-contain-many-inaccuracies-and-duplicates/. Accessed 21 February 2022
Hobbes, T. (1999). *De Corpore*. Vrin.
Hochberg, L. R., Serruya, M. D., Friehs, G. M., Mukand, J. A., Saleh, M., Caplan, A. H., et al. (2006). Neuronal ensemble control of prosthetic devices by a human with tetraplegia. *Nature*, *442*(7099), 164–171. https://doi.org/10.1038/nature04970

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No.101006249.

Hoffman, H. G., Sharar, S. R., Coda, B., Everett, J. J., Ciol, M., Richards, T., & Patterson, D. R. (2004). Manipulating presence influences the magnitude of virtual reality analgesia. *Pain*, *111*(1–2), 162–168.
Höjer, M., & Mattsson, L.-G. (2000). Determinism and backcasting in future studies. *Futures*, *32*(7), 613–634. https://doi.org/10.1016/S0016-3287(00)00012-4
Holmes, D. S. (1978). Projection as a defense mechanism. *Psychological Bulletin*, *85*(4), 677.
Horton, J. B., Reynolds, J. L., Buck, H. J., Callies, D., Schäfer, S., Keith, D. W., & Rayner, S. (2018). Solar Geoengineering and Democracy. *Global Environmental Politics*, *18*(3), 5–24. https://doi.org/10.1162/glep_a_00466.
Horton, J., & Keith, D. (2016). Solar Geoengineering and Obligations to the Global Poor. In C. J. Preston (Ed.), *Climate Justice and Geoengineering. Ethics and Policy in the Atmospheric Anthropocene* (pp. 79–92). London ; New York: Rowman & Littlefield.
Hourdequin, M. (2018). Climate Change, Climate Engineering, and the 'Global Poor': What Does Justice Require?" Ethics. *Policy & Environment*, *21*(3), 270–88. https://doi.org/10.1080/21550085.2018.1562525.
Howell, S. (2019). *Big Data and Monopolization* (SSRN Scholarly Paper No. 3123976). Rochester, NY: Social Science Research Network. https://doi.org/10.2139/ssrn.3123976
Hsieh, K. (2019). *Transformer Poetry: Classic Poetry Reimagined by Artificial Intelligence*. Paper Gains Publishing.
Huang, S. A., & Bailenson, J. (2019). Close Relationships and Virtual Reality. In T. D. Parsons, L. Lin, & D. Cockerham (Eds.), *Mind, Brain and Technology: Learning in the Age of Emerging Technologies* (pp. 49–65). Cham: Springer International Publishing. https://doi.org/10.1007/978-3-030-02631-8_4
Hughes, J. (2013). Transhumanism and Personal Identity. In *The Transhumanist Reader* (pp. 227–233). John Wiley & Sons, Ltd. https://doi.org/10.1002/9781118555927.ch23
Hui, Y. (2016). *On the Existence of Digital Objects*. U of Minnesota Press.
Husserl, E. (1936, 1970). The crisis of European sciences and transcendental phenomenology, Evanston: Northwestern University Press (Translated from the German original "Die Krisis der europäischen Wissenschaften und die transzendentale Phänomenologie."
Hyde, M., & Power, D. (2006). Some Ethical Dimensions of Cochlear Implantation for Deaf Children and Their Families. *The Journal of Deaf Studies and Deaf Education*, *11*(1), 102–111. https://doi.org/10.1093/deafed/enj009
Ienca, M., & Andorno, R. (2017). Towards new human rights in the age of neuroscience and neurotechnology. *Life Sciences, Society and Policy*, *13*(1), 5. https://doi.org/10.1186/s40504-017-0050-1
Ienca, M., Haselager, P., & Emanuel, E. J. (2018). Brain leaks and consumer neurotechnology. *Nature Biotechnology*, *36*(9), 805–810. https://doi.org/10.1038/nbt.4240
Illich, I. (1976). *Medical Nemesis: The Expropriation of Health*. Pantheon Books.
Jamieson, D. (1996). Ethics and Intentional Climate Change. *Climatic Change*, *33*, 323–36.
Jansen, Philip, Henschke, Adam, Erden, Yasemin, Marchiori, Samuela, Brey, Philip, & Hoefsloot, Marit. (2021). *D5.7 Ethics by Design and Research Ethics for AI*. De Montfort University. https://figshare.dmu.ac.uk/articles/online_resource/D5_7_Ethics_by_Design_and_Research_Ethics_for_AI/16912345. Accessed 26 April 2022
Jarrin, S., & Finn, D. P. (2019). Optogenetics and its application in pain and anxiety research. *Neuroscience & Biobehavioral Reviews*, *105*, 200–211. https://doi.org/10.1016/j.neubiorev.2019.08.007
Jasnow, A. M., Ehrlich, D. E., Choi, D. C., Dabrowska, J., Bowers, M. E., McCullough, K. M., et al. (2013). Thy1-Expressing Neurons in the Basolateral Amygdala May Mediate Fear Inhibition. *The Journal of Neuroscience*, *33*(25), 10396–10404. https://doi.org/10.1523/JNEUROSCI.5539-12.2013
Jelinek, F. (1976). Continuous speech recognition by statistical methods. *Proceedings of the IEEE*, *64*(4), 532–556.
Jensen, S. R. (2020). SIENNA D3.4: Ethical Analysis of Human Enhancement Technologies. https://doi.org/10.5281/ZENODO.4068071
Jensen, S. R., Nagel, S., Brey, P., Ditzel, T., Rodrigues, R., Broadhead, S., & Wright, D. (2018). SIENNA D3.1: State-of-the-art Review: Human Enhancement. https://doi.org/10.5281/zenodo.4066557
Jeon, H., Youn, H., Ko, S., & Kim, T. (2022). Blockchain and AI Meet in the Metaverse. *Advances in the Convergence of Blockchain and Artificial Intelligence*, 73.
Jia, R., & Liang, P. (2017). Adversarial Examples for Evaluating Reading Comprehension Systems. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing* (pp.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No.101006249.

2021–2031). Presented at the EMNLP 2017, Copenhagen, Denmark: Association for Computational Linguistics. https://doi.org/10.18653/v1/D17-1215
Jiang, Y., Raghupathi, V., & Raghupathi, W. (2009). Content and design of corporate governance web sites. *Information Systems Management*, *26*(1), 13–27.
Jiménez-Alonso, B., & Brescó de Luna, I. (2022). Griefbots. A New Way of Communicating With The Dead? *Integrative Psychological and Behavioral Science*. https://doi.org/10.1007/s12124-022-09679-3
Johansen, J. P., Hamanaka, H., Monfils, M. H., Behnia, R., Deisseroth, K., Blair, H. T., & LeDoux, J. E. (2010). Optical activation of lateral amygdala pyramidal cells instructs associative fear learning. *Proceedings of the National Academy of Sciences of the United States of America*, *107*(28), 12692–12697. https://doi.org/10.1073/pnas.1002418107
Jonas, H. (1976). Responsibility Today: The Ethics of an Endangered Future. *Social Research*, *43*(1), 77–97.
Jonas, H. (1985). *The Imperative of Responsibility: In Search of an Ethics for the Technological Age*. University of Chicago Press.
Jones, M. (2003). Overcoming the Myth of Free Will in Criminal Law: The True Impact of the Genetic Revolution. *Duke Law Journal*, *52*(5), 1031–1053.
Joshi, P., Santy, S., Budhiraja, A., Bali, K., & Choudhury, M. (2021). The State and Fate of Linguistic Diversity and Inclusion in the NLP World. *arXiv:2004.09095 [cs]*. http://arxiv.org/abs/2004.09095. Accessed 5 April 2022
Kahane, G. (2011). Mastery Without Mystery: Why there is no Promethean Sin in Enhancement. *Journal of Applied Philosophy*, *28*(4), 355–368. https://doi.org/10.1111/j.1468-5930.2011.00543.x
Kamenetz, A. (2016, March 28). Software Flags "Suicidal" Students, Presenting Privacy Dilemma. *NPR*. https://www.npr.org/sections/ed/2016/03/28/470840270/when-school-installed-software-stops-a-suicide. Accessed 7 April 2022
Kant, I. (1998). Critique of pure reason. In *Critique of Pure Reason* (1st ed.). Cambridge: Cambridge University Press.
Kant, I. (2012). *Kant: Groundwork of the Metaphysics of Morals*. Cambridge University Press.
Kaplan, A. D., Cruit, J., Endsley, M., Beers, S. M., Sawyer, B. D., & Hancock, P. A. (2021). The effects of virtual reality, augmented reality, and mixed reality as training enhancement methods: A meta-analysis. *Human factors*, *63*(4), 706–726.
Kellner, D. (2020). Jean Baudrillard. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Winter 2020.). Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/archives/win2020/entries/baudrillard/. Accessed 22 February 2022
Kim, Y., & Sundar, S. S. (2012). Anthropomorphism of computers: Is it mindful or mindless? *Computers in Human Behavior*, *28*(1), 241–250.
Kircaburun, K., & Griffiths, M. D. (2018). Instagram addiction and the Big Five of personality: The mediating role of self-liking. *Journal of behavioral addictions*, *7*(1), 158–170.
Klein, E. (2017). Neuromodulation ethics: Preparing for brain–computer interface medicine. In *Neuroethics*. Oxford: Oxford University Press. https://doi.org/10.1093/oso/9780198786832.003.0007
Klibansky, R., Panofsky, E., & Saxl, F. (2019). *Saturn and melancholy: Studies in the history of natural philosophy, religion, and art*. McGill-Queen's Press-MQUP.
Knight, F. H. (1921). *Risk, uncertainty and profit*. New York: Sentry Press.
Kocoń, J., Figas, A., Gruza, M., Puchalska, D., Kajdanowicz, T., & Kazienko, P. (2021). Offensive, aggressive, and hate speech analysis: From data-centric to human-centered approach. *Information Processing & Management*, *58*(5), 102643. https://doi.org/10.1016/j.ipm.2021.102643
Kögel, J., Jox, R. J., & Friedrich, O. (2020). What is it like to use a BCI? – insights from an interview study with brain-computer interface users. In *BMC Med Ethics 21* (p. 2). https://doi.org/10.1186/s12910-019-0442-2
Køster, A., & Kofod, E. H. (Eds.). (2021). *Cultural, Existential and Phenomenological Dimensions of Grief Experience* (1st edition.). Routledge.
Kotchetkov, I. S., Hwang, B. Y., Appelboom, G., Kellner, C. P., & Connolly, E. S., Jr. (2010). Brain-computer interfaces: military, neurosurgical, and ethical perspective. *Neurosurgical Focus FOC*, *28*(5), 25.
Krauss, J. K., Lipsman, N., & Aziz, T. (2021). Technology of deep brain stimulation: current status and future directions. *Nat Rev Neurol*, *17*, 75–87. https://doi.org/10.1038/s41582-020-00426-z

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No.101006249.

Kucewicz, M. T., Berry, B. M., Kremen, V., Miller, L. R., Khadjevand, F., Ezzyat, Y., et al. (2018). Electrical Stimulation Modulates High γ Activity and Human Memory Performance. *eNeuro*, *5*(1), ENEURO.0369-17.2018. https://doi.org/10.1523/ENEURO.0369-17.2018
Kugler, L. (2021). The state of virtual reality hardware. *Communications of the ACM*, *64*(2), 15–16.
Lampe, K. (2017). *The Birth of Hedonism: The Cyrenaic Philosophers and Pleasure as a Way of Life*. Princeton University Press.
Landauer, R. (1991). Information is physical. *Physics Today*, *44*(5), 23–29.
Landhuis, E. (2017). Do D.I.Y. Brain-Booster Devices Work? *Scientific American*. https://www.scientificamerican.com/article/do-diy-brain-booster-devices-work/. Accessed 21 June 2022
Langener, S., VanDerNagel, J., Klaassen, R., Van der Valk, P., & Heylen, D. (2021). "Go up in smoke": Feasibility and initial acceptance of a virtual environment to measure tobacco craving in vulnerable individuals. In *2021 IEEE 9th International Conference on Serious Games and Applications for Health (SeGAH)* (pp. 1–8). IEEE.
Larson, W. (2019). Optimism'. In H. Paul (Ed.), *Critical Terms in Futures Studies* (pp. 209–213). Cham: Springer International Publishing. https://doi.org/10.1007/978-3-030-28987-4_33.
Laugier, S. (2011). Le care comme critique et comme féminisme. *Travail, genre et sociétés*, *26*(2), 183–188.
Lea, M., & Spears, R. (1995). Love at first byte? Building personal relationships over computer networks. In *Under-studied relationships: Off the beaten track* (pp. 197–233). Thousand Oaks, CA, US: Sage Publications, Inc.
Lebeck, K., Kohno, T., & Roesner, F. (2016). How to safely augment reality: Challenges and directions. In *Proceedings of the 17th International Workshop on Mobile Computing Systems and Applications* (pp. 45–50).
Leckie, A., Santos, A., Thévoz, L.-A., Epron, B., & Gaudinat, A. (2020). *Assistants vocaux, enceintes connectées et recherche d'information*. Haute école de gestion de Genève.
Lele, A. (2013). Virtual reality and its military utility. *Journal of Ambient Intelligence and Humanized Computing*, *4*(1), 17–26. https://doi.org/10.1007/s12652-011-0052-4
Lenzi, D. (2018). The Ethics of Negative Emissions. *Global Sustainability*, *1*(e7), 1–8. https://doi.org/10.1017/sus.2018.5.
Lenzi, D. (2021). On the Permissibility (Or Otherwise) of Negative Emissions. *Ethics, Policy & Environment*, *February*, 1–14. https://doi.org/10.1080/21550085.2021.1885249.
Lenzi, D., Jakob, M., Honegger, M., Droege, S., Heyward, J. C., & Kruger, T. (2021). Equity Implications of Net Zero Visions. *Climatic Change*, *169*(3), 20. https://doi.org/10.1007/s10584-021-03270-2.
Lenzi, D., Lamb, W. F., Hilaire, J., Kowarsch, M., & Minx, J. C. (2018). Don't Deploy Negative Emissions Technologies without Ethical Analysis. *Nature*, *561*(7723), 303. https://doi.org/10.1038/d41586-018-06695-5.
Lesniak, I. (2022, March 23). La death tech ou l'avenir radieux des pompes funèbres. *Les Echos*. https://www.lesechos.fr/weekend/business-story/la-death-tech-ou-lavenir-radieux-des-pompes-funebres-1395621. Accessed 7 April 2022
Levac, D. E., & Galvin, J. (2013). When is virtual reality "therapy"? *Archives of physical medicine and rehabilitation*, *94*(4), 795–798.
Levy, N. (2007). The responsibility of the psychopath revisited. *Philosophy, Psychiatry, & Psychology*, *14*(2), 129–138.
Levy, N. (2011). *Hard Luck: How Luck Undermines Free Will and Moral Responsibility*. Oxford: Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199601387.001.0001
Lewis, M., Yarats, D., Dauphin, Y. N., Parikh, D., & Batra, D. (2017). Deal or No Deal? End-to-End Learning for Negotiation Dialogues. *arXiv:1706.05125 [cs]*. http://arxiv.org/abs/1706.05125. Accessed 5 April 2022
Libet, B. (1985). Unconscious cerebral initiative and the role of conscious will in voluntary action. *Behavioral and brain sciences*, *8*(4), 529–539.
Libet, B. (1999). Do We Have Free Will? *Journal of Consciousness Studies*, *6*(8–9), 47–57.
Libet, B., Gleason, C. A., Wright, E. W., & Pearl, D. K. (1993). Time of Conscious Intention to Act in Relation to Onset of Cerebral Activity (Readiness-Potential). In B. Libet (Ed.), *Neurophysiology of Consciousness* (pp. 249–268). Boston, MA: Birkhäuser. https://doi.org/10.1007/978-1-4612-0355-1_15
Liu, B. (2012). Sentiment analysis and opinion mining. *Synthesis lectures on human language technologies*, *5*(1), 1–167.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No.101006249.

Liu, J. (2021). Social Robots as the bride?: Understanding the construction of gender in a Japanese social robot product. *Human-Machine Communication*, *2*, 105–120.
Lombard, M., & Ditton, T. (2006). At the Heart of It All: The Concept of Presence. *Journal of Computer-Mediated Communication*, *3*(2), 0–0. https://doi.org/10.1111/j.1083-6101.1997.tb00072.x
Lucy, L., & Bamman, D. (2021). Gender and Representation Bias in GPT-3 Generated Stories. In *Proceedings of the Third Workshop on Narrative Understanding* (pp. 48–55). Presented at the NAACL-NUSE 2021, Virtual: Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.nuse-1.5
Luhmann, N. (1984, 1995). Social systems, Stanford: Stanford University Press (Translated from the German original "Soziale Systeme: Grundriss einer allgemeinen Theorie."
Luhmann, N. (1986, 1989). Ecological communication, Chicago: University of Chicago Press (Translated from the German original "Ökologische Kommunikation."
Luhmann, N. (1997, 2012). Theory of society. Stanford: Stanford University Press.
Luong, P., Glorioso, T. J., Grunwald, G. K., Peterson, P., Allen, L. A., Khanna, A., et al. (2021). Text Message Medication Adherence Reminders Automated and Delivered at Scale Across Two Institutions: Testing the Nudge System: Pilot Study. *Circulation: Cardiovascular Quality and Outcomes*, *14*(5), e007015. https://doi.org/10.1161/CIRCOUTCOMES.120.007015
Mackenzie, S. (2022). Criminology towards the metaverse: Cryptocurrency scams, grey economy and the technosocial. *The British Journal of Criminology*, azab118. https://doi.org/10.1093/bjc/azab118
Mader, B., Banks, M. S., & Farid, H. (2017). Identifying computer-generated portraits: The importance of training and incentives. *Perception*, *46*(9), 1062–1076.
Mai, J.-E. (2019). Situating Personal Information: Privacy in the Algorithmic Age. In *Human rights in the age of platforms* (pp. 95–116). The MIT Press.
Makazhanov, A., Rafiei, D., & Waqar, M. (2014). Predicting political preference of Twitter users. *Social Network Analysis and Mining*, *4*(1), 193. https://doi.org/10.1007/s13278-014-0193-5
Mandarelli, G., Moretti, G., Pasquini, M., Nicolò, G., & Ferracuti, S. (2018). Informed Consent Decision-Making in Deep Brain Stimulation. *Brain sciences*, *8*(5), 84. https://doi.org/10.3390/brainsci8050084
Margalit, A. (1998). *The Decent Society*. Harvard University Press.
McGee, J., Brent, K., McDonald, J., & Heyward, C. (2021). International Governance of Solar Radiation Management: Does the ENMOD Convention Deserve a Closer Look? *Carbon & Climate Law Review*, *14*(4), 294–305. https://doi.org/10.21552/cclr/2020/4/8.
McIntosh, T., DuBois, J. M., & Perlmutter, J. S. (2022). Ethical Challenges in the Commercialization of Neurotechnology: Contending with Competing Priorities. *AJOB Neuroscience*, *13*(1), 60–62. https://doi.org/10.1080/21507740.2021.2001083
McKenna, K. Y., & Bargh, J. A. (2000). Plan 9 from cyberspace: The implications of the Internet for personality and social psychology. *Personality and social psychology review*, *4*(1), 57–75.
McKinnon, C. (2020). The Panglossian Politics of the Geoclique. *Critical Review of International Social and Political Philosophy*, *23*(5), 584–99. https://doi.org/10.1080/13698230.2020.1694216.
McLellan, H. (1996). Virtual realities. *Handbook of research for educational communications and technology*, 457–487.
Metaverse Property. (2020, September 4). Buy in Decentraland - Metaverse Property. https://metaverse.properties/buy-in-decentraland/. Accessed 14 April 2022
Meyers, C. (2021). Deception and the Clinical Ethicist. *The American Journal of Bioethics*, *21*(5), 4–12. https://doi.org/10.1080/15265161.2020.1863513
Milgram, P., Takemura, H., Utsumi, A., & Kishino, F. (1995). Augmented reality: A class of displays on the reality-virtuality continuum. In *Telemanipulator and telepresence technologies* (Vol. 2351, pp. 282–292). International Society for Optics and Photonics.
Mill, J. S. (1863). *Utilitarianism*. Parker, Son and Bourn.
Mill, J. S. (1978). *On Liberty*. Hackett Publishing.
Millar, L. (2019). Revolution'. In H. Paul (Ed.), *Critical Terms in Futures Studies* (pp. 253–259). Cham: Springer International Publishing. https://doi.org/10.1007/978-3-030-28987-4_39.
Miller, S., & Selgelid, M. J. (2008). Ethics and the dual-use dilemma in the life sciences. In *Physicians at War* (pp. 195–211). Springer.
Mills, C. (2005). "Ideal theory" as an Ideology. *Hypatia*, *20*(3), 165–184.
Miner, A. S., Milstein, A., Schueller, S., Hegde, R., Mangurian, C., & Linos, E. (2016). Smartphone-based conversational agents and responses to questions about mental health, interpersonal violence, and physical health. *JAMA internal medicine*, *176*(5), 619–625.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No.101006249.

Minx, J. C., Lamb, W. F., Callaghan, M. W., Fuss, S., Hilaire, J., Creutzig, F., & Amann, T. (2018). Negative Emissions—Part 1: Research Landscape and Synthesis. *Environmental Research Letters*, *13*(6), 063001. https://doi.org/10.1088/1748-9326/aabf9b.
Mirzayi, P., Shobeiri, P., Kalantari, A., Perry, G., & Rezaei, N. (2022). Optogenetics: implications for Alzheimer's disease research and therapy. *Molecular Brain*, *15*(1), 20. https://doi.org/10.1186/s13041-022-00905-y
Mittelstadt, B. (2019). Principles alone cannot guarantee ethical AI. *Nature Machine Intelligence*, *1*, 501–507.
Moncur, W., Masthoff, J., Reiter, E., Freer, Y., & Nguyen, H. (2014). Providing adaptive health updates across the personal social network. *Human–Computer Interaction*, *29*(3), 256–309.
Mori, M., MacDorman, K. F., & Kageki, N. (2012). The uncanny valley [from the field]. *IEEE Robotics & automation magazine*, *19*(2), 98–100.
Morrow, D. R. (2020). A Mission-Driven Research Program on Solar Geoengineering Could Promote Justice and Legitimacy. *Critical Review of International Social and Political Philosophy*, *23*(5), 618–40. https://doi.org/10.1080/13698230.2020.1694220.
Muñoz, K. A., Kostick, K., Sanchez, C., Kalwani, L., Torgerson, L., Hsu, R., et al. (2020). Researcher Perspectives on Ethical Considerations in Adaptive Deep Brain Stimulation Trials. In *Frontiers in Human Neuroscience* (p. 14). https://doi.org/10.3389/fnhum.2020.578695
Munyon, C. N. (2018). Neuroethics of Non-primary Brain Computer Interface: Focus on Potential Military Applications. *Front. Neurosci*, *12*, 696. https://doi.org/10.3389/fnins.2018.00696
Murphy, H. (2022). Facebook patents reveal how it intends to cash in on metaverse. *Financial Times*. https://www.ft.com/content/76d40aac-034e-4e0b-95eb-c5d34146f647. Accessed 16 February 2022
Nagel, T. (1974). What is it like to be a bat. *Readings in philosophy of psychology*, *1*, 159–168.
Nahmias, E. (2012). Free will and responsibility. *WIREs Cognitive Science*, *3*(4), 439–449. https://doi.org/10.1002/wcs.1181
Nahmias, E. (2014). Is Free Will an Illusion? Confronting Challenges from the Modern Mind Sciences. In *Moral Psychology, Volume 4*. The MIT Press. https://doi.org/10.7551/mitpress/9780262026680.003.0002
Nemet, G. F., Callaghan, M. W., Creutzig, F., Fuss, S., Hartmann, J., Hilaire, J., et al. (2018). Negative Emissions—Part 3: Innovation and Upscaling. *Environmental Research Letters*, *13*(6), 063003. https://doi.org/10.1088/1748-9326/aabff4.
Neuronews. (2022, May 10). Synchron enrols first patient in US COMMAND study of endovascular brain-computer interface. *NeuroNews International*. https://neuronewsinternational.com/synchron-enrols-first-patient-in-us-command-study-of-endovascular-brain-computer-interface/. Accessed 13 May 2022
Nishimoto, S., Vu, A. T., Naselaris, T., Benjamini, Y., Yu, B., & Gallant, J. L. (2011). Reconstructing Visual Experiences from Brain Activity Evoked by Natural Movies. *Current Biology*, *21*(19), 1641–1646. https://doi.org/10.1016/j.cub.2011.08.031
Norris, K. O. (2004). Gender stereotypes, aggression, and computer games: An online survey of women. *Cyberpsychology & Behavior*, *7*(6), 714–727.
North, M. M., North, S. M., & Coble, J. R. (1997). Virtual reality therapy: An effective treatment for psychological. *Virtual reality in neuro-psycho-physiology: Cognitive, clinical and methodological issues in assessment and rehabilitation*, *44*, 59.
Nozick, R. (1974). *Anarchy, state, and utopia* (Vol. 5038). new york: Basic Books.
OECD. (2019). Recommendation of the Council on Responsible Innovation in Neurotechnology. OECD Publishing Paris, France.
Olson, D., & Che, T. (2022). "The Problem With NFTs": A Crypto Expert Responds to a Viral Takedown. *Time*. https://time.com/6144332/the-problem-with-nfts-video/. Accessed 16 February 2022
Paikari, E., & van der Hoek, A. (2018). A Framework for Understanding Chatbots and Their Future. In *2018 IEEE/ACM 11th International Workshop on Cooperative and Human Aspects of Software Engineering (CHASE)* (pp. 13–16). Presented at the 2018 IEEE/ACM 11th International Workshop on Cooperative and Human Aspects of Software Engineering (CHASE).
Pamplany, A., Gordijn, B., & Brereton, P. (2020). The Ethics of Geoengineering: A Literature Review. *Science and Engineering Ethics*, *26*(6), 3069–3119. https://doi.org/10.1007/s11948-020-00258-6.
Pardes, A. (2018). The Emotional Chatbots Are Here to Probe Our Feelings. *Wired*. https://www.wired.com/story/replika-open-source/. Accessed 24 March 2022
Parikka, J. (2015). *A Geology of Media*. U of Minnesota Press.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No.101006249.

Parker, A., & Irvine, P. J. (2018). The Risk of Termination Shock From Solar Geoengineering. *Earth's Future*, *6*(3), 456–67. https://doi.org/10.1002/2017EF000735.
Parkin, S. (2022, January 9). The trouble with Roblox, the video game empire built on child labour. *The Observer*. https://www.theguardian.com/games/2022/jan/09/the-trouble-with-roblox-the-video-game-empire-built-on-child-labour. Accessed 14 April 2022
Pasquale, F. (2015). *The black box society*. Harvard University Press.
Pereboom, D. (2014). *Free Will, Agency, and Meaning in Life*. Oxford: Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199685516.001.0001
Perez, C. C. (2019). *Invisible Women: the Sunday Times number one bestseller exposing the gender bias women face every day*. Random House.
Perez-Marin, D., & Pascual-Nieto, I. (2011). *Conversational agents and natural language interaction: Techniques and effective practices: Techniques and effective practices*. IGI Global.
Persily, N., & Tucker, J. A. (2020). *Social Media and Democracy: The State of the Field, Prospects for Reform*. Cambridge University Press.
Peters, L. (2018). The History Of Furby, The Electronic Pet That Took The Late '90s By Storm. *Bustle*. https://www.bustle.com/p/the-history-of-furby-the-electronic-pet-that-took-the-late-90s-by-storm-8080509. Accessed 24 March 2022
Petschick, G. (2015). Ethnographic panels for analyzing innovation processes. *Historical Social Research/Historische Sozialforschung*, 210–232.
Pink, S., Horst, H., Postill, J., Hjorth, L., Lewis, T., & J, T. (Eds.). (2016). *Digital ethnography: principles and practice*. Los Angeles: SAGE.
Pink, S., & Morgan, J. (2013). Short-Term Ethnography: Intense Routes to Knowing: Short-Term Ethnography'. *Symbolic Interaction*, *36*(3), 351–361. https://doi.org/10.1002/symb.66.
Piumsomboon, T., Lee, G. A., Hart, J. D., Ens, B., Lindeman, R. W., Thomas, B. H., & Billinghurst, M. (2018). Mini-Me: An Adaptive Avatar for Mixed Reality Remote Collaboration. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (pp. 1–13). New York, NY, USA: Association for Computing Machinery. https://doi.org/10.1145/3173574.3173620. Accessed 8 March 2022
Piumsomboon, T., Lee, Y., Lee, G., & Billinghurst, M. (2017). CoVAR: a collaborative virtual and augmented reality system for remote collaboration. In *SIGGRAPH Asia 2017 Emerging Technologies* (pp. 1–2). New York, NY, USA: Association for Computing Machinery. https://doi.org/10.1145/3132818.3132822
Politi, V., & Grinbaum, A. (2020). The distribution of ethical labor in the scientific community. *Journal of Responsible Innovation*, *7*(3), 263–279.
Polybius. (1980). *The Rise of the Roman Empire*. (I. Scott-Kilvert, Trans., F. W. Walbank, Ed.) (Reprint edition.). Harmondsworth ; New York: Penguin Classics.
Ponnusamy, S., Iranmanesh, M., Foroughi, B., & Hyun, S. S. (2020). Drivers and outcomes of Instagram Addiction: Psychological well-being as moderator. *Computers in Human Behavior*, *107*, 106294.
Potts, J. (2018). Futurism, Futurology, Future Shock, Climate Change: Visions of the Future from 1909 to the Present. *PORTAL Journal of Multidisciplinary International Studies*, *15*(1–2), 99–116. https://doi.org/10.5130/portal.v15i1-2.5810
Prasad, P. (1997). Systems of Meaning: Ethnography as a Methodology for the Study of Information Technologies'. In A. S. Lee, J. Liebenau, & J. I. DeGross (Eds.), *Information Systems and Qualitative Research* (pp. 101–118). Boston, MA: Springer US. https://doi.org/10.1007/978-0-387-35309-8_7.
Preston, C. J. (2013). Ethics and Geoengineering: Reviewing the Moral Issues Raised by Solar Radiation Management and Carbon Dioxide Removal. *Wiley Interdisciplinary Reviews: Climate Change*, *4*(1), 23–37. https://doi.org/10.1002/wcc.198.
Putnam, H. (1979). *Philosophical Papers: Volume 2, Mind, Language and Reality* (Vol. 2). Cambridge University Press.
Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). Language models are unsupervised multitask learners. *OpenAI blog*, *1*(8), 9.
Rae, J., Irving, G., & Weidinger, L. (2021). Language modelling at scale: Gopher, ethical considerations, and retrieval. https://www.deepmind.com/blog/language-modelling-at-scale-gopher-ethical-considerations-and-retrieval. Accessed 7 April 2022
Rae, J. W., Borgeaud, S., Cai, T., Millican, K., Hoffmann, J., Song, F., et al. (2021). Scaling language models: Methods, analysis & insights from training gopher. *arXiv preprint arXiv:2112.11446*.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No.101006249.

Ramirez, E. J. (2019). Ecological and ethical issues in virtual reality research: A call for increased scrutiny. *Philosophical Psychology*, *32*(2), 211–233. https://doi.org/10.1080/09515089.2018.1532073
Ramirez, E. J., Elliott, M., & Milam, P.-E. (2021). What it's like to be a _____: why it's (often) unethical to use VR as an empathy nudging tool. *Ethics and Information Technology*, *23*(3), 527–542. https://doi.org/10.1007/s10676-021-09594-y
Rauschnabel, P. A., Rossmann, A., & tom Dieck, M. C. (2017). An adoption framework for mobile augmented reality games: The case of Pokémon Go. *Computers in Human Behavior*, *76*, 276–286.
Ray, S. G. (2003). *Gender Inclusive Game Design: Expanding the Market (Advances in Computer Graphics and Game Development Series)*. Charles River Media, Inc.
Rayner, S. (2010). Trust and the Transformation of Energy Systems." Energy Policy, The Role of Trust in Managing Uncertainties in the Transition to a Sustainable Energy Economy. *Special Section with Regular Papers*, *38*(6), 2617–23. https://doi.org/10.1016/j.enpol.2009.05.035.
Reiners, D., Davahli, M. R., Karwowski, W., & Cruz-Neira, C. (2021). The Combination of Artificial Intelligence and Extended Reality: A Systematic Review. *Frontiers in Virtual Reality*, 114.
Reisach, U. (2021). The responsibility of social media in times of societal and political manipulation. *European Journal of Operational Research*, *291*(3), 906–917. https://doi.org/10.1016/j.ejor.2020.09.020
Rességuier, A., & Rodrigues, R. (2020). AI ethics should not remain toothless! A call to bring back the teeth of ethics. *Big Data & Society*, *7*(2), 2053951720942541. https://doi.org/10.1177/2053951720942541
Resseguier, A., & Rodrigues, R. (2021). Ethics as attention to context: recommendations for the ethics of artificial intelligence. *Open Research Europe*.
Ricci, G. R. (2013). *Culture and Civilization: Cosmopolitanism and the Global Polity*. Transaction Publishers.
Riley, J. (2013). Isaiah Berlin's "Minimum of Common Moral Ground." *Political Theory*, *41*(1), 61–89.
Roach, J. (2021, November 2). Mesh for Microsoft Teams aims to make collaboration in the 'metaverse' personal and fun. *Microsoft: Innovation Stories*. https://news.microsoft.com/innovation-stories/mesh-for-microsoft-teams/. Accessed 8 March 2022
Roblox. (2022). DevHub. https://developer.roblox.com/en-us/. Accessed 14 April 2022
Rochfeld, J. (2022). Les avatars d'éternité (ou prolongations numériques des défunts) : vers de nouvelles personnes résiduelles compassionnelles ? In *Mélanges en l'honneur du professeur Catherine Labrusse-Riou*. https://www.lgdj.fr/melanges-en-l-honneur-du-professeur-catherine-labrusse-riou-9782850020520.html. Accessed 24 June 2022
Roesner, F., Kohno, T., & Molnar, D. (2014). Security and privacy for augmented reality systems. *Communications of the ACM*, *57*(4), 88–96. https://doi.org/10.1145/2580723.2580730
Rosa, H. (2016). *Resonanz: Eine Soziologie der Weltbeziehung*. Berlin: Suhrkamp.
Roskies, A. (2002). Neuroethics for the new millenium. *Neuron*, *35*(1), 21–23. https://doi.org/10.1016/s0896-6273(02)00763-8
Ruane, E., Birhane, A., & Ventresque, A. (2019). Conversational AI: Social and Ethical Considerations. In *AICS* (pp. 104–115).
Rudinow, J. (1978). Manipulation. *Ethics*, *88*(4), 338–347.
Sadowski, J. (2019). When data is capital: Datafication, accumulation, and extraction. *Big Data & Society*, *6*(1), 2053951718820549. https://doi.org/10.1177/2053951718820549
Sætra, H. S. (2019). When nudge comes to shove: Liberty and nudging in the era of big data. *Technology in Society*, *59*, 101130. https://doi.org/10.1016/j.techsoc.2019.04.006
Sagnier, C., Loup-Escande, É., & Valléry, G. (2021). Virtual Reality: Definitions, Characteristics and Applications in the Workplace. In *Digital Transformations in the Challenge of Activity and Work* (pp. 31–44). John Wiley & Sons, Ltd. https://doi.org/10.1002/9781119808343.ch3
Sandel, M. (2007). *The Case Against Perfection: Ethics in the Age of Genetic Engineering*. Cambrige: Harvard University Press.
Sardar, Z. (2010). The Namesake: Futures; futures studies; futurology; futuristic; foresight—What's in a name?'. *Futures*, *42*(3), 177–184. https://doi.org/10.1016/j.futures.2009.11.001.
Sax, M. (2021). *Between Empowerment and Manipulation: The Ethics and Regulation of For-Profit Health Apps*. Kluwer Law International B.V.
Schirrmeister, R. T., Springenberg, J. T., Fiederer, L. D. J., Glasstetter, M., Eggensperger, K., Tangermann, M., et al. (2017). Deep learning with convolutional neural networks for EEG decoding and visualization. *Human brain mapping*, *38*(11), 5391–5420.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No.101006249.

Schmidt, A. T., & Engelen, B. (2020). The ethics of nudging: An overview. *Philosophy Compass*, *15*(4), e12658.
Schmidt, M. W., Köppinger, K. F., Fan, C., Kowalewski, K.-F., Schmidt, L. P., Vey, J., et al. (2021). Virtual reality simulation in robot-assisted surgery: meta-analysis of skill transfer and predictability of skill. *BJS open*, *5*(2), zraa066.
Schneider, R. (2019). Gesture'. In H. Paul (Ed.), *Critical Terms in Futures Studies* (pp. 145–149). Cham: Springer International Publishing. https://doi.org/10.1007/978-3-030-28987-4_23.
Schoemaker, P. J. H. (1995). Scenario planning: A tool for strategic thinking. *MIT Sloan Management Review*, *36*.
Schrader, S., Jones, N., & Shattell, M. (2013). Mad Pride: Reflections on Sociopolitical Identity and Mental Diversity in the Context of Culturally Competent Psychiatric Care. *Issues in Mental Health Nursing*, *34*(1), 62–64. https://doi.org/10.3109/01612840.2012.740769
Schübel, H., & Wallimann-Helmer, I. (2021). Food Security and the Moral Differences between Climate Mitigation and Geoengineering: The Case of Biofuels and BECCS. In *Justice and Food Security in a Changing Climate, 71–76. Conference Proceedings*. Wageningen Academic Publishers. https://doi.org/10.3920/978-90-8686-915-2_8.
Schwägerl, C., & Crutzen, P. (2018). *The Anthropocene: The Human Era and How It Shapes Our Planet*. Synergetic Press.
Scotus, J. D. (1997). *Duns Scotus on the Will and Morality*. CUA Press.
Searle, J. R. (1980). Minds, brains, and programs. *Behavioral and brain sciences*, *3*(3), 417–424.
Sensi, M., Eleopra, R., Cavallo, M. A., Sette, E., Milani, P., Quatrale, R., et al. (2004). Explosive-aggressive behavior related to bilateral subthalamic stimulation. *Parkinsonism & Related Disorders*, *10*(4), 247–251. https://doi.org/10.1016/j.parkreldis.2004.01.007
Sententia, W. (2004). Neuroethical Considerations: Cognitive Liberty and Converging Technologies for Improving Human Cognition. *Annals of the New York Academy of Sciences*, *1013*(1), 221–228. https://doi.org/10.1196/annals.1305.014
Sessa, C., Cassolá, D., & Kienegger, M. (2021). Foresight on demand: Development and enrichment of key factor descriptions for the the project - Scenarios for a sustainable Europe in.
Sethumadhavan, A., & Phisuthikul, A. (2019). Can Machines Detect Emotions? *Ergonomics in Design*, *27*(3), 30–30. https://doi.org/10.1177/1064804619847190
Sharma, M., & Kaur, M. (2022). A Review of Deepfake Technology: An Emerging AI Threat. In G. Ranganathan, X. Fernando, F. Shi, & Y. El Allioui (Eds.), *Soft Computing for Security Applications* (pp. 605–619). Singapore: Springer. https://doi.org/10.1007/978-981-16-5301-8_44
Sharma, U. (2022, February 5). What Is an Avatar in the Metaverse? *Beebom*. https://beebom.com/metaverse-avatars-explained/. Accessed 16 February 2022
Shepherd, J., Caldeira, K., Cox, P., Haigh, J., Keith, D., Launder, B., & Mace, G. (2009). Geoengineering the Climate: Science, Governance, and Uncertainty. http://royalsociety.
Shibata, S. B., Cortez, S. R., Beyer, L. A., Wiler, J. A., Di Polo, A., Pfingst, B. E., & Raphael, Y. (2010). Transgenic BDNF induces nerve fiber regrowth into the auditory epithelium in deaf cochleae. *Experimental Neurology*, *223*(2), 464–472. https://doi.org/10.1016/j.expneurol.2010.01.011
Sholeh, A., & Rusdi, A. (2019). A new measurement of Instagram addiction: psychometric properties of The Instagram Addiction Scale (TIAS). *feedback*, *737*, 499.
Shope, M. L. (2021). Lawyer and Judicial Competency in the Era of Artificial Intelligence: Ethical Requirements for Documenting Datasets and Machine Learning Models. *Georgetown Journal of Legal Ethics*, *34*, 191.
Shotbolt, P., Moriarty, J., Costello, A., Jha, A., David, A., Ashkan, K., & Samuel, M. (2012). Relationships between deep brain stimulation and impulse control disorders in Parkinson's disease, with a literature review. *Parkinsonism & Related Disorders*, *18*(1), 10–16. https://doi.org/10.1016/j.parkreldis.2011.08.016
Shue, H. (2017). Climate Dreaming: Negative Emissions, Risk Transfer, and Irreversibility. *Journal of Human Rights and the Environment*, *8*(2), 203–16. https://doi.org/10.4337/jhre.2017.02.02.
Shue, H. (2018). Mitigation Gambles: Uncertainty, Urgency and the Last Gamble Possible. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, *376*(2119), 20170105. https://doi.org/10.1098/rsta.2017.0105.
Skarbez, R., Smith, M., & Whitton, M. C. (2021). Revisiting milgram and kishino's reality-virtuality continuum. *Frontiers in Virtual Reality*, *2*, 27.
Slater, M., & Wilbur, S. (1997). A framework for immersive virtual environments (FIVE): Speculations on the role of presence in virtual environments. *Presence: Teleoperators & Virtual Environments*, *6*(6), 603–616.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No.101006249.

Smith, A. (2010). *The Theory of Moral Sentiments*. Penguin.
Smith, P. (2021). Who May Geoengineer: Global Domination, Revolution, and Solar Radiation Management. *Global Justice : Theory Practice Rhetoric*, *13*, 138–165. https://doi.org/10.21248/gjn.13.01.237
Song, S. (2019, December 14). CGI Influencer Lil Miquela Criticized For "Sexual Assault" Vlog. *PAPER*. https://www.papermag.com/lil-miquela-sexual-assault-vlog-2641593301.html. Accessed 23 March 2022
Southgate, E., Smith, S. P., & Scevak, J. (2017). Asking ethical questions in research using immersive virtual and augmented reality technologies with children and youth. In *2017 IEEE Virtual Reality (VR)* (pp. 12–18). Presented at the 2017 IEEE Virtual Reality (VR). https://doi.org/10.1109/VR.2017.7892226
Speicher, M., Hall, B. D., & Nebeling, M. (2019). What is mixed reality? In *Proceedings of the 2019 CHI conference on human factors in computing systems* (pp. 1–15).
Spengler, B. (2019). Imagination'. In H. Paul (Ed.), *Critical Terms in Futures Studies* (pp. 163–169). Cham: Springer International Publishing. https://doi.org/10.1007/978-3-030-28987-4_26.
Spinoza, B. de. (2018). *Spinoza: Ethics: Demonstrated in Geometric Order*. Cambridge University Press.
Stahl, B. C., Timmermans, J., & Flick, C. (2017). Ethics of Emerging Information and Communication TechnologiesOn the implementation of responsible research and innovation. *Science and Public Policy*, *44*(3), 369–381.
Stanton, S. J., Sinnott-Armstrong, W., & Huettel, S. A. (2017). Neuromarketing: Ethical Implications of its Use and Potential Misuse. *Journal of Business Ethics*, *144*(4), 799–811. https://doi.org/10.1007/s10551-016-3059-0
Stein, S. (2022). Magic Leap 2 Hands-On: AR Glasses That Can Dim the Real World. *CNET*. https://www.cnet.com/tech/computing/features/magic-leap-2-hands-on-ar-glasses-that-can-dim-the-real-world/. Accessed 16 March 2022
Strawson, P. (1962). Freedom and Resentment. *Proceedings of the British Academy*, *48*, 187–211.
Susser, D., Roessler, B., & Nissenbaum, H. (2018). *Online Manipulation: Hidden Influences in a Digital World* (SSRN Scholarly Paper No. ID 3306006). Rochester, NY: Social Science Research Network. https://doi.org/10.2139/ssrn.3306006
Swan, M. (2015). *Blockchain*. O'Reilly Media.
Swierstra, T., & Rip, A. (2007). Nano-ethics as NEST-ethics: Patterns of Moral Argumentation About New and Emerging Science and Technology. *NanoEthics*, *1*(1), 3–20. https://doi.org/10.1007/s11569-007-0005-8
Takahashi, D. (2022, April 23). Craig Donato interview: How Roblox navigates brands, UGC, and the metaverse. *VentureBeat*. https://venturebeat.com/2022/04/23/craig-donato-interview-how-roblox-navigates-brands-ugc-and-the-metaverse/. Accessed 25 April 2022
Tännsjö, T. (2007). Narrow hedonism. *Journal of Happiness Studies*, *8*(1), 79–98.
Thaler, R. H., & Sunstein, C. R. (2008). *Nudge: Improving Decisions about Health, Wealth, and Happiness*. Yale University Press.
Tham, J., Duin, A. H., Gee, L., Ernst, N., Abdelqader, B., & McGrath, M. (2018). Understanding Virtual Reality: Presence, Embodiment, and Professional Practice. *IEEE Transactions on Professional Communication*, *61*(2), 178–195. Presented at the IEEE Transactions on Professional Communication. https://doi.org/10.1109/TPC.2018.2804238
Theis, & Köppe. (2018). Peace operations 2025: From shaping factors to scenarios. In *Envisioning uncertain futures: Scenarios as a tool in security, privacy and mobility research* (pp. 155–173). Wiesbaden: Springer.
Thoppilan, R., De Freitas, D., Hall, J., Shazeer, N., Kulshreshtha, A., Cheng, H.-T., et al. (2022). LaMDA: Language Models for Dialog Applications. *arXiv:2201.08239 [cs]*. http://arxiv.org/abs/2201.08239. Accessed 9 February 2022
Tiku, N. (2022, June 11). The Google engineer who thinks the company's AI has come to life. *Washington Post*. https://www.washingtonpost.com/technology/2022/06/11/google-ai-lamda-blake-lemoine/. Accessed 24 June 2022
Turing, A. M. (1936). On computable numbers, with an application to the Entscheidungsproblem. *J. of Math*, *58*(345–363), 5.
Underwood, P. J. R., Teehan, G. D. J., III, G. L. K., Holland, J., & Westerhold, K. M. (2021, October 26). Display device with graphical user interface. https://patents.google.com/patent/USD934286S1/en?q=%22Display+device+with+graphical+user+interface%22+underwood&oq=%22Display+device+with+graphical+user+interface%22+underwood. Accessed 11 March 2022

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No.101006249.

UNESCO. (2008). *On consent: report*. Paris: Unesco.
UNESCO. (2021). Report of the International Bioethics Committee of UNESCO (IBC) on the ethical issues of neurotechnology.
UNESCO, U. (1982). Mexico City declaration on cultural policies. In *World Conference on Cultural Policies*.
Van den Hoven, J., Miller, S., & Pogge, T. (2017). The design turn in applied ethics. *Designing in ethics*, 11–31.
Van den Hoven, J., Vermaas, P. E., & Van de Poel, I. (2015). *Handbook of ethics, values, and technological design: Sources, theory, values and application domains*. Springer.
Van Dijk, T. A. (2006). Discourse and manipulation. *Discourse & society*, *17*(3), 359–383.
Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., et al. (2017). Attention is all you need. *Advances in neural information processing systems*, *30*.
Veit, W. (2018). Cognitive Enhancement and the Threat of Inequality. *Journal of Cognitive Enhancement*, *2*(4), 404–410. https://doi.org/10.1007/s41465-018-0108-x
Vézina, B., & Hinchliff Pearson, S. (2021, March 4). Should CC-Licensed Content be Used to Train AI? It Depends. *Creative Commons*. https://creativecommons.org/2021/03/04/should-cc-licensed-content-be-used-to-train-ai-it-depends/. Accessed 30 March 2022
Vishwakarma, R., Khwaja, H., Samant, V., Gaude, P., Gambhir, M., & Aswale, S. (2020). EEG Signals Analysis And Classification For BCI Systems: A Review. In *2020 International Conference on Emerging Trends in Information Technology and Engineering (ic-ETITE)* (pp. 1–6). Presented at the 2020 International Conference on Emerging Trends in Information Technology and Engineering (ic-ETITE). https://doi.org/10.1109/ic-ETITE47903.2020.066
Von Neumann, J. (1945). First Draft of a Report on the EDVAC. *IEEE Annals of the History of Computing*, *15*(4), 27–75.
Wagner, B. (2018). Ethics as an Escape from Regulation: From ethics-washing to ethics-shopping. In E. Bayamlioglu, I. Baraliuc, L. Janssens, & M. Hildebrandt (Eds.), *Being Profiled: Cogitas Ergo Sum: 10 Zears of Profiling the European Citizen*. Amsterdam: Amsterdam University Press.
Wallace, E., Zhao, T. Z., Feng, S., & Singh, S. (2021). Concealed Data Poisoning Attacks on NLP Models. *arXiv:2010.12563 [cs]*. http://arxiv.org/abs/2010.12563. Accessed 7 April 2022
Walther, J. B. (1996). Computer-Mediated Communication: Impersonal, Interpersonal, and Hyperpersonal Interaction. *Communication Research*, *23*(1), 3–43. https://doi.org/10.1177/009365096023001001
Walton, S., P, O. 'Kane, & Ruwhiu, D. (2019). Developing a theory of plausibility in scenario building: Designing plausible scenarios. *Futures*, *111*, 42–56.
Wang, Q., Li, R., Wang, Q., & Chen, S. (2021). Non-Fungible Token (NFT): Overview, Evaluation, Opportunities and Challenges. *arXiv:2105.07447 [cs]*. http://arxiv.org/abs/2105.07447. Accessed 16 February 2022
Wardle, H. (2021). Challenging "Play." In H. Wardle (Ed.), *Games Without Frontiers? Socio-historical Perspectives at the Gaming/Gambling Intersection* (pp. 79–101). Cham: Springer International Publishing. https://doi.org/10.1007/978-3-030-74910-1_4
Warnke, M. (2016). On the Spot: The Double Immersion of Virtual Reality. In *Immersion in the Visual Arts and Media* (pp. 204–213). Brill Rodopi.
Waterman, A. S. (2008). Reconsidering happiness: A eudaimonist's perspective. *The Journal of Positive Psychology*, *3*(4), 234–252.
Wegner, D. M. (2017). *The Illusion of Conscious Will, New Edition*. MIT Press.
Wehmeier, S., & Raaz, O. (2012). Transparency matters: The concept of organizational transparency in the academic discourse. *Public Relations Inquiry*, *1*(3), 337–366.
Weidinger, L., Mellor, J., Rauh, M., Griffin, C., Uesato, J., Huang, P.-S., et al. (2021). Ethical and social risks of harm from Language Models. *arXiv preprint arXiv:2112.04359*.
Weizenbaum, J. (1976). Computer power and human reason: From judgment to calculation.
Welbl, J., Glaese, A., Uesato, J., Dathathri, S., Mellor, J., Hendricks, L. A., et al. (2021). Challenges in Detoxifying Language Models. *arXiv:2109.07445 [cs]*. http://arxiv.org/abs/2109.07445. Accessed 31 March 2022
Wender, R., Hoffman, H. G., Hunner, H. H., Seibel, E. J., Patterson, D. R., & Sharar, S. R. (2009). Interactivity influences the magnitude of virtual reality analgesia. *Journal of cyber therapy and rehabilitation*, *2*(1), 27.
Westin, A. F. (1968). Privacy and freedom. *Washington and Lee Law Review*, *25*(1), 166.
Westphal, G., Garner, J., & McGirl, N. (2021). *Leveraging Industrial Mixed Reality Successes for Safeguards*. Oak Ridge National Lab.(ORNL), Oak Ridge, TN (United States).

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No.101006249.

Whalley, L. J. (1995). Ethical issues in the application of virtual reality to medicine. *Computers in biology and medicine*, *25*(2), 107–114.
Whitman, M. E. (2003). Enemy at the gate: threats to information security. *Communications of the ACM*, *46*(8), 91–95. https://doi.org/10.1145/859670.859675
Whyte, K. P. (2012). Now This! Indigenous Sovereignty, Political Obliviousness and Governance Models for SRM Research. *Ethics, Policy & Environment*, *15*(2), 172–87. https://doi.org/10.1080/21550085.2012.685570.
Whyte, K. P. (2017). Is It Colonial Déjà vu? Indigenous Peoples and Climate Injustice. In J. Adamson & M. Davis (Eds.), *Humanities for the Environment: Integrating Knowledge, Forging New Constellations of Practice* (pp. 88–105). London ; New York: Routledge.
Williams, B. (2006). *Ethics and the Limits of Philosophy*. Routledge.
Williamson, B. (2019). Brain Data: Scanning, Scraping and Sculpting the Plastic Learning Brain Through Neurotechnology. *Postdigital Science and Education*, *1*(1), 65–86. https://doi.org/10.1007/s42438-018-0008-5
Winner, L. (1980). Do artifacts have politics? *Daedalus*, 121–136.
Witmer, B. G., & Singer, M. J. (1998). Measuring presence in virtual environments: A presence questionnaire. *Presence*, *7*(3), 225–240.
Wolf, S. (1993). *Freedom within Reason*. Oxford University Press.
Wolf, S., & Schoeman, F. (1987). Sanity and the Metaphysics of Responsibility. *Ethical Theory: An Anthology*.
World Health Organization & Council for International Organizations of Medical Sciences. (2017). *International ethical guidelines for health-related research involving humans*. Geneva: CIOMS.
Wu, Y., Mo, J., Sui, L., Zhang, J., Hu, W., Zhang, C., et al. (2021). Deep Brain Stimulation in Treatment-Resistant Depression: A Systematic Review and Meta-Analysis on Efficacy and Safety. *Frontiers in neuroscience*, *15*, 655412. https://doi.org/10.3389/fnins.2021.655412
Yee, N., & Bailenson, J. (2007). The Proteus effect: The effect of transformed self-representation on behavior. *Human communication research*, *33*(3), 271–290.
Yee, N., Bailenson, J. N., & Ducheneaut, N. (2009). The Proteus effect: Implications of transformed digital self-representation on online and offline behavior. *Communication Research*, *36*(2), 285–312.
Yeung, A. W. K., Tosevska, A., Klager, E., Eibensteiner, F., Laxar, D., Stoyanov, J., et al. (2021). Virtual and augmented reality applications in medicine: analysis of the scientific literature. *Journal of medical internet research*, *23*(2), e25499.
Yoshida, S., Teshima, T., Kuribayashi-Shigetomi, K., & Takeuchi, S. (2016). Mobile Microplates for Morphological Control and Assembly of Individual Neural Cells. *Advanced Healthcare Materials*, *5*(4), 415–420. https://doi.org/10.1002/adhm.201500782
Zagal, J. P., & Mateas, M. (2010). Time in Video Games: A Survey and Analysis. *Simulation & Gaming*, *41*(6), 844–868. https://doi.org/10.1177/1046878110375594
Zeide, E. (2017). *Unpacking Student Privacy* (SSRN Scholarly Paper No. 2970767). Rochester, NY: Social Science Research Network. https://papers.ssrn.com/abstract=2970767. Accessed 7 April 2022
Zhai, X., Asmi, F., Yuan, J., Anwar, M. A., Siddiquei, N. L., Ahmad, I., & Zhou, R. (2021). The Role of Motivation and Desire in Explaining Students' VR Games Addiction: A Cognitive-Behavioral Perspective. *Mathematical Problems in Engineering*, *2021*, e5526046. https://doi.org/10.1155/2021/5526046
Zhou, M. X., Mark, G., Li, J., & Yang, H. (2019). Trusting Virtual Agents: The Effect of Personality. *ACM Transactions on Interactive Intelligent Systems*, *9*(2–3), 10:1-10:36. https://doi.org/10.1145/3232077
Zuk, P., & Lázaro-Muñoz, G. (2021). DBS and Autonomy: Clarifying the Role of Theoretical Neuroethics. *Neuroethics*, *14*, 83–93. https://doi.org/10.1007/s12152-019-09417-4

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No.101006249.

Annex: First wave of TechEthos WP2 interviews

The following is an edited transcript of answers provided by the TechEthos Experts (denoted A to H) during the first way of WP2 interviews. The interviews were conducted and analyzed by DMU in Spring 2022.

As a result of technological innovation in the area of (technology family) how do think the world will change by 2045?

A) This area would become more invasive as a result of technological innovation. The areas of functional and cognitive neuro technology. There will be an enhancement of hybridisation between the human and technology but this will mainly be cognitive. It's a hybridisation between humans and neurotech. Functional hybridisation but also physical hybridisation

B) Do not know - any predictions made need to be cautious as it cannot be foreseen and it is not easy to make specific predictions, although scenario building can help.

C) By 2045 it will be an increase in technology in the world and large amounts of data available. The advantages would probably be in medicine where there will be increased access to patient data history, where the doctors can use algorithms and AI to predict or suggest personalised treatment. Essentially the computer could aid your recovery because it could do a patient centred treatment and analysis of the exactly what the problem is and this is presented to the clinician. However, there could also be stereotyping (which opens humans to abuse for example in insurance purposes) and over-reliant use of algorithms which can lead to mis-diagnose diseases. It is important to build trust and ensure privacy of the patient. This is a double face problem, as we need to be patient centred and respect patient goals i.e. consent (i.e. who will have access to their data and who will use it), so there is infringement of freedom and ethics.

D) There will be an increased convergence towards emerging technologies, neuroscience/neurotechnology with the use of AI & biology may progress towards maturity, for example in year the 2000 to make a human genome was a matter of months, but to make a human genome is today a matter of a few hours. In 2045 it could be a matter of two minutes.

E) There will an optimisation of algorithm and AI (with increased convergence and maturity), leading to increase speed in the analysis of brain activity, privacy and autonomy (which already exists) at decision making might be more powerful. Uptake has been slow to start with but overtime there could be Increased amount of research and funding into immersive technologies, and feels there will be more of a presence in the workplace for this type of technology, it could even replace mobile phone.

F) It is difficult to forecast; AI systems will reach maturity at some stage and the availability for AI tools for SEMs will increase.

G) There will be an increased difference between countries and how this is managed will depend on the compatibility with ethical values, if we leave the technology firms to it then this will lead to a decrease in social justice. Being optimistic, hopefully the European Commission can steer technology firms so that they are useful and for everyone (inclusive) i.e. they will benefit everyone and not just the super-rich. For addressing societal problems like climate change, energy efficiency, social justice, better access to education for people from different backgrounds. I also see a great risk, especially in those countries where either human rights do not count anyway, or they're corporation, they're just too powerful that they will innovate in areas where they can make profits. Which will mean serving the interests of groups with purchasing power and neglecting any other values. Basically, I see the risk of a huge divergences there, and that will then of course also play out in geopolitical terms and so on.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No.101006249.

H) I think we cannot really imagine the world of 2045 because the climate conditions will be so fundamentally different from today. I do not think we can fully imagine what [the world] will do culturally. So that part will be very hard to determine. I think it's fairly clear that through technological innovation but also the dreams and sort of cultural aspirations, that carbon capture and storage will become a massive industry. I'm afraid that solar geoengineering will also become a major part of the conversation even though I might not want it to.

In your view, what do you think are the benefits associated with this technology by 2045?

A) The ability to facilitate the cognitive effect is a benefit. In addition, further research in this allows for a better exploration of the cognitive potential but feels that there is great benefit from an education perspective for example an improvement in educational practices and protocols. For example, children with learning disabilities to benefit, if there is an increase in research around cognitive development this will help their learning and disability.

B) Do not know and prefers not to comment - there could be benefits.

C) I highlighted some in previous question but there will be an extension to the capability in patient care and lots of potential in terms of extending human capability and increase in efficiency. We have several monitoring systems that normally on unknowingly are deployed in our world for example Alexa or Apple products that record what you do, what you say and so we are constantly monitored. Overall there is an infringement of the freedom of people and so this is a huge problem and an ethical problem too. Even if we give the consent as we do at a moment for GDPR, I do not know who is using my data.

D) In terms of health - for example patients with Parkinson disease, we hope that there could be less invasive techniques. Education of children with disabilities for example those with dyslexia, dyscalculia and autistism technology could help them improve their skills and allow them more autonomy. However, the more intrusive these technologies are there is a potential of interference on our thoughts and our behaviours. Therefore, this area needs to be tightly controlled in terms of who is collecting the data and has access to the data, the transparency of algorithms, accountability and collection of data.

E) Benefits include in the areas of medical applications for example pain mitigation during surgery, phobias, education using virtual worlds, treating psychopathology.

F) There is a potential for economic growth, the amount of AI tools available for innovation and intervention will benefit the area of medicine and drug discovery, there could be improvements in environment, poverty

G) It depends very much on where you use technology, but I think there are huge potentials not only in in the area of health i.e. from a public health perspective. A view on how can these technologies be developed such that they serve all members of society that can benefit from them and try to not reinforce social injustice, but rather try to address it. new possibilities of communication and information sharing have huge potential of empowering individuals to really lead their own lives to join forces with others to run things bottom up in society. There is a huge potential for democratising various processes, including in the realm of work. [Politics and also ethical decision making at the political level is important].

H) That really depends on what you define as geoengineering or climate engineering. If we manage to get a carbon capture and storage system up and running, that is both fair and just as effective, (all three of those are very debatable) then it may help us reach climate targets. By 2045, I do not anticipate us using SRM on any major scale, but maybe some climate engineering technologies will be used to cool regionally and could potentially provide the benefit of less intense heat waves or cooling specific regions for specific reasons.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No.101006249.

Can you anticipate what risks and harms might arise?

A) The disparities and inequalities around the world will increase. Technologies are not democratic (the forefront of technologies are not democratic) - This is much dependant on economic and financial power on a macro level (considering the economy as a whole) and micro level (considering communities and individual). There is economic disparity at different levels for countries and this could get worse.

B) Direct air capture - can involve costs and affect political priorities. For Ocean based - changes to ocean chemistry and biodiversity, changes to livelihood and land use.

C) Privacy, and this can be used as manipulation which steers humans in a particular direction to act like puppets. There is freedom of information but once that information is released then you do not know what the future use of that data will be? The harms is that we (humans) are machines and are slaves of whatever power decides how to utilise my data, to manipulate what I do, because you have seen already, how they have tampered with the US election, and opinions in the UK about Brexit. Social interactions have now created a fake world and anti-social behaviour and a decrease in social interaction - a distortion of what the real world is about for example sexting, bullying, aggression, and a loss of manners. Furthermore, you can probably work from home and never see anybody and never be with anybody, and so we are becoming antisocial.

D) Privacy - for example how data will be kept secret as this can determine decision making, ensuring no interference and how to retain autonomy. Risks of discrimination of access to treatment if it is expensive. Application of technology to children and teenagers whose brains are still developing - could this lead to brain interference. Therefore we clearly need to oversee the use and the impact of these technologies.

E) Privacy of personal information since overtime this will lead to sharing of even more personal information. For example, having an avatar of ourselves in real time we will learn more about ourselves/our micro-movements but it could lead to opportunities in intervening or manipulation or nudging users to behave in a certain way. In terms of children - we do not know what the technology does to developing minds. Neglect of the physical environment, neglecting one's biological body. For example, for massively multiplayer online roles - playing games where people feel more comfortable in the virtual world as opposed to the real world. An illusion is created which results in a loss of connection one example is when COVID-19 happened.

F) Inequality amongst countries can cause a dynamic effect, there could be risks to certain jobs which will create a destructive effect. Issues around transparency and explainability, AI mis-use for example in the area of warfare. The over-reliance of AI technology could be at the detriment of doing things fairly. In addition, bias and further risk of erosion to society values.

G) The big challenges are going to be using these new technologies in ways that help address climate change and improve energy efficiency, because what is often not part of the discussion, in my observation is that lots of these new technologies actually require a lot of energy. So when you talk about replacing all kinds of jobs by robots, no one talks about the energy for these robots. One of the big questions I think is going to be how to organise the division of Labour between humans and technologies and what kind of energy supply this will require. In terms of countries, US innovation is very much driven by companies in order to make profits. China uses technology for state purposes and surveillance. In Europe there is this ideal that it's somehow meant to serve society, but it is quite hard to do this and also how to organise this in terms of institutions. If [technology] is in the hands of more state powers without any counter power then then it can just make dictatorships. This can lead to increase surveillance and decrease spaces for personal interaction. At least at the moment, there seems to be relatively little attention to the long-term effects on children being exposed to new technologies, there is a lot of enthusiasm about introducing digital technologies in schools. For

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No.101006249.

example, without really thinking through what it means for someone's education and development, physical, psychological and social dynamics.

H) The major risk is that all of these technologies distract from actual climate policy or actual mitigation policies, we have already seen that to some extent happen. With this assumption that we can do carbon capture on a massive scale. We have had all kinds of climate scenarios projecting into the future, and CO2 capturing on the scale that we do not know is possible at the moment. Actually, most people say that it probably is not possible, but still our climate policy is based on those assumptions and it's still allows us or it allows our politicians and our policymakers to postpone the very hard cuts that need to be made. Another aspect of justice around carbon capture and storage, and especially what we call natural solutions as non-based natural solutions, is that these take an awful lot of land and this may create land pressures for people who are now using it for their own agriculture for their own subsistence. Another form of what people would call neo-colonialism where we compensate for our own emissions by pushing people off their lands far away. These technologies and technology engineering particularly bring a lot of risks, such as technical risks, geopolitical risk. They may lead to tensions and even war.

Who are the main beneficiaries of this [technology family]? And who will be excluded in your view?

A) The main beneficiaries are those with economic power, they will benefit the most from this technology. Also culture plays an important role as to how these technologies will be implemented and play an important role in our lives. However, this type of technology will not be available to everyone in particular those from a less economically developed countries. [Economy and Cultural are the main determinants of the beneficiaries and those who will be excluded].

B) Anyone and everyone - politics which addresses climate change - concept of 'green washing'. Other disadvantages for indigenous populations. (Reference to SCoPEx : Stratospheric Controlled Perturbation Experiment).

C) Beneficiaries are the pharmaceutical companies from a sales of medicine point of view, the state and cyberspace via twitter, and for everyone else there's a positive and negative to everything, it's a double edge sword but it is important to ensure regulation and ethics from a global perspective. The basic rules must be followed due to consequences of information misuse.

D) Hard to say but hoping that patients will be beneficiaries possibly those with brain disorders (1/3 of expenditure is in the field of health). So, beneficiaries could be people with Alzheimer disease as well as people with the autistic spectrum disorder.

E) Classically research & innovation is expensive so who can have access to the technology is based on money, distribution and remoteness of populations - these people are then left vulnerable and [hence seen as excluded]. The COVID-19 vaccine was a perfect illustration of the challenge for innovation. The cost but also the amount of production, and the possibility of distribution and access to the people that really need them.

F) Beneficiaries can include venture capitalist, medical patients, those seeking help in health & fitness, those excluded in a sense of being victimised using the technology as a tool for oppression. Exclusion in the sense of someone not having access to immersive technology that's not too bad but being coerced into using it is say at work by an employer can lead to are gives you depersonalisation symptoms.

G) Beneficiaries include society at large, populations as a whole, the area of medical innovation. Although the technology can be expensive so there will be a equality gap as to who can have access to it and at an international level. [Artificial intelligence requires vast amounts of computing power, data, and expertise to train and deploy the massive machine learning

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No.101006249.

models behind the most advanced research. But access is increasingly out of reach for most colleges and universities. [A National Research Cloud (NRC) would provide academic and non-profit researchers with the compute power and government datasets needed for education and research at Stanford University].

H) The COVID-19 vaccine illustrated this, we have the technological capacity to make such innovation, but then we lack the social innovations and the social institutions for getting these innovations to everyone. There are many other examples that people in poorer countries or poorer layers in even richer countries tend to be left out and not benefit so much. Groups that are disadvantaged along all kinds of lines like race, gender, and class in richer countries. It is such a broad set of technologies, it's very hard to pinpoint single groups. Specifically, when we're talking about land use, the main beneficiaries have historically been large companies who have been using land grabbing techniques to make sure that they own land that was previously owned by community or was used by smaller communities. In addition, especially around carbon capture and storage and with the economical setup that we have now where carbon will be priced. The main beneficiaries will be a whole new economic sector that needs to be built around this. In terms of the solar geoengineering side, it is hard to tell who the beneficiaries would be. But again, it's probably people. historically, all of these technologies have been used militarily or by major powers. I think a lot of inconvenient voices will be excluded. What we see in the policy process and about in all of these processes is that you need to be able to speak a certain language and need to phrase your concerns in a certain way. And if it doesn't fit into the dominant way of thinking about approaching problems.

Considering the global interest in the issue of ethics what do you predict to be the ethical issues that could arise by 2045?

A) Not sure whether there will be new ethical issues that will arise, perhaps the same ethical concerns will be present however further exploration and study will go into the 'fake use of ethics' - i.e. 'ethics washing' for example companies and industry will try to give an importance to ethics by giving this more attention - as a way of being perceived to care and be interested but in reality, this will be fake. This concept will grow more in the future.

B) Seems no new ethical issues will arise but perhaps the old ones expressed in a new way. Potential issues, such as resources (nature and otherwise) moral hazard, land use which impacts biodiversity, IP patents, funding, (reference to hubris).

C) Apart from the ones mentioned in the previous questions, there is a need to agree on basic rules and there needs to be a meeting of cultures otherwise there will be a loss of identification and variety, and a loss of cultural. We will end up as human machines so it is important to retain diversity and innovation in order to keep an individual's identity.

D) Equal moral worth - we are humans so should be treated equally such that no one is excluded. There is an increase in inequality (justice), take the COVID-19 example in the previous question and it could be worse for technology. [Other ethical issues such as privacy has been mentioned previously in Qus].

E) Already mentioned some in previous questions, but the right to one's likeness. but generating an avatar opens questions about digital property - there are some countries that do have laws about protecting, but because the servers are international, so this is not a trivial problem to solve. What about the right to be 'offline' as society we are overwhelmed by expectations our entire lives, so right to be offline is good for human flourishing.

F) Ethical issues will depend on how advance the AI system is and how it is being used. Also, the relationship between government and the private sector (which is more advanced) is important, the power of companies will play a role so focus should be on the political system and finding the right policy. Other ethical issues include labour displacement and ensuring that machine learning values are aligned with human values.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No.101006249.

G) One thing is what I already mentioned that lots of technologies that are truly beneficial will simply not reach people in poorer countries. Questions arise about economic development and global justice. New forms with new technologies is that you use people in poorer countries for studies like Guinea pigs for new technologies to see what it does, an example of this was with the Covid-19 vaccine where studies were being carried out in India. In terms of climate engineering, there is a huge question around the risks and benefits of these be evaluated in ways it really takes the interests of people in poorer countries or in coastal areas into account?

H) I think the main discussion for at least the next 10 years and then probably going forward will be environmental justice or climate justice. Can we here in the north keep living the lifestyles that we are living? That will be one of the major issues, but also who suffers from our lifestyles.

Do you think we have gone past the point of reversibility & irreversibility of this technology? And please explain why?

A) Yes, we have gone past irreversibility because we are in a new evolution phase of human kind. The advancement of technology is very different from what it was 10 years ago and we cannot go back. For instance, smartphone play it is so crucial role in our life that I do not think we can get rid of this for evolutionary reasons. We are functionally already hybrid with this technology. So we cannot really be separated from it anymore.

B) N/A

C) No - it has just started and there is a need for data curation, things are still evolving.

D) I do not know what will happen next, it depends on the convenience, accessibility and usefulness with respect to health. The technology where there is clearly a technological addiction because of its usefulness will likely remain. If tomorrow the technology appears not useful or even detrimental, you will have a rejection of the technology. I am not really afraid by the technology itself I'm afraid by the human beings that are using the technology and the recent images from Ukraine just illustrate that it's not a matter of technology, it's a matter of human beings or people just forgetting that they are human beings.

E) Immersive technology is not widely accepted yet, its interesting because there is funding and interest in the development of this technology but the consumer excitement is not quite there. Historical trends show that development of technology does not stop unless there's a reversal of attitudes.

F) Yes - there is alot of AGI (artificial general intelligence) research although we are still in the early days it is important that government policy mechanisms are right so they can nudge towards society benefits.

G) I do not think its technology in itself that is reversible. It's always a sort of constellation of technology and social structures and power relations that can create certain path dependencies that are very hard to revert. The kind of socio technological barrier that these companies have become so powerful that they might just have too much impact and that's what makes it irreversible.

H) I would say definitely for carbon capture as an idea and as a set of technologies. We have definitely past the point of irreversibility. It's been embedded in the so many of our climate policies at the moment, and the dominant narrative around it is that is just needed. Then we cannot do without and economic system is being built around it through a lot of investment. The conversation is really heating up, so I do not think we are going back on any of that as a sort of overarching category. On solar radiation management. We are definitely not past the point of irreversibility, and we definitely also shouldn't be. It's still very contentious whether or not we should even do research on this.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No.101006249.

Is there anything else you would to add which we have not covered already?

A) A new topic at the moment is Brain inspired AI. There has been struggle with this kind of question, whether, for instance, brain inspire AI raises new kind of ethical issues or traditional ethical issues, and we still do not have a final answer to that, so we are unable to say whether there will be new kind of ethical issues. Because after all, humankind will be always the same. Fundamentally speaking, ethics will continue to be what it is already today. However, levels of ethical problems will depend on the economics (and culture) of the country. I feel that philosophy will be at the forefront of a fair ethical analysis moving forward.

B) N/A

C) Always think positive and do positive with a good intention and try to help others, this will lead to the better good of society, change rules - change mindset - it should be globally unacceptable to misuse data. Time for re-education in new ways that are ethically and morally viable.

D) Brain surveillance is future - how things are governed will depend of whether we will in a democratic society or a dictatorship society. We cannot disconnect science and technology from the from the economy at large. Neoliberalism is dominating the economy and that anything you can sell anything you can market is something that will gain support.

E) N/A

F) It is very difficult to predict these things because so much that will also depend on the potential breakthroughs that we might have in terms of the safety and ethics in the next 10 years so you know to the extent that you solve things like explainability and you know being able to reverse engineer AI outputs and kind of improve on robustness and alignment.

G) I would like to add, which is the need for an intersectional and feminist perspective. Technology is mostly developed by men for men and that's one of the ways in which biases. If the power structures are still very much dominated by patriarchy, on example is working from home and digital work. Which on the one hand is great as it opens up new possibilities, but I see a real risk that it will enable one and without caring responsibilities to marginalise women who work more from home because they have family responsibilities in the workplace.

H) N/A

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No.101006249.