TECHETHOS

FUTURE ○ TECHNOLOGY ○ ETHICS

Suggestions for the revision of existing operational guidelines for climate engineering, neurotechnologies and digital XR technologies

Deliverable 5.3

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.

WP5 (D5.3)

Work Package	WP5		
Lead Partner	De Montfort University, DMU		
Author(s)	Sara Cannizzaro (DMU), Laurence Brooks (DMU), Kathleen Richardson (DMU), Nitika Bhalla (DMU), Bennet Francis (UT), Dominic Lenzi (UT)		
Contributor(s)	Laurynas Adomaitis (CEA), Steven Umbrello (TUD)		
Acknowledgements	Michel Bourban, Andy Parker, Matthias Honegger, Marie-Valentine Florin, Maria Rementeria, David Winickoff, Thomas Stieglitz, Laura Kreiling, Douglas Robinson, Philip Brey, Alexei Grinbaum		
Due date	30th September 2023		
Submitted date			
Version number	2	Status	draft

Project Information

Grant Agreement 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

The information, documentation and figures in this deliverable were produced by the TechEthos project consortium under EC grant agreement 101006249 and do not necessarily reflect the views of the European Commission. The European Commission is not liable for any use that may be made of the information contained herein.

Dissemination Level	
PU: Public	☒
PP: Restricted to other programme participants (including the European Commission)	☐
RE: Restricted to a group specified by the consortium (including the European Commission)	☐
CO: Confidential, only for members of the consortium (including the European Commission)	☐

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

Quality Control

Reviewed by:	Review date:
Laurynas Adomaitis (CEA)	02/10/23
Pieter Vermaas (TUD)	23/09/23

Revision history

Version	Date	Description
1.0	15/06/23	Converted to native Google doc
2.0	20/07/23	Version for sharing with ADIM board

Keywords

Operational guidelines, ethical issues, climate engineering, neurotechnologies, digital extended reality

How to cite

If you are using this document in your own writing, our preferred citation is:

Cannizzaro, S., Bhalla, N., Brooks, L., Richardson, K., Francis, B. and Lenzi, D. (2023), *TechEthos Deliverable D5.3: Suggestions for the revision of existing operational guidelines for climate engineering, neurotechnologies and digital XR technologies. Available at* *www.techethos.eu*.

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

The TechEthos Project

Short project summary

TechEthos is an EU-funded project that deals with the ethics of the new and emerging technologies anticipated to have high socio-economic impact. The project involves ten scientific partners and six science engagement organisations and runs from January 2021 to the end of 2023.

TechEthos aims to facilitate "ethics by design", namely, to bring ethical and societal values into the design and development of new and emerging technologies from the very beginning of the process. Technologies covered are "climate engineering", "digital extended reality" and "neuro-technologies". The project will produce operational ethics guidelines for these technologies for users such as researchers, research ethics committees and policy makers. To reconcile the needs of research and innovation and the concerns of society, the project will explore the awareness, acceptance and aspirations of academia, industry and the general public alike and reflect them in the guidelines.

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.

Table of contents

Short project summary 4

Executive Summary 9

1. Introduction 10

1.1 Background 10

1.2 Structure of the report 11

2. Methodology for guidelines improvement 12

Definitions of ethics guidelines/frameworks/codes 16

3. Proposed Guideline Improvements 19

3.1 Climate Engineering 19

3.1.1 Introduction 19

3.1.2 Selected guidelines for Climate Engineering (CE) 25

Original Tollgate Principles for SRM 25

Original ABCs of Carbon Dioxide Removal (CDR) 26

3.1.3 CE guidelines' gaps 27

3.1.4 Proposed improvements to Guidelines (CE): 28

Proposed improvement to The Tollgate Principles for SRM 29

Addendum to CDR Guideline: Policy relevance of normative approach 37

Proposed improvements for ABCs of Carbon Dioxide Removal 40

3.2 Digital Extended Reality 50

3.2.1 Introduction 50

3.2.2 Selected guidelines for dXR 50

Original ALTAI Guidelines (shortened) 50

Original Microsoft Guidelines for Human-AI Interaction 55

3.2.3 dXR guidelines' gaps 58

3.2.4 Proposed improvements to guidelines (XR) 60

Refined ALTAI Guidelines 60

Refined Microsoft Guidelines for Human-AI Interaction 62

3.3 Neurotechnology 67

3.3.1 Introduction 67

3.3.2 Selected Guidelines for Neurotechnologies 68

Original OECD Recommendation on Responsible Innovation in Neurotechnologies 68

Original Neuroethics Guiding Principles for the NIH BRAIN Initiative 71

3.3.3 Neurotechnologies guidelines' gaps 74

3.3.4 Proposed improvements to Guidelines (NT) 76

Revised OECD Guidelines on Responsible Innovation in Neurotechnologies 77

Revised Neuroethics Guiding Principles for the NIH BRAIN Project 82

4 Discussion 83

5 References 88

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

List of tables

Table 1: List of Definitions

Table 2: List of Abbreviations

Table 3: Original guidelines revised for each technology family

Table 4: Refined Tollgate principles for SRM

Table 5: Refined ABCs principle guidelines for CDR

Table 6: AI Design Guidelines and their applications

Table 7: Refined ALTAI guidelines for XR

Table 8: Refined Microsoft Guidelines for Human-AI Interaction, for XR

Table 9: Refined OECD Guidelines on Responsible Innovation in Neurotechnologies

Table 10: Refined Neuroethics Guiding Principles for the NIH BRAIN Project

List of figures

Figure 1. Mindmap of operational guidelines, frameworks and codes for Climate Engineering

Figure 2. Mindmap of operational guidelines, frameworks and codes for Digital Extended Reality

Figure 3. Mindmap of operational guidelines, frameworks and codes for Neurotechnologies

Figure 4: Section of Miro board which we used during the CE guidelines development/refinement workshop (22 February 2023, DMU) to lay out gaps next to TechEthos findings.

Figure 5: Innovation ecosystem of climate engineering technologies.

Figure 6: Microsoft Guidelines for Human-AI Interaction

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

Definitions and abbreviations

Term	Explanation
Climate Engineering	Climate engineering is a family of technologies that enables the modification of natural processes and human activities looking to address and mitigate climate change locally and globally.
Digital Extended Reality	Digital Extended Reality refers to AI-powered digital technologies (hardware and software) capable of perceiving and processing human sensorial outputs, e.g., voice, gestures, language, movement, emotions and other elements of human communication, as well as responding to these types of signals by creating an extended visual, audio, linguistic or haptic digital environment for users.
Neurotechnologies	Neurotechnologies are technologies that aim at affecting and emulating human-brain capabilities and functions through artificial replacements or add-ons in a two-way interaction between the brain and the external environment or systems.

Table 1: List of Definitions

Term	Explanation
ADIM Board	Advisory and Impact Board
AI	Artificial Intelligence
ALTAI	Assessment List for Trustworth Artificial Intelligence
CCS	Carbon Capture and Storage
CDR	Carbon Dioxide Removal
CE	Climate Engineering
ChatGPT	Chat Generative Pre-Trained Transformer
ESG	Environmental, Social and Governance
EU	European Union
dXR	Digital Extended Reality

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

Term	Explanation
HIC	Human-in-Command
HITL	Human-in-the-loop
HOTL	Human-on-the-loop
LLM	Large Language Model
ML	Machine Learning
NIH	National Institute of Health
NT	Neurotechnology
OECD	Organisation for Economic Co-operation and Development
R&I	Research and Innovation
SAI	Stratospheric Aerosol Injection
SRM	Solar Radiation Management
TEAeM	TechEthos Anticipatory Ethics Matrix
UNFCCC	United Nations Framework Convention on Climate Change
VR	Virtual Reality
WP	Work Package
XR	Extended Reality

Table 2: List of Abbreviations

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

Executive Summary

This report presents reflections on existing guidelines and proposes improvements to existing ethical guidelines based on work carried out in WPs 2, 3 and 5 (T5.1). This report explores the needs and gaps in current guidelines in order to reflect on and make suggestions for improvements to them for selected technologies. Drawing on ethics by design, this report incorporates findings from the stakeholder activities including the underrepresented group (D3.5) and expert interviews (D2.2).

This report builds on WP1 D1.1 (Technology Families) and the consortium selected technology families. These are:

- Climate Engineering Technologies
- Digital Extended Reality
- Neurotechnologies

Specifically we have explored the gaps in current operational ethical guidelines. The report discusses the potential improvements to selected guidelines for each technology family using the ethics by design approach, while taking into account the expectations of different stakeholder groups.

The proposed improvements to ethical guidelines is based on (i) desk analysis, taking advantage of existing ethical guidelines, policy, industry and non-governmental organisations and governmental at international, EU and national levels (ii) search documents with relevant keywords (iii) an adapted mapping analysis approach enhanced by expert consultations, (iv) incorporated findings from stakeholder engagements and (v) expert interviews and consultation on refinement.

The TechEthos proposals for improvements to ethical guidelines is drawn from a novel approach in grouping and clustering families of technologies, based on the functions, applications, ethical and societal challenges, and the identification of criteria for assessing potential socio-economic impacts of these technology families.

While there is no universal ethical guidance across the three TechEthos technology families and beyond, we have synthesised a set of key recommendations that can be used for proposed improvements to guidelines:

- **Bespoke governance/institutional infrastructures** - relevant administrative bodies to ensure the guidelines are properly applied, training and support in how to interpret and use the guidelines
- **Diverse stakeholder participation** - enable engagement with broadest range of stakeholders, including co-creation, co-decision making
- **Impact** - testing the efficacy of the outcomes, from use of the guidelines, with real-world examples
- **Inter-sector skills and knowledge exchange** - institutionalise cooperation between technology providers and policy makers
- **Responsibility to the future** - responsible forecasting, ethical defensibility, sustainability
- **Social and communicative awareness** - enable the developers and technologists to be socially aware, for example in terms of making language more accessible and gaining feedback

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 Background

This report is the culmination of an interdisciplinary collaboration involving academics, experts and the wider community carried out in person and online, by the authors of this deliverable. The task (5.2) was to reflect on existing guidelines and make suggestions for improving these operational guidelines for Research and Innovation (R&I) for our three selected technology families: 1) Climate Engineering 2) Digital Extended Reality and 3) Neurotechnologies. In approaching this task, a number of criteria were considered in relation to the systematising of decisions-tree process including identifying the range of frameworks, guidelines and codes identified in previous tasks (D2.1) and assess their appropriateness in relation to the ethical issues arising from the technological families' issues.

Several themes identified and discussed in D5.3 are worth noting. While several frameworks, codes and guidelines do exist tangentially connected to the three technology families, the vast majority have no statutory or legal regulatory status. Hence, they are developed by independent experts - sometimes companies, as with the case of the Microsoft Guidelines for Human-AI Interaction (Microsoft, 2017), or as the product of research groups, such as the Oxford Principles and Tollgate Principles for SRM (Rayner et al. 2013; Gardiner & Fragnière, 2018). Guidelines acquire status on the basis of other factors, such as reference to them by state and governmental actors, or respected independent bodies. There can also be the thorny subject of scope and generality. Many guidelines are tailored for specific problems and thus are non transferable to other areas, even if they fall in the remit of their technology family. A case in point would be the application of the Oxford Principles (Rayner et al., 2013) which were developed in response to the Royal Society's 2009 report (Shepherd et al., 2009, p. 17) on geoengineering, and thus their applicability to particular domains of climate engineering and to adjacent technology fields is underspecified. Ethical guidelines, as particular types of cultural artefacts, are developed usually in response to something. Subsequently, as TechEthos reflects on these guidelines, the transferability or usefulness of them beyond a niche technological arena is something we have tried to avoid. Needless to say this is not a failsafe task, as unique variabilities between emerging technologies, even within the same family, must allow for flexibility, relevance and robustness.

In preparing this document, we consulted with experts to identify *gaps* in existing guidelines and to contribute, where possible, to improving a guideline's principle. For example, in the ABC's guidelines - principle (e) *Policy mixes should include international cooperation to improve CDR efficiency,* our experts identified gaps relevant to it, and developed the meaning to add more nuance to the principle *'Cooperation amongst stakeholders and wider reach to increase desirability of the technology'.* Also, this could include collaboration of ethics and policy making, for the purpose of merging governance and ethics.

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

The TechEthos approach was to build on the methodology advanced in the Sienna-SHERPA projects, and use these as the building blocks for refining ethical guidelines in these areas. Taking as a starting point existing guidelines in the three technology families, and requesting an additional three guidelines from ADIM Board experts. This was followed by creating a rubric listing each item with the given explanation from its original authors. This was then supplemented by further information from TechEthos consortium members and ADIM board experts. Moreover, these additions were enriched by drawing on the findings from WP3 which incorporated the findings from the linked third party consultations with under-represented groups. We wanted to test the efficacy of the process by identifying two guidelines, and following the procedure for each item. Moving beyond the limits of prescription to operationalisation we drew on the Sienna project's deliverable D6.3 *Methods for translating ethical analysis into instruments for the ethical development and deployment of emerging technologies* (p.15 this volume has further explanation of the method). In addition to a set of guidelines, an assistive administrative structure is recommended including the development of new roles and responsibilities to assist organisations in how to implement ethics by design. Of course, such an approach puts additional costs and requirements on actors involved in developing new and emerging technologies to "shadow" the work of professionals at key stages of the process: from inception of the idea (what is its value, who will it help, hinder or harm) to long term considerations about social and psychological welfare of future users or directly and indirectly-affected stakeholders.

The two existing operational guidelines for each of the technology families selected are:

Technology Family	**Original Guideline**
Climate Engineering	Tollgate Principles for SRM (Gardiner & Fragnière, 2018)
	ABCs of Carbon Dioxide Removal (CDR) (Honegger et al., 2022)
digital Extended Reality	ALTAI Guidelines (European Commission. Directorate General for Communications Networks, Content and Technology., 2020)
	Microsoft Guidelines for Human-AI Interaction (https://www.microsoft.com/en-us/research/project/guidelines-for-human-ai-interaction/)
Neurotechnology	OECD Recommendation on Responsible Innovation in Neurotechnologies (OECD, 2019)
	Neuroethics Guiding Principles for the NIH BRAIN Initiative (Greely et al., 2018)

Table 3: Original guidelines revised for each technology family

1.2 Structure of the report

The structure of the report is as follows: this deliverable (D5.3) initially presents the background to the TechEthos T5.2 task concerning the development/refinement of operational ethical guidelines and codes of conduct for R&I in the identified technology family

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

and with relation to research integrity. This report then outlines the methodology for suggesting improvements of existing guidelines, which starts by a) providing an overview of existing operational guidelines for each technology family, including the results for the guidelines/codes/frameworks scanning exercise we carried out for T2.1 and included in D2.2 enriched by discussion with experts; b) identifying gaps in current operational guidelines, frameworks and codes of conduct for each technology family. For Climate Engineering, we have selected the Tollgate principles (applicable to Solar Radiation Management) and ABCs Carbon Dioxide Removal governance principles for CDR). For Digital extended reality, we have included the ALTAI guidelines and Microsoft guidelines for human-AI interaction. Finally for Neurotechnologies we selected the NIH brain initiative and the OECD recommendation on Responsible Innovation in Neurotechnologies; c) reviewing each of these current operational guidelines in consultation with internal and external experts in combination with empirical results from the TechEthos project to identify gaps; d) identifying where the current guidelines can be improved with respect to new and emerging technologies. Finally this report presents suggestions concerning the refinement of operational ethical guidelines based on the outlined methodology. It finally indicates the implications of this guidelines enhancement for each of the technology families. This deliverable will help to support the research community including academia, industry, policy makers and society (such as under-represented groups) to refer to the refined operational guidelines enhanced by the TechEthos results for R&I surrounding the ethical development and potentially, deployment of each of the technology families.

2. Methodology for guidelines improvement

As part of the methodology for proposing improvements to guidelines, we prepared the 'T5.2 - Approach document - Preparing mindmaps of operational guidelines, code and frameworks with respect to Neurotechnology or Extended digital reality (with a focus on ethics)'. We used the methodology contained in this document to set up the first part of the process to develop/refine guidelines for one technology family, namely climate engineering (CE). We then distributed the document to the other partners involved in developing/refining the guidelines for the other technology families i.e. neurotechnologies (NT) and digital XR. The T5.2 approach document reads as follows:

1) **Identifying existing guidelines, codes and frameworks:**

- Start with deliverable D2.1 and locate the existing guidelines, codes and frameworks for your allocated technology family.
- Identify one or two experts (possibly from within the project or ADIM board) within your technology family to review the existing selection of guidelines, codes and frameworks, and then with their expertise identify at least 3 additional guidelines, codes and frameworks. This is to ensure we have a more comprehensive set.
- Build a mind map (or use a software of choice) of your selected technology family - please see the link to the CE mindmap as an example. The mindmap can be built including title and reference to the code/guideline/framework being reviewed.
- We then sought to **identify gaps with regards to ethical issues in the current operational guidelines, codes and frameworks, such that the ethical issues and gaps could be integrated using the TechEthos project results**. We devised the

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

following methodology which we then distributed to the other partners involved in developing/refining the guidelines for the other technology families i.e. neurotechnologies and digital XR.

- With the help of your experts (can be the same ones from before, or others, but at least need to be in the relevant field), **select (one or two) operational guidelines, code or framework** that has ethical gaps and therefore can be refined.
- We suggest you organise a **meeting** where you can discuss with your experts the existing ethical gaps in the chosen guidelines and identify ways in which these can be refined with the TechEthos project results.
- **Capture these contributions using a template/table/spreadsheet**. Here is an example of how we started to do this for Climate Engineering – note this is still work in progress T5.2 Gap_raw data for climate engineering.xlsx .

2) **We then requested for the completed mind map and gaps template to be returned via email by 20 February 2023, to the team at DMU.**

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

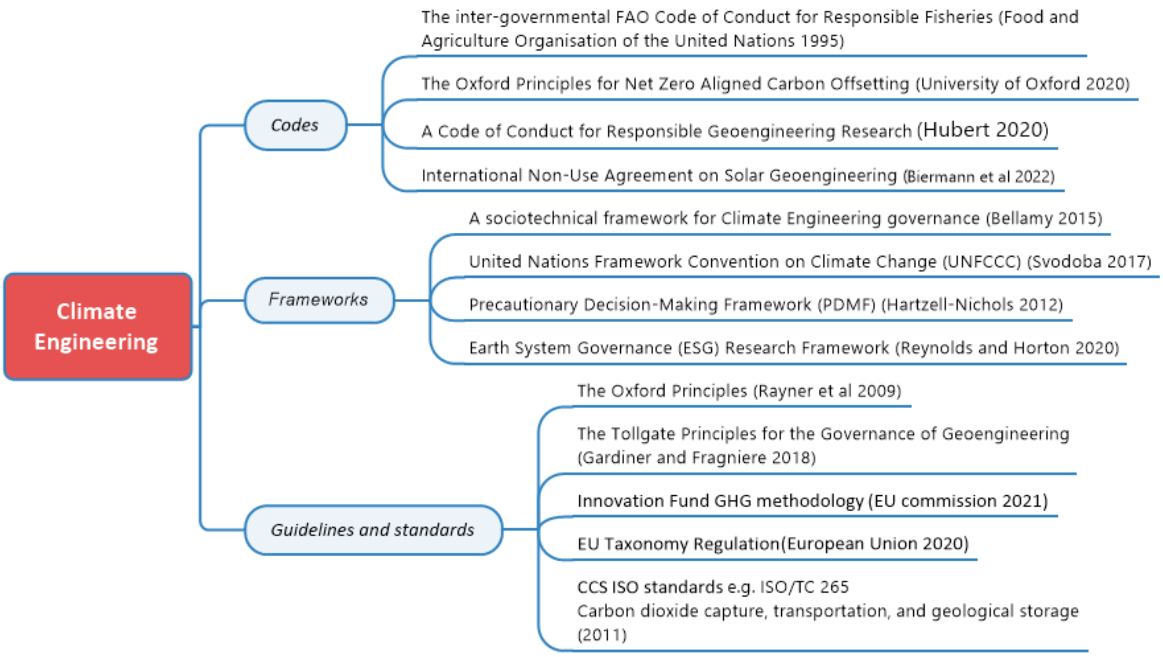

Climate Engineering
Codes
The inter-governmental FAO Code of Conduct for Responsible Fisheries (Food and Agriculture Organisation of the United Nations 1995)
The Oxford Principles for Net Zero Aligned Carbon Offsetting (University of Oxford 2020)
A Code of Conduct for Responsible Geoengineering Research (Hubert 2020)
International Non-Use Agreement on Solar Geoengineering (Biermann et al 2022)
Frameworks
A sociotechnical framework for Climate Engineering governance (Bellamy 2015)
United Nations Framework Convention on Climate Change (UNFCCC) (Svodoba 2017)
Precautionary Decision-Making Framework (PDMF) (Hartzell-Nichols 2012)
Earth System Governance (ESG) Research Framework (Reynolds and Horton 2020)
Guidelines and standards
The Oxford Principles (Rayner et al 2009)
The Tollgate Principles for the Governance of Geoengineering (Gardiner and Fragniere 2018)
Innovation Fund GHG methodology (EU commission 2021)
EU Taxonomy Regulation(European Union 2020)
CCS ISO standards e.g. ISO/TC 265 Carbon dioxide capture, transportation, and geological storage (2011)

Figure 1. Mindmap of operational guidelines, frameworks and codes for Climate Engineering

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

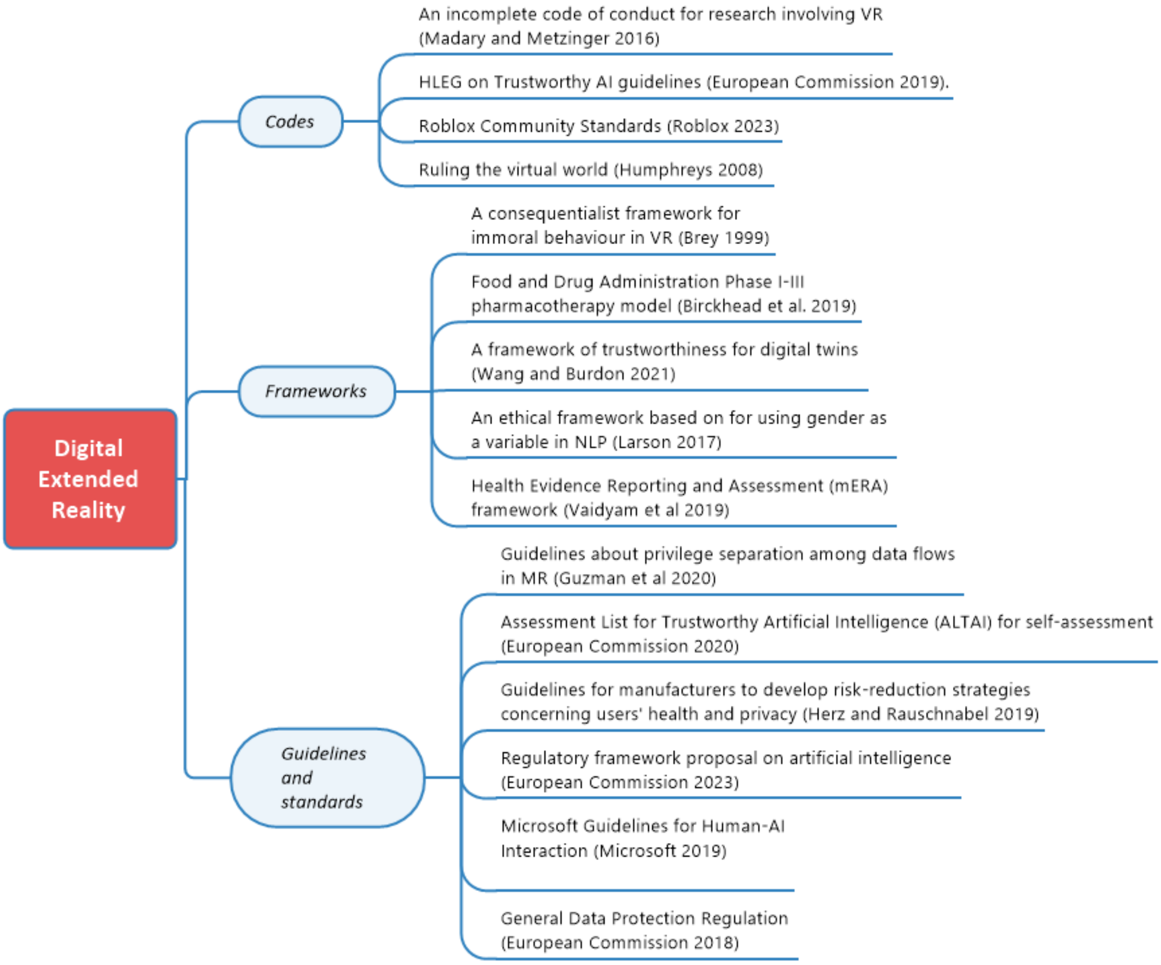

Digital Extended Reality
Codes
An incomplete code of conduct for research involving VR (Madary and Metzinger 2016)
HLEG on Trustworthy AI guidelines (European Commission 2019).
Roblox Community Standards (Roblox 2023)
Ruling the virtual world (Humphreys 2008)
Frameworks
A consequentialist framework for immoral behaviour in VR (Brey 1999)
Food and Drug Administration Phase I-III pharmacotherapy model (Birckhead et al. 2019)
A framework of trustworthiness for digital twins (Wang and Burdon 2021)
An ethical framework based on for using gender as a variable in NLP (Larson 2017)
Health Evidence Reporting and Assessment (mERA) framework (Vaidyam et al 2019)
Guidelines and standards
Guidelines about privilege separation among data flows in MR (Guzman et al 2020)
Assessment List for Trustworthy Artificial Intelligence (ALTAI) for self-assessment (European Commission 2020)
Guidelines for manufacturers to develop risk-reduction strategies concerning users' health and privacy (Herz and Rauschnabel 2019)
Regulatory framework proposal on artificial intelligence (European Commission 2023)
Microsoft Guidelines for Human-AI Interaction (Microsoft 2019)
General Data Protection Regulation (European Commission 2018)

Figure 2. Mindmap of operational guidelines, frameworks and codes for Digital Extended Reality

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

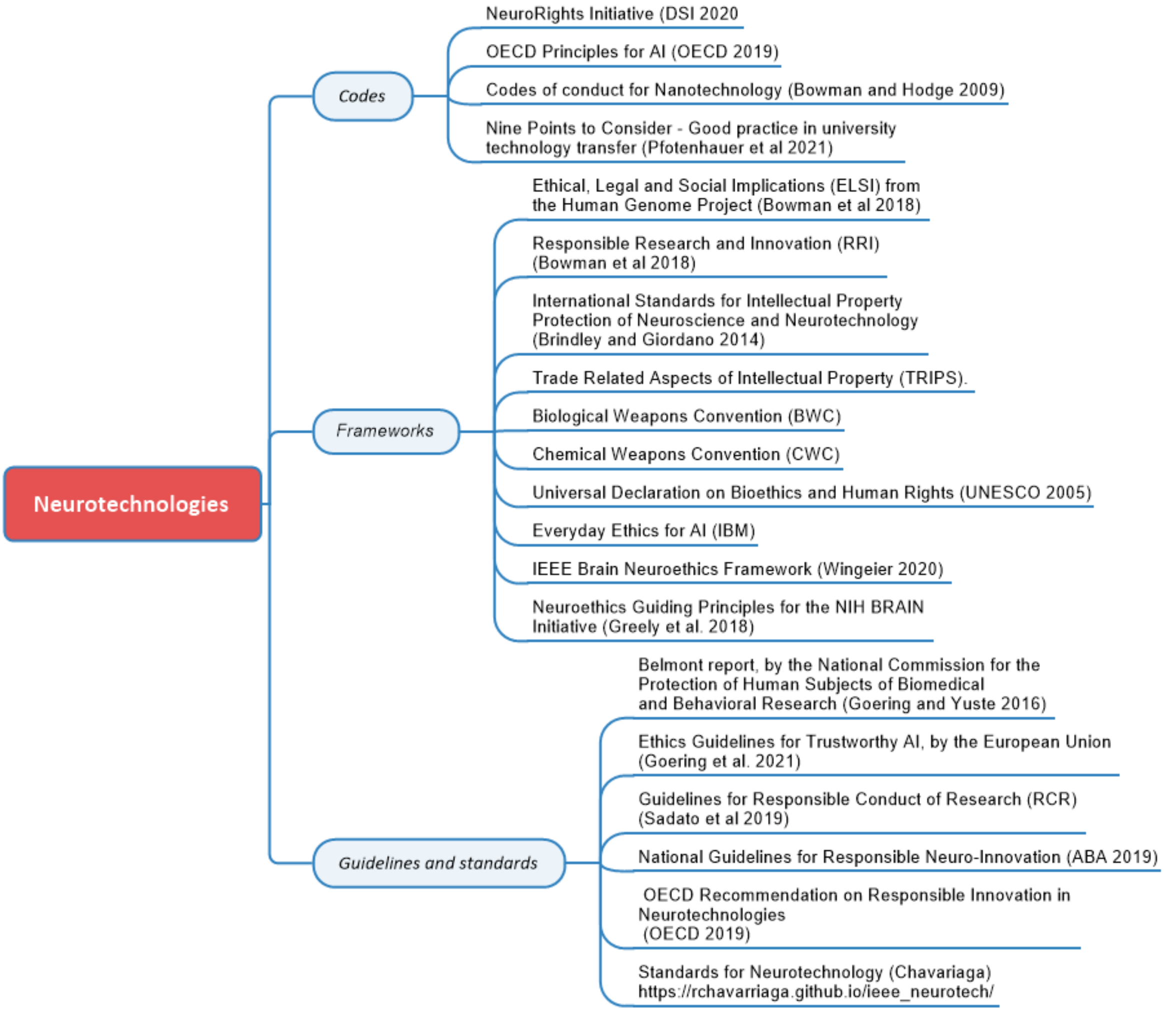

Neurotechnologies
Codes
NeuroRights Initiative (DSI 2020
OECD Principles for AI (OECD 2019)
Codes of conduct for Nanotechnology (Bowman and Hodge 2009)
Nine Points to Consider - Good practice in university technology transfer (Pfotenhauer et al 2021)
Frameworks
Ethical, Legal and Social Implications (ELSI) from the Human Genome Project (Bowman et al 2018)
Responsible Research and Innovation (RRI) (Bowman et al 2018)
International Standards for Intellectual Property Protection of Neuroscience and Neurotechnology (Brindley and Giordano 2014)
Trade Related Aspects of Intellectual Property (TRIPS).
Biological Weapons Convention (BWC)
Chemical Weapons Convention (CWC)
Universal Declaration on Bioethics and Human Rights (UNESCO 2005)
Everyday Ethics for AI (IBM)
IEEE Brain Neuroethics Framework (Wingeier 2020)
Neuroethics Guiding Principles for the NIH BRAIN Initiative (Greely et al. 2018)
Guidelines and standards
Belmont report, by the National Commission for the Protection of Human Subjects of Biomedical and Behavioral Research (Goering and Yuste 2016)
Ethics Guidelines for Trustworthy AI, by the European Union (Goering et al. 2021)
Guidelines for Responsible Conduct of Research (RCR) (Sadato et al 2019)
National Guidelines for Responsible Neuro-Innovation (ABA 2019)
OECD Recommendation on Responsible Innovation in Neurotechnologies (OECD 2019)
Standards for Neurotechnology (Chavariaga) https://rchavarriaga.github.io/ieee_neurotech/

Figure 3. Mindmap of operational guidelines, frameworks and codes for Neurotechnologies

Definitions of ethics guidelines/frameworks/codes

In TechEthos Deliverable 2.1, Section 3.4, we sought to identify relevant ethics guidelines/frameworks/codes within the selected sources for each technology family (Cannizzaro et al., 2021). Here we noted that the terms guidelines/frameworks/codes were used interchangeably in the literature and that guidelines/frameworks/codes can indeed be interrelated to each other in a complex manner, sometime hierarchically (for example codes and guidelines are considered by some as components of frameworks), hence are not strictly reducible to paradigmatic, self-contained definitions. However, for the purpose of this deliverable we did not aim to delineate such interrelations nor the hierarchical levels to which guidelines/frameworks/codes pertain but aimed to identify the main difference between these terms to lie in their level of generality i.e. ethical codes have a narrower and more specific focus and guidelines have a broader scope, with frameworks laying somewhere in the middle in terms of level of generality.

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

We capture and articulate further the distinction amongst these terms based on the example set by Rothenberg et al. (2019). Hence, we generated definitions of these terms with the purpose of defining in a clear-cut manner what constitute ethical codes, guidelines and frameworks:

- *Ethical codes* set forth responsibilities to which individuals and groups or organisations hold themselves to account. Compliance with codes may be enforced with socially mediated consequences for non-compliance or rewards for compliance. Related to emerging technologies, ethical codes elevate individual responsibility to promote desirable and/or minimise undesirable developments in the field.
- *Ethical frameworks* set forth general or specific principles to which countries, organisations, or research communities hold themselves to account. Frameworks arise in otherwise unregulated situations where groups of actors seek to alter the development trajectory of a field. Compliance with frameworks may be enforced with socially mediated consequences for non-compliance or rewards for compliance. Related to emerging technologies, ethical frameworks seek to coordinate alignments of the behaviour of collectives of individuals to promote desirable and/or minimise undesirable developments in the field.
- *Ethical guidelines* collect general or specific principles specifying how a technology or field ought to develop. Guidelines may be generated through concerted collective action of individuals or organisations. Compliance is not usually considered with guidelines. Related to emerging technologies, ethical guidelines propose development pathways intended to enhance desirable and/or minimise undesirable outcomes of a field.

Bearing this general level of distinction in mind, we note that guidelines are more future oriented and so more suitable to emerging technologies. Therefore, when consulting with experts, we found that they focused just on guidelines and so led to our selection of the key guidelines for each technology family. Having outlined their principles and identified gaps in the principles, we then devised a methodology for proposing improvements to the guidelines, which was applied to all three technology families, with the help of relevant experts.

3) Developing/refining operational guidelines using ethics by design (use and development) using WP2 results - How to link the gaps and the process of refining the guidelines? (continued).

The first step is establishing a set of principles constituting a methodology for reflecting and proposing improvements for existing operational guidelines. We firstly considered the Ethics by Design drawn from the methodology developed within the SHERPA and SIENNA projects. This consists of the following steps:

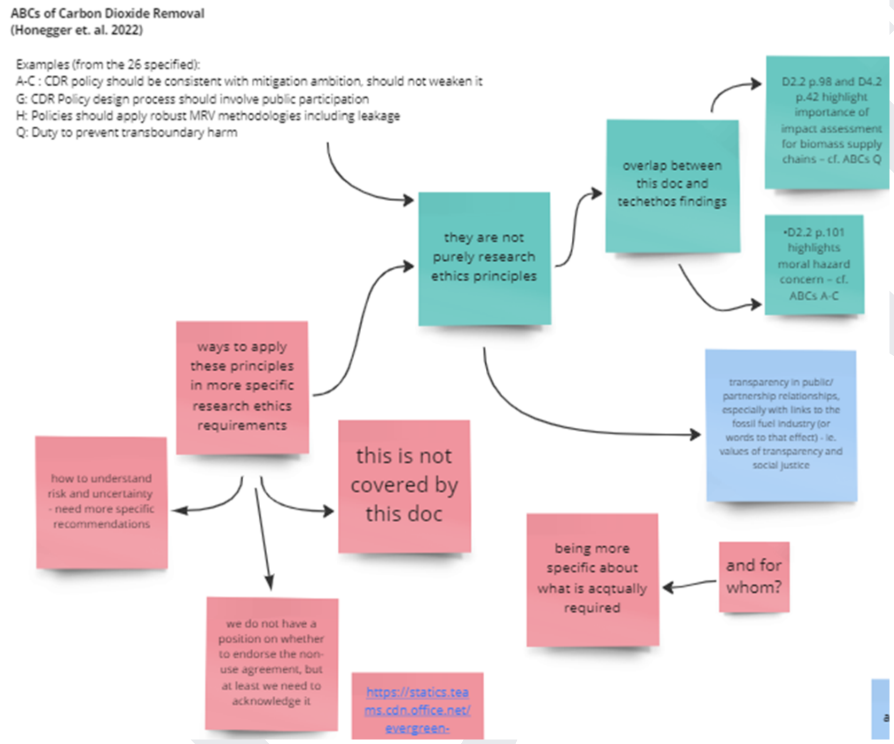
- Step 1: Reach consensus on the key moral values and principles that apply to the technology field.
- Step 2: Derive ethical requisites (or norms) from these values.
- Step 3: Choose and describe an established design methodology for the development of technology in the technology field.

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

- Step 4: Develop suggestions for improvements to operational ethics guidelines that involve a translation of the ethical requisites to actionable methodological guidelines.
- Step 5: Develop tools, methods and special topics.

Within this ethics by design approach, we adapted step 4, which addresses developing operational guidelines, to reflect on and propose improvements to existing guidelines.

We reviewed two existing ethical guidelines for each technology family, for example, within climate engineering we selected the Oxford Principles (Rayner et al 2013), and a critical response to them, Tollgate Principles (Gardiner & Fragniere 2018), for SRM, and ABCs of Carbon Dioxide Removal (Honegger et al. 2022) governance principles for CDR (for a more detailed explanation of this selection see below). After this we selected relevant sections that would serve as starting points for suggesting improvement of the existing guidelines, within this deliverable.



ABCs of Carbon Dioxide Removal
(Honegger et. al. 2022)

Examples (from the 26 specified):
A-C : CDR policy should be consistent with mitigation ambition, should not weaken it
G: CDR Policy design process should involve public participation
H: Policies should apply robust MRV methodologies including leakage
Q: Duty to prevent transboundary harm

they are not purely research ethics principles

overlap between this doc and techethos findings

D2.2 p.98 and D4.2 p.42 highlight importance of impact assessment for biomass supply chains - cf. ABCs Q

•D2.2 p.101 highlights moral hazard concern - cf. ABCs A-C

transparency in public/ partnership relationships, especially with links to the fossil fuel industry (or words to that effect) - ie. values of transparency and social justice

ways to apply these principles in more specific research ethics requirements

how to understand risk and uncertainty - need more specific recommendations

this is not covered by this doc

we do not have a position on whether to endorse the non-use agreement, but at least we need to acknowledge it

https://statics.teams.cdn.office.net/evergreen-

being more specific about what is acqtually required

and for whom?

Figure 4. Section of Miro board which we used during the CE guidelines development/refinement workshop (22 February 2023, DMU) to lay out gaps next to TechEthos findings.

To structure this part of the methodology, we sought to apply the operationalisation methodology suggested in this passage from the Sienna project (from Sienna 6.3 - refining privacy):

> “General ways to make to turn ethical requisites into concrete guidelines include dividing the ethical requisites into their component parts, specifying any actions to be taken in order to realise the ethical requisites, referencing actors who should take these actions, and relating the requisites to a concrete practice or set of practices. For example, “Adequate privacy protections should be put in place.” could thus become

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

> "Action X should be taken by person X so that adequate privacy protections are in place for data subjects during development, deployment and use of the product."

We sought to use the same operationalisation method by breaking down the selected principles within each of the guidelines into key component parts which were previously identified as bearing gaps. We then enhanced these components using the TechEthos empirical findings concerning both theoretical contributions (e.g. relevant theoretical principles as identified in WP2, D.2.2) as well as empirical results (e.g. from the stakeholder engagement part of the project as in WP3).

To put these steps into practice, we conducted workshops with internal and external experts either face to face (as for Climate Engineering on 22nd February 2023 at De Montfort University, 4 experts) and online (for Neurotechnologies on 10th July 2023, 4 experts & and digital XR on 12th July 2023, 2 experts). Furthermore, we conducted two online meetings on 2nd/3rd August 2023 with partners from WP3 on the data analysed from the consultation exercises with under-represented groups to use their findings to further enrich the development refinement of guidelines. Overall, to conduct the main reflection on existing guidelines and to propose improvements, we consulted 10 experts.

Furthermore, as we identified the need for a code of responsible conduct which was realised as guidelines for responsible research (in academia and industry) for each of the three technology families, we also sought to meet stakeholder expectations by 1) consulting with internal experts responsible for the analysis of data in WP3 - public awareness and attitudes of various stakeholder groups (2 experts) to gain an overview of expectations of under-represented groups; and 2) by soliciting experts from 21 members of the TechEthos Admin Board and gain an overview of experts' expectations to determine the way forward with the guidelines' refinement (see discussion for more details). In this way, we have solicited responses from a total of 32 experts, across the technology families.

3. Proposed Guideline Improvements

3.1 Climate Engineering

3.1.1 Introduction

A mind-map exercise on guidelines relevant to climate engineering research and development was conducted. This literature was divided into "codes", "frameworks" and "guidelines/standards".

On the basis of the mind-map, expert consultation was used to determine which of these existing documents most closely approximated current best practice. This process encompassed two stages. In the first instance, reference was made to the expert consultation exercise already conducted under Task 3.4. The purpose of the earlier exercise was to elicit responses to a series of scenarios representing possible futures in the context of imagined research and innovation pathways, with three scenarios for each of the three technology

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

families. The two-fold objective was first to refine the scenarios to ensure they interrogate the most salient ethical intuitions as precisely as possible, and second to detect expert attitudes to the scenarios themselves, in order to identify concerns to be addressed through operational guidance (see TechEthos Deliverable D3.6).[1]

These exercises surfaced a number of directly relevant considerations. In particular, a key cross-cutting concern was the need for a holistic, policy-level perspective on technological innovation and development. This was manifested most forcefully in discussions of SRM and CDR, where the importance of viewing the role of technology in the context of an overall climate strategy was emphasised, a strategy encompassing the energy, industrial policy, food, built environment and transportation sectors. The principle was assigned general significance across technology families, however. For instance, experts noted that neurotechnology should be recognized as one class of potential mental and neurological health intervention among many, and that commercial motives may accelerate the introduction of Digital Extended Reality technologies where there is no genuine need, in contexts in which it may be preferable to prioritise more low-tech solutions to social problems.

In a related concern, expert workshop participants also foregrounded the importance of ensuring ethical guidelines do not promote or presuppose a conception of “development” which privileges a narrow technical-elite perspective. This overarching consideration might find expression in ethical guidelines’ recognition of potential power imbalances between R&I actors on the one hand, and communities who might be affected by research outcomes on the other; including recognition of the need to respect global and ideally intergenerational perspectives in stakeholder analyses, actively empowering marginalised communities where necessary to ground meaningful participation.

These considerations guided the process of guideline selection on the basis of the mind-map. In the first instance, priority was given to those proposals which bore most directly on the field of CE in particular, rather than adjacent or superordinate fields. This narrowed the mind-map significantly, as several of the documents identified related to, for instance, CCS, offsetting or renewable energy. All of these have a bearing on CE in the broader policy context, but as the corresponding documents address technical aspects of the adjacent technology fields, they do not in themselves promote engagement with that broader context. Focusing on explicitly CE-oriented guidelines ensured relevance while not limiting researchers from taking the kind of synoptic view the expert workshop participants envisaged.

In the second stage, of the shortlisted guidelines, further expert consultation was used to guide selection, this time through discussion with internal experts. The Oxford Principles (Rayner et al., 2013) were initially determined to be the most appropriate starting point. This set of principles was originally produced in 2009 by an interdisciplinary group centred around the University of Oxford. It was produced, in part, pursuant to the Royal Society’s recommendation for ‘[t]he development and implementation of governance frameworks to guide both research and development in the short term, and possible deployment in the longer term’ (Shepherd, 2009). Although an early contribution to that project, it remains one with unique status.

The initial selection of the Oxford Principles was made according to the following criteria. Firstly, the Oxford Principles enjoy quasi-institutionalized recognition in policy circles, insofar

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

as they are, and remain, the only set of CE ethics principles to receive (qualified) endorsement by a government (at least until the release of the Office of Science and Technology Policy report (OSTP 2023) in by the White House, USA, in June 2023), as well as a committee of a national legislature. Following the publication of the Royal Society's report in September 2009, the House of Commons Science of Technology Select Committee in the UK convened an inquiry on the regulation of geo-engineering, inviting expert submissions. The principles were submitted in evidence to the committee, and were endorsed in the committee's report, which was then endorsed by the UK government.

Although more recent proposals have been advanced as part of government-commissioned research, notably in Germany (Bodle & Oberthür, 2014) and in the United States (National Academies of Sciences, Engineering and Medicine 2021), these proposals remain advisory. A development in the debate around SRM noted by the National Academies report which was not included in the Oxford Principles is the suggestion that SRM policy should be assessed according to a risk-risk framing, meaning the risks of SRM research as well as any eventual deployment should be assessed against the risks of not progressing with research, and thus depriving policymakers of a potentially vital shield against climate-induced harms. As the National Academies report notes, 'in practice such assessments will be extremely complex, because there are many possible combinations of climate response, some with and some without [SRM], and all will be under-described (i.e., one cannot fully know how any of the options will play out, and there will be risks and uncertainties associated with each). Additionally, each option will have benefits and drawbacks that may be difficult to assess on a single scale.' (National Academies of Sciences, Engineering and Medicine 2021, p.77). Thus, while the recognition of a potential risk-risk framing is accounted for in the below analysis, the analysis further notes the importance of assessing complex climate policy scenarios, rather than SRM development vs. its absence. The approach taken by the Oxford-Tollgate principles thus enjoys continued importance in relation to some of the most recent major analyses. On the other hand, the more recent White House OSTP report takes a much stronger view of the risk-risk framing, giving it a central place in their proposal (OSTP 2023, p.5). These contrasting approaches from such closely allied agencies indicate the framing is still a matter of open debate in the expert community, which should be acknowledged while pointing to the need for further deliberation (the White House report was published too late to be fully included in the present analysis). A further reason for the selection of the Oxford principles is the influential position they retain in academia, with an authoritative 2019 review of solar geoengineering governance literature in *Proceedings of the Royal Society A* referring to them as 'the most influential set of principles on climate engineering' (Reynolds, 2019). Rayner et al. (2013) has the highest citation count of surveyed publications directly on CE ethics and governance (Web of Science - 122, Scopus - 147, Google Scholar – 253, retrieved 3 March 2023.).

However, as an early contribution to this literature, the Oxford Principles have since their original formulation and subsequent publication received sustained critical evaluation. It was judged important to reflect these developments, while continuing to acknowledge the principles' ongoing significance. In this respect, the Tollgate Principles (Gardiner & Fragnière, 2018) were identified as a complement to the Oxford principles. The proposal was that Tollgate principles should not be treated as a novel competing framework, but rather as an updating and correction of the Oxford principles in light of compelling philosophical analysis,

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

which retains the earlier proposal's basic structure and central concerns while adding a further level of ethical precision.

A method of expert consultation was applied to identify gaps in the selected guidance documents, the Oxford Principles and the Tollgate principles. The most significant gap identified was the treatment these documents gave to the distinction between the ethics and governance of Solar Geoengineering, as against Carbon Dioxide Removal techniques. As noted, the Oxford Principles were in part a response to the Royal Society's 2009 report, which defined its subject matter as 'geoengineering' understood as 'the deliberate large-scale manipulation of the planetary environment to counteract anthropogenic climate change' (Shepherd, 2009), and thus the principles' authors retained this designation of their subject matter.

Despite formally being directed at both CDR and SRM, the Oxford principles have little to say on specificities of CDR governance. Their application to CDR may also be misleading. For instance, it is implausible the authors intended through principle 5 – 'Governance before deployment' (Rayner et al., 2013) – to promote a moratorium on all carbon removal activities until legal governance frameworks were in place, including, for instance, all tree-planting and land-use changes. At the time, the field of CDR ethics and governance was less advanced than it is at present, and the disadvantages of attempting to capture both CDR and SRM under a common framework were less clearly identified.

The Tollgate authors explicitly restrict the focus of their concern to SRM. They state that they 'aim to sidestep definitional discussions' by 'assuming that we are discussing the paradigm case of stratospheric sulphate injection (SSI)', adding '[t]he extent to which other interventions share the features that make all or some of the Tollgate Principles appropriate…are topics for another occasion' (Gardiner & Fragnière, 2018, p. 145). In other words, they make explicit what the Oxford Principles arguably leave implicit: that the principles are formulated with SRM in mind (in particular, Stratospheric Aerosol Injection), and leave open the question of whether they apply to other interventions standardly termed "geoengineering".

Neither of these approaches to the question of the guidelines' interoperability between CDR and SRM applications was determined to be entirely satisfactory in expert consultation. Since the publication of the Oxford Principles, a norm has emerged across disciplines according to which CDR and SRM are treated as fundamentally different categories of intervention. The IPCC's 6th Assessment report analyses CDR under Working Group III, thus regarding it as mitigation, while SRM is analysed separately (IPCC 2022, 2022; 14.4.5). CDR is regarded as a form of mitigation in international law, while the status of SRM in international law is uncertain (Honegger et al., 2021). There is also philosophical literature arguing that CDR and SRM should not for most purposes be analysed together under the term "geoengineering", for example an influential intervention by one of the Oxford Principles' authors (Heyward, 2013). In the Climate Ethics literature specifically, the ethics of SRM and the ethics of CDR are increasingly becoming two separate sets of literature (while there used to be a subfield in climate ethics called "the ethics of geoengineering/climate engineering", it is increasingly standard practice to distinguish between the two different subfields of the ethics of SRM and the ethics of CDR).

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

This norm reflects substantive ethical, technical and policy considerations. For instance, while drawing a sharp distinction between the ethical and regulatory requirements of testing vs deployment in relation to CDR technologies is relatively straightforward, the same cannot be said for SRM without qualification. This is because evidence suggests that any test of the scale and duration required to produce data sufficient to predict the effects of a full-scale deployment of SAI would not fit plausible definitions of a test, given it would likely require decades of operation and have global impact (Lenferna et al., 2017), meaning the testing vs deployment distinction requires careful regulatory attention.

Thus, across Science Policy, Law and Ethics, the fields of CDR and SRM research have diverged so markedly as to make a unified research guidance framework for CDR and SRM impracticable, despite the attempt of the Oxford authors. Experts noted that the Tollgate authors' decision to treat Stratospheric Aerosol Injection as a 'paradigm' offers some guidance as to an appropriate approach: moving forward, the principles should be treated as applying only to SRM. If certain forms of SRM diverge markedly from the SAI paradigm, further specification of operational guidelines should determine how they are to be applied.

The decision to limit the application of the selected template guidelines to SRM revealed the need to survey existing guidelines relevant to CDR, to determine whether any extant proposal was suitable for use as a template for initial analysis. This was again achieved via internal expert consultation. The literature on guidelines for CDR research and development is much less advanced than the literature on climate engineering in general or SRM in particular, with a number of calls for such guidelines to be formulated having been issued (*AGU Climate Intervention Engagement: Leading the Development of an Ethical Framework*, 2022; Cox et al., 2018; Loomis et al., 2022), but few fully developed proposals thus far introduced. Important progress is being made on issues adjacent to R&I, including accounting standards for removals (European Commission 2022, Procedure 2022/0394/COD). Sector specific standards for CCS have been formulated (*ISO/TC 265 - Carbon Dioxide Capture, Transportation, and Geological Storage*, 2023). There have also been moves within the carbon removals industry to develop self-administered ethics standards and/or codes of conduct, although these appear somewhat cursory at present (Carbon Business Council, n.d.; Planetary Technologies, n.d.). These contributions were determined to be either too restricted in their application, or insufficiently developed, to serve as a template for guidance.

Honegger et al. (2022) was identified as the most comprehensive and developed extant proposal relevant to operational guidance of CDR. The proposal was produced under the project CDR-PoET (Carbon Dioxide Removal Options, Policy and Ethics), funded by the German Federal Ministry of Education and Research under the CDRterra research programme[1]. In addition to their being the most developed extant framework of their kind, a further criterion for their selection was the methodology used to produce them: the authors review existing hard law directives and soft law guidance relevant to CDR policy, which are then used to abstract away a 'conceptual framework regarding norms and principles relevant to CDR' (Honegger et al., 2022, p. 2). The document thus constitutes both an up-to-date survey of relevant extant instruments, and an important contribution towards organising guidance in an operationally usable form. As a policy-oriented set of governance principles, the framework's

[1] CDR-PoEt - CDRterra. Accessed 7 March 2023. https://cdrterra.de/en/consortia/cdr-poet

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

primary relevance in the present context is in providing guidance for the determination of research objectives and insuring that R&I goals are compatible with overall CDR policy goals.

It should be noted that the Oxford and Tollgate principles on the one hand, and the ABCs on the other, represent different methodological approaches which lead to a correspondingly different institutional status. As noted, the ABCs did not introduce any novel principles but sought to systematise what principles could already be found rooted in existing applicable governance contexts as well as the emerging CDR governance literature. They are therefore intended as a clarification of norms to which, in some cases at least, relevant actors already have legal reasons to adhere. The authors of both the Oxford Principles and the Tollgate Principles, meanwhile, were operating in an environment in which there was much less existing relevant international regulation and policy with respect to SRM (a situation which has not much changed). While there were arguments that norms for the regulation of SRM could be derived from current practice, there were also strong reasons to believe that current practice was not well suited to this task and required revision. Furthermore, the Oxford principles were intended to serve as scaffolding for intergovernmental and interdisciplinary development of formal governance regimes, as this collaborative process was considered key to securing their legitimacy. The Tollgate principles, as an intellectual descendant of the Oxford principles, should be read in the same context. Thus, though the principles go beyond currently accepted international norms, they represent a significant current of expert opinion making them appropriate for this scaffolding role.

A brief comment on a methodological disagreement between the Oxford and Tollgate principles: Heyward, Rayner and Savelescu (2017) defending the Oxford Principles (of which they were co-authors) against rival proposals, point out that the authors deliberately confined themselves to appeal to procedural values, on the grounds that sets of principles which go beyond procedural values and appeal to substantive values fail to cohere with norms of procedural justice. The claim is that it would not be compatible with procedural justice to compel people to adopt principles based on substantive values they themselves do not acknowledge. The Tollgate Principles appeal to substantive ethical values. It can therefore be argued that they fail to cohere with norms of procedural justice.

It was determined this methodological concern was overstated, on the following grounds. The Oxford Principles' authors' claim that the principles are not substantive but merely procedural is contentious. The argument that 'ethical worldviews… are to be expressed in the public discussion which the Principles make space for, rather than being assumed at the outset' (Heyward et al., 2017, p. 112) obscures the difficulty in distinguishing supposedly uncontroversial procedural values from values which are substantive and therefore essentially contested. For instance, there are implicit ethical assumptions embodied in the constitution of publics and the selection of deliberative procedures. The Tollgate authors contend that the relevant public for discussions of geoengineering governance is 'global, intergenerational and ecological' (Gardiner & Fragnière, 2018, p. 155). This indeed is a substantive ethical claim, but it also sheds light on the substantive ethical claim the Oxford Principles make implicitly: that the relevant public is not global, intergenerational and ecological, but restricted to geographically demarcated, currently extant human individuals whose interests might be materially affected by geoengineering deployment. The decision to treat substantive ethical values as unavoidable coheres with the mode of ethical analysis adopted under TechEthos Deliverable

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

2.2, which highlights the tendency of approaches which purport to be neutral to 'rationalize the status quo' (Adomaitis et al., 2022, p. 22; Mills, 2005, p. 181).

That being said, one implication of the argument put forward by (Heyward et al. 2017) is the critique that the Tollgate authors' substantively normative approach excludes other reasonable approaches. This report contextualises the position taken in relation to this critique below, in the section headed (Addendum to CDR Guideline: Policy relevance of normative approach). At this point, it is necessary to note that where alternative framings are available in the literature, this is acknowledged in the course of the guideline revision process. This constitutes an important contribution of the present report.

The two selected guidelines for development and refinement for climate engineering were the Tollgate Principles for SRM and ABCs of Carbon Dioxide Removal governance principles for CDR. Within these guidelines we selected relevant sections that would serve as starting points for the TechEthos refined ethical guideline, within this deliverable.

3.1.2 Selected guidelines for Climate Engineering (CE)

Below we present the Tollgate principles which we used as a starting point for the first guideline for climate engineering.

Original Tollgate Principles for SRM

(Gardiner & Fragnière, 2018)

1. Framing: Geoengineering should be administered by or on behalf of the global, intergenerational and ecological public, in light of their interests and other ethically relevant norms.
2. Authorization: Geoengineering decision-making (e.g. authorising research programs, large-scale field trials, deployment) should be done by bodies acting on behalf of (e.g. representing) the global, intergenerational and ecological public, with appropriate authority and in accordance with suitably strong ethical norms, including of justice and political legitimacy.
3. Consultation: Decisions about geoengineering research activities should be made only after proper notification and consultation of those materially affected and their appropriate representatives, and after due consideration of their self-declared interests and values.
4. Trust: Geoengineering policy should be organized so as to facilitate reliability, trust and accountability across nations, generations and species."
5. Ethical Accountability: Robust governance systems (including of authority, legitimacy, justification and management) are increasingly needed and ethically necessary at each stage from advanced research to deployment.
6. Technical Availability: For a geoengineering technique to be policy-relevant, ethically defensible forms of it must be technically feasible on the relevant timeframe.
7. Predictability: For a geoengineering technique to be policy relevant, ethically defensible forms of it must be reasonably predictable on the relevant timeframe and in relation to the threat being addressed.

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

8. Protection: Climate policies that include geoengineering schemes should be socially and ecologically preferable to other available climate policies, and focus on protecting basic ethical interests and concerns (e.g. human rights, capabilities, fundamental ecological values).
9. Respecting General Ethical Norms: Geoengineering policy should respect general ethical norms that are well-founded and salient to global environmental policy (e.g. justice, autonomy, beneficence).
10. Respecting Ecological Norms: Geoengineering policy should respect well-founded ecological norms, including norms of environmental ethics and governance (e.g. sustainability, precaution, respect for nature, ecological accommodation).

Below we present the second starting set of guidelines for climate engineering, using the ABC of CDR.

Original ABCs of Carbon Dioxide Removal (CDR)

(Honegger et al., 2022)

a) CDR should be considered in NDCs,
b) CDR policies should not weaken other mitigation efforts
c) Resulting CDR efforts should be commensurate with the long-term collective mitigation ambition
d) Policy mixes should include technology-transfer to help strengthen capacities for CDR
e) Policy mixes should include international cooperation to improve CDR efficiency
f) Policy mixes should include international climate finance transfers to mobilize CDR.
g) Policies should ensure consistent accounting for CDR results applying conservative baselines and including leakage
h) Policies should apply robust MRV methodologies including on leakage
i) CDR policies should fulfil principles of inter- and intragenerational equity (e.g., Polluter Pays or Ability to Pay).
j) Efforts should internationally be differentiated per common-but-differentiated responsibilities
k) Efforts should internationally be differentiated by respective capacities and (national) circumstances.
l) Policies should include a national determination of clear objectives, policies, and metrics for CDR
m) Policies should consider both short- and long-term effectiveness and efficiency
n) Policies should be procedurally just
o) The policy design process should involve public participation and stakeholder involvement
p) Policies should contribute to sustainable development
q) Policies should prevent transboundary harm > Q: Duty to prevent transboundary harm
r) Policies should prioritize rectifying damage at source
s) Policy designs should reflect multi-risk trade-offs including policy or technology failure risks as well as countervailing risks of omitting policy steps.

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

t) Anticipation of longer-term CDR needs incl. toward net-zero or net-negative emissions targets
u) Policy mixes should include strategies for preventing over-promise and under-delivery
v) Policies should include intermittent targets and policy objectives
w) Policies should be adapted upon missing intermittent targets and objectives.
x) Policies should involve increasingly specific targets for various CDR and emission reduction methods
y) Policies should reflect CDR methods' specificities
z) Policy ensembles should meet the needs of the targeted methods

3.1.3 CE guidelines' gaps

The template guidelines selected represent high-level policy guidance for SRM and CDR. No operational guideline documents for R&I in these fields are currently extant in a sufficiently developed form. The guidelines selected therefore represent best practice in terms of high-level guidance - some of the principles listed have direct implications for the operational guidance of research and development, while for others, further work is needed to operationalise the guidance for the R&I context in particular, distinguishing this function from legislative and policy guidance. As high-level guidance, it has been remarked (Morrow, 2018; Nericcio, 2018) that putting the Oxford and Tollgate principles into practice in the context of project development, public policy, etc. is not necessarily straightforward.

This criticism should not be overstated, since the authors' aim to set out a framework which applies to a broad range of institutional contexts - from international policy, to internal science policy and funding, to research - necessarily precludes detailed context-specific guidance. Nevertheless, the principles' authors themselves acknowledge that further work is needed to translate the high-level principles into operational guidance. Furthermore, as noted, the Tollgate principles constitute a moment in an ongoing academic debate, awareness of which is important when interpreting the principles' content. Some alterations to the text are necessary to bring this context into the foreground, in order to produce self-standing guidance that can be applied by an end-user without the need to do additional interpretive work. Honegger et al. (2022) is a guideline for policy specifically; only a subset of principles listed will be of direct operational use to R&I actors.

The process of identifying gaps therefore has two components: one, the identification of gaps in the ethical content of the principles themselves (on the basis of TechEthos findings and more recent developments in the literature), two, reframing the updated sets of principles in a more operationally accessible form.

One means of moving from high-level guidance to operational guidance is via the promotion of a code of conduct for researchers. The possibility of developing a code of conduct for climate engineering research has for a long time been a feature of the literature on climate engineering ethics and governance. The call to develop a code of practice for geoengineering research was one of the recommendations of the 2009 Royal Society Report (Shepherd, 2009, p. xii). This call was taken up by the Geoengineering Research Governance Project, a

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

collaboration between the University of Calgary (Canada), the Institute of Advanced Sustainability Studies Potsdam (Germany) and the University of Oxford (UK). This project produced (Hubert & Reichwein, 2015), which was further refined as (Hubert, 2021), *Code of Conduct for Responsible Geoengineering Research.*

The method of identifying gaps in the existing operational guidance did not identify a need to develop a novel code of conduct, for two reasons. Firstly, codes of conduct are only one element of operational guidance, which circumscribe the norms of acceptable individual behaviour in a given context. A code of conduct is a voluntary instrument which may give the impression that it exhausts the domain of ethical action and responsibility with respect to a given field, leading to 'ethics washing' concerns. Hubert (2021) already exists as a voluntary code of conduct to which R&I actors may subscribe if they wish, this should be in addition to, and not instead of, TechEthos operational guidance.

3.1.4 Proposed improvements to Guidelines (CE):

Our strategy is to build on existing guidelines in order to reflect and to suggest improvements. As such, we built on Gardiner and Fragnière (2018), hereafter the Tollgate principles (applicable to SRM) and Honegger et al. (2022), hereafter the ABCs (applicable to CDR). To that end, we organised a hybrid workshop in which we identified potential gaps in the selected guidelines, drawing on the expert knowledge of the consortium members, plus outputs from WP2, WP3 and WP4. In the following section, we juxtapose the principles identified within the two existing ethical guidelines for both CDR and for SRM with the gaps identified during expert consultation.

Both the Tollgate Principles and the ABCs constitute normative models to which CE policy ought to conform. They leave open, to some extent, the question of how policy can be made to conform to the principles, or what concrete actions need to be carried out and by whom. In order to complete the process of operationalising the guidelines, it is necessary more precisely to specify which parties have primary responsibility for ensuring the fulfilment of each principle, or at least, who are the most relevant actors in relation to each principle. While it is not possible to specify concrete actions for every actor connected to the innovation ecosystem for SRM and CDR, TechEthos is in a position to make an important contribution to the debate on guidance by connecting specific principles to specific actors at specific stages in the R&I process, an application of an Ethics by Design methodology.

Building on the innovation ecosystem mappings carried out by TechEthos under D3.1, §2.3.1, three potential stakeholder groups have been identified as most relevant for the application of operational guidelines: researchers, manufacturers and technology providers, and policymakers.

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

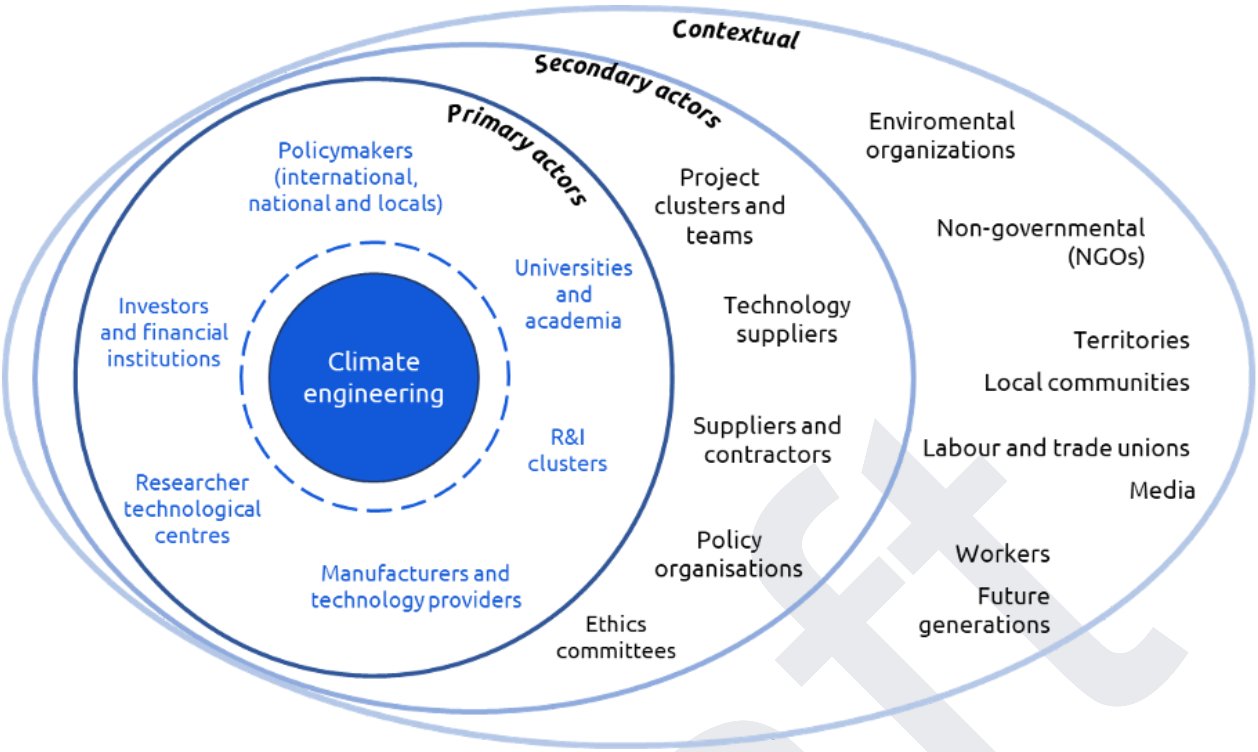

Contextual
Secondary actors
Primary actors
Climate engineering
Policymakers (international, national and locals)
Investors and financial institutions
Universities and academia
R&I clusters
Researcher technological centres
Manufacturers and technology providers
Project clusters and teams
Technology suppliers
Suppliers and contractors
Policy organisations
Ethics committees
Enviromental organizations
Non-governmental (NGOs)
Territories
Local communities
Labour and trade unions
Media
Workers
Future generations

Figure 5: Innovation ecosystem of climate engineering technologies.

These have been identified as core primary actors, accepting some overlap between the three categories. R&I clusters and researcher technological centres bridge academia and industry, thus some guidance focused on the universities sector will be appropriate to actors in R&I clusters, while for others, guidance for manufacturers will be more appropriate. Investors and financial institutions will have some overlap with policymakers, although in many areas specific guidance is required, in particular, the updating of ESG frameworks. This is a project that falls outside the scope of TechEthos's core technology ethics remit.

Proposed improvement to The Tollgate Principles for SRM

Starting with the Tollgate principles, the juxtaposition works as follows - the outlined principle is matched up with a corresponding "gap" or suggestions for refinement.

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

Tollgate principle guidelines	**Annotations for the refinement of the operational guidelines**	**Proposed refined operational guidelines**
1. Framing: Geoengineering should be administered by or on behalf of the global, intergenerational and ecological public, in light of their interests and other ethically relevant norms.	The term "Geoengineering" in this principle is ambiguous as to whether it refers to deployment or research. The principle should be specified as "SRM as a research and policy programme", thereby applying to the administration of research as well as any potential deployment. This principle must be relativised to particular contexts to determine what administration on behalf of the global, intergenerational and ecological public demands for actors in those specific spheres, with concurrent consideration of the global, intergenerational and ecological concerns related to foregoing SRM and allowing the planet to continue warming. It should contemplate an economical, technological and political feasibility of the technology family, that is also ethically acceptable. At a minimum, administration on behalf of the public precludes administration on behalf of an interest group, for instance, a group of shareholders or private investors in a given project. **Policymakers** should direct research strategy so as to ensure the acquisition of intellectual property in SRM technologies is managed in the public interest, in certain cases restricting the conditions under which private entities can acquire intellectual property. **Technology providers** should manage intellectual property in a manner which serves the public interest. **Researchers** should be cognisant that their primary responsibility is to the global, intergenerational and ecological public, and not to any private interest.	1. Framing: SRM as a research and policy programme should be administered by or on behalf of the global, intergenerational and ecological public, in light of their interests and other ethically relevant norms. This should contemplate an economical, technological and political feasibility of SRM as a research and policy programme, that is also ethically acceptable.

Tollgate principle guidelines	**Annotations for the refinement of the operational guidelines**	**Proposed refined operational guidelines**
2. Authorization: Geoengineering decision-making (e.g. authorising research programs, large-scale field trials, deployment) should be done by bodies acting on behalf of (e.g. representing) the global, intergenerational and ecological public, with appropriate authority and in accordance with suitably strong ethical norms, including of justice and political legitimacy.	The norms of political legitimacy need to be even handed and imply an extremely demanding standard for the authorisation of any open-air tests of stratospheric SRM techniques that would be expected to have significant transboundary physical risks. (Morrow et al., 2013) This implies an effective moratorium on deployment, and on open-air experiments that have significant risk of transboundary harm **Policymakers, with contributions from the public:** should co-produce appropriate research ethics and governance standards, prior to any decision to expand SRM research. **Technology providers:** should not enter into partnerships which envisage deployment until there is international recognition that the conditions for legitimating an SRM policy programme set out in this document have been met. **Researchers:** Since there remains great polarisation about SRM research, researchers may choose to endorse the view of two open letters from prominent groups of researchers which call for expanded SRM research to ensure future decision-making on deployment can be made on an informed basis (Open Letter, 27 Feb 2023; Call for Balance, March 2023). Alternately, researchers may choose to endorse the proposal by one group of researchers which calls upon states not to fund the development of solar geoengineering technologies, the so-called non-use agreement (Biermann et al., 2022). The present document	2. Authorization: SRM decision-making (e.g. authorising research programs, large-scale field trials, deployment) should be done by bodies acting on behalf of (e.g. representing) the global, intergenerational and ecological public, with appropriate authority and in accordance with suitably strong ethical norms, including of justice and political legitimacy.

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

Tollgate principle guidelines	Annotations for the refinement of the operational guidelines	Proposed refined operational guidelines
	acknowledges but does not actively endorse these proposals.	
3. Consultation: Decisions about geoengineering research activities should be made only after proper notification and consultation of those materially affected and their appropriate representatives, and after due consideration of their self-declared interests and values.	Different levels of consultation are appropriate alongside different levels of research: formulation of research policy, lab research and modelling, open-air experimentation. TechEthos's engagement with public groups across Europe, including under-represented groups, has identified key values in relation to SRM that should be taken into consideration in future consultations.Most notably the key values are: Ecosystem health, safety and reliability, effectiveness and efficiency, justice (both global distribution of justice & intergenerational justice, from D3.1, p.100), and naturality. TechEthos consultation identified key concerns in relation to the unilateral deployment of SRM technologies (SAI). It also identified a call to evaluate SRM in comparison to the alternative, the negative impacts of non-deployment. Alongside consultation, capacity-building would be should be pursued - so that participants have knowledge and understanding to inform their contributions to the consultation process. One of such ways is gamification as used in TechEthos (D3.2) to engage with citizens from different backgrounds. Policymakers and researchers should engage in cooperation to ensure that sufficient capacity for relevant public bodies to meaningfully participate in consultation is in place. This capacity building process should run in parallel to	3. Consultation: Decisions about SRM research activities should be made only after proper notification and consultation of those materially affected and their appropriate representatives, and after due consideration of their self-declared interests and values. Alongside consultation, capacity-building should be pursued. Policymakers and researchers should engage in cooperation to ensure that sufficient capacity for relevant public bodies to meaningfully participate in consultation is in place

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

Tollgate principle guidelines	Annotations for the refinement of the operational guidelines	Proposed refined operational guidelines
	consultation, rather than acting as a barrier to it. Steps should be taken to ensure the process of constituting deliberative democratic bodies is itself sufficiently inclusive and unbiased.	
4. Trust: Geoengineering policy should be organized so as to facilitate reliability, trust and accountability across nations, generations and species.	When considering trust, there is the need to underline/foreground the values of transparency and social justice. Particularly, trust needs to be complemented with transparency (and explainability) in public/private partnership relationships, especially when it comes to the links to the fossil fuel industry/parties with vested interests in the development of SRM. SRM research programmes, including social science and modelling, and certainly any experimentation, should as far as possible be internationally cooperative, and should be subject to public periodic reporting, and open publication of results, including negative results and balanced reporting of positive results[2]. Policymakers and researchers should cooperate to ensure these conditions are implemented. There are different types of trust - in the technologies and in the institutions that develop and those that regulate it.	4. Trust: SRM policy should be organized so as to facilitate reliability, trust, transparency and social justice and accountability across nations, generations and species.

[2] For example, see: https://peteirvine.substack.com/p/are-srm-scientists-boosters-or-blockers?utm_source=profile&utm_medium=reader2

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

Tollgate principle guidelines	Annotations for the refinement of the operational guidelines	Proposed refined operational guidelines
5. Ethical Accountability: Robust governance systems (including of authority, legitimacy, justification and management) are increasingly needed and ethically necessary at each stage from advanced research to deployment.	**Policymakers**: Should ensure that governance institutions for SRM are constituted to manifest, inter alia, the virtues of transparency and accountability (see Morrow, Kopp & Oppenheimer 2013). This implies an effective mechanism for affected parties to collectively monitor and collectively approve or reject the decisions and actions of those governance institutions. Monitoring requires that institutions provide information on their goals and behaviour in a format intelligible to all relevant global publics.	5. Ethical Accountability for SRM: Robust governance systems (including of authority, legitimacy, transparency, justification and management) are increasingly needed and ethically necessary at each stage from advanced research to deployment.
6. Technical Availability: For a geoengineering technique to be policy-relevant, ethically defensible forms of it must be technically feasible on the relevant timeframe.	Reference to "ethically defensible" forms seems to threaten circularity in the context of an ethical guidance document. "Ethically defensible forms" should here be understood as forms that can reasonably be defended against the charge they violate the norms underlying Tollgate principles 9 and 10 (for this argument see Morrow 2018).	6. Technical Availability: For a SRM technique to be policy-relevant, ethically defensible forms of it must be technically feasible on the relevant timeframe.
7. Predictability: For a geoengineering technique to be policy relevant, ethically defensible forms of it must be reasonably predictable on the relevant timeframe and in relation to the threat being addressed.	This principle should include a re-elaboration of the concern of irreversibility from using or rejecting SRM (D2.2). Irreversibility refers to irreversible harm and irreversible changes in the Earth system linked to the crossing of climate tipping points, which are a function both of climate change with or without SRM. The likelihood of tipping points is known to increase with warming. Irreversibility also links to the termination-shock concern, i.e. if SRM is deployed to a high level and then suddenly and permanently stopped, the rebound effect may be worse than the situation that was initially tackled by the first intervention.	Predictability: For a SRM technique to be policy relevant, ethically defensible forms of it must be reasonably predictable on the relevant timeframe and in relation to the threat being addressed. Re-elaboration of the concern of irreversibility from using or rejecting SRM

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

Tollgate principle guidelines	Annotations for the refinement of the operational guidelines	Proposed refined operational guidelines
8. Protection: Climate policies that include geoengineering schemes should be socially and ecologically preferable to other available climate policies, and focus on protecting basic ethical interests and concerns (e.g. human rights, capabilities, fundamental ecological values).	Note that comparisons between policies should factor in consideration of the risk of warming arising from any decision not to develop SRM. The risk-risk framing should be attached to "ecological guardrails". Risks of non-implementation of SRM should be considered in the context of climate policy scenarios which are compatible with human rights/human wellbeing, and fundamental ecological values, rather than taking for granted scenarios in which policymakers fail to protect these values.	8. Protection: Climate policies that include SRM schemes should be socially and ecologically preferable to other available climate policies, and focus on protecting basic ethical interests and concerns (e.g. human rights, capabilities, fundamental ecological values). Non-implementation of SRM should be considered in the context of climate policy scenarios
9.Respecting General Ethical Norms: Geoengineering policy should respect general ethical norms that are well-founded and salient to global environmental policy (e.g. justice, autonomy, beneficence).	Note that "geoengineering policy" also comprises decisions to constrain or delay SRM research.	9.Respecting General Ethical Norms: SRM policy should respect general ethical norms that are well-founded and salient to global environmental policy (e.g. justice, autonomy, beneficence). Constrain or delay SRM research should also be considered.

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

Tollgate principle guidelines	Annotations for the refinement of the operational guidelines	Proposed refined operational guidelines
10. Respecting Ecological Norms: Geoengineering policy should respect well-founded ecological norms, including norms of environmental ethics and governance (e.g. sustainability, precaution, respect for nature, ecological accommodation).	See above clarification of term "geoengineering policy".	10. Respecting Ecological Norms: SRM policy should respect well-founded ecological norms, including norms of environmental ethics and governance (e.g. sustainability, precaution, respect for nature, ecological accommodation). Constrain or delay SRM research should also be considered.

Table 4: Refined Tollgate principles for SRM

As already intimated, to understand how these principles can be operationalised, it is necessary to recognise the extent to which they form a dialogue with the Oxford principles. The Oxford principles state that 'Wherever possible, those conducting geoengineering research should be required to notify, consult, and ideally obtain the prior informed consent of those affected by the research activities' (Rayner et al., 2013). The Tollgate authors argue for the modification of this principle, insofar as it limits the scope of public participation to 'research activities' and 'those affected' by them. This excludes from the scope of public participation, for example, oversight of research policy programmes, or indeed, actual deployment of CE. It also excludes from participation those people who are not affected by research, which the Tollgate authors interpret to mean those not whose interests are not materially set back by experimental activities in themselves, which they argue is an unwarranted restriction.

At least one critic (Nericcio, 2018) has claimed that the Tollgate authors' case for expanding the scope of public participation implies similar wide scope public participation requirements would be necessary for any intervention in the climate system, even for purposes other than climate engineering - for instance, greenhouse gas emissions. The criticism is that if we think such requirements implausible in the case of greenhouse gas emissions, then the same requirements in the case of CE must be too demanding. Nerricio (2018) reads the Tollgate authors' rhetorical question "is it not our planet too?" as expressing their argument for wide scope participation. He compares the argument to the claim that any resident of a city should have participatory rights with respect to planning approval for development in the city, on the grounds that it is "their city", even if it is an area of the city they never visit and with which they have no connection.

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

The Tollgate authors have resources to defend themselves against this argument. The claim is not that people who are completely unaffected by an intervention in some domain should nevertheless have the right to participate in decision making about it if that domain is in any sense “theirs”. The claim is rather that geoengineering in particular, as a research programme, significantly affects extant and future people on a global scale, even if those effects are indirect or not ‘material’. It is therefore appropriate that they be subject to democratic legitimation of some kind.

However, the criticism gives rise to the important observation (see also Morrow, Kopp & Oppenheimer 2013) that deliberation is not the only potential mechanism for legitimation: legitimate institutions can in relevant cases be empowered to make decisions on behalf of the publics they represent. This consideration supports the argument (see Morrow 2020), that a mission-driven SRM research program, with objectives and constraints clearly established by legitimate institutions, could enable research projects to achieve the required standards of legitimacy. This would mean that decisions to limit or slow SRM research and development, which also significantly affect extant and future people on a global scale (even if those effects are indirect or not ‘material’) could also be subject to legitimation through democratic channels. It would of course be important to consider the extent to which the legitimacy of mission-driven research programmes would be constrained by the authority of the institution coordinating such programmes. For instance, a single state or limited group of states would not have the authority to act on behalf of global and intergenerational publics.

The above updated principles are informed by TechEthos expert consultation under D3.5 in a number ways. For one, there are concerns, particularly related to unilateral deployment of technologies, like stratospheric aerosol injection (SAI), where regional consequences may play out beyond the zone of technology deployment. Closely related to the issue of deploying SAI was a call, still, to consider the alternative—the negative impacts of non-deployment.

Participants reinforced the importance of establishing governance regimes commensurate with the scale of SRM challenges. Such systems of governance might include international agreements to address decision-making procedures that strongly attend to unequal power relations (either across nations or between large multinational private actors and public entities). Agreements might also be considered within this context around SAI use with expanded research and collaboration, so that all actors might better understand potential implications of SAI and CDR deployment and use. In addition, experts discussed the importance of parallel and related, empowered social dialogues (among civil society, small businesses, researchers, and publics) to articulate forward-looking, inclusive governance goals for CE. These concerns have been reflected in the above principles; the principles should also be read in light of them.

Addendum to CDR Guideline: Policy relevance of normative approach

The analysis presented here, on the basis of an updating of the Oxford and Tollgate principles, contains the foundations of an approach to defining the application of a precautionary principle to the domain of SRM research. A key question in determining the application of the precautionary principle to any domain is the question of the level of risk that can be considered societally tolerable. As EU communication documentation, as well as case law, have made clear, this is a constitutively political question that can only be determined through political channels (European Commission COM/2000/0001 final;).

Although it is not explicitly defined in the European Treaties, the precautionary principle as stated in EU communication documentation, simply stated, is that if there is a significant risk of harm to public health or the environment associated with some phenomenon, process or product, and there is no scientific consensus regarding it, that phenomenon, process or product should be regulated, up to and including full prohibition (European Commission COM/2000/0001 final). The key reason for the operation of the precautionary principle in the EU context is to ensure that a lack of scientific consensus is not allowed to present a barrier to necessary regulation, especially regulation for which there is a clear public demand.

The Tollgate principles' framing of SRM governance debate - in terms of the conditions under which institutions have the legitimate authority to institute SRM research and deployment - draws on the tradition of normative political theory. This tradition often proceeds by scrutinising whether given policy interventions, conceived as exercises of political power, can be justified according to given political-normative theories, including social contract theory, democratic theory, and the liberal theory of rights. This approach can be contrasted against approaches which seek to derive guidance for action from existing explicit or implicit norms in international practice, which proceed from the assumption that any action is permitted if it is not restricted.

As a principle which explicitly requires a political determination of its applicability conditions, as well as its effects, the precautionary principle represents a kind of bridge between these two approaches. On the one hand, the application of the principle needs to be refined through the analysis of case law and arguments by analogy with other domains of practice. On the other hand, normative theory has a clear place in determining the application of the principle, insofar as normative theory constitutes an intervention in a necessary political debate. A key policy-relevant contribution of the present analysis is that it offers a normative theoretical grounding for an operationalization of the precautionary principle for the regulation of SRM research, which arises from a critical analysis of the Oxford and Tollgate principles.

If a political decision is made to apply the precautionary principle to questions of policy in relation to SRM research and deployment, much would need to be clarified. A prima-facie examination of the precautionary principle might lead to the view that it is unsuited to the governance of SRM, because an intuitive interpretation of the principle gives rise to a basic tension. Explaining this tension helps to clarify how a version of the principle with the proper political parameters can form the foundation of meaningful SRM governance.

The tension consists in the thought that, if we apply the principle to the assessment of the policy to pursue SRM research, we find that there are at least some significant risks of harm to the environment or public health associated with the development of SRM, together with a lack of scientific consensus, implying that the policy should be delayed. At the same time, if we apply the principle to the policy of not pursuing SRM research, or blocking SRM research, we

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

also find *that* policy associated with a significant risk of harm to public health and the environment, and the same lack of scientific consensus invoked in the previous case. Thus, the precautionary principle can apparently be invoked both to justify blocking SRM research, and to justify blocking the erection of barriers to SRM research.

Politically determined normative parameters for the application of the principle can, and should, be invoked to overcome this tension. First, we need to constrain the kinds of phenomenon to which the principle should be applied. As proposed in the updated guidelines, risk analysis, a practice which includes consideration of the precautionary principle, should be applied not to the isolated question of whether SRM research should be pursued, but to the question of which of a range of climate policy mixes should be pursued. This study should involve an analysis of a range of theoretical scenarios, including a range of projections for global and European socio-economic evolution, analogous to the Shared Socioeconomic Pathways used by the IPCC.

Crucially, however, these guidelines propose that there need to be normative constraints placed on the comparisons between such scenarios. First, it is not justifiable for a policymaker to consider a scenario in which a future deployment of SRM would be strongly preferred, if that is a scenario in which *they themselves* (for instance, the European Union itself) have failed, and continue to fail, in standing duties to protect human rights and the environment. This reflects an important insight of precautionary approaches to the regulation of risk, namely that in a risk management it is preferable that the decision-makers and the potential beneficiaries of the object of assessment should not be separate agents from those people exposed to risk, or else there is a structural tendency to dump risk on parties that have no recourse (this is a way of framing a "moral hazard" concern) (Hermansson & Hanson 2007).

Second, as well as the well-established requirement to politically determine the level of risk that is socially acceptable when applying the precautionary principle, this guideline proposes the need to also define the *kind* of risk that is socially acceptable. The policymaker conducting the assessment may not consider interventions that would constitute human rights violations, or would significantly harm the interests of parties not represented by the institution making the decision (e.g. the EU), in particular, under-represented substate minorities including indigenous people, and future generations. Furthermore, minimal standards for the protection of ecosystems and non-human animals must be defined.

Thirdly, there needs to be a safeguard in terms of distributive justice, consisting of some kind of minimal protection for the most vulnerable, which cannot be outweighed by predicted advantages across the rest of the vulnerability distribution. This would mean, for example, that a policy mix should not be considered if it is likely to subject the most vulnerable to significant risk of harm, even if everyone above some threshold level of vulnerability would very likely be better off under that policy mix.[3] Thus, if there was some level of risk that the most vulnerable would be worse off under worst-case-scenarios for some policy mix involving SRM than they would be under other available policy mixes that do not involve SRM, those policy mixes involving SRM should not be considered. This safeguard reflects the basic

[3] While this principle is largely inspired by Jonathan Wolff's Minimum Equity Principle (Wolff 1996, 2020), acknowledgement is also due to John Shepherd for ideas in private correspondence

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

intuition of fairness that the interests of the most vulnerable should not be traded-off against the interests of the better off. Policy mixes that involve compensation to ensure the worst off meet the threshold for benefit may be considered, but only if it is possible to ensure the parties being compensated are the same as the parties being exposed to risk.

These constraints are referred to in this guideline under the term *guardrails*. With, and only with, these guardrails in place, it may be appropriate to carry out a risk analysis that takes on board both the risks of policy mixes that include SRM research, and policy mixes that do not include SRM research. These guardrails embed a precautionary approach, which prevents an unconstrained risk-benefit analysis of SRM from invoking the risk of total climate breakdown to justify a virtually unbounded set of interventions, including interventions which are incompatible with basic justice, basic ecological standards, and reasonable risk management norms.

Proposed improvements for ABCs of Carbon Dioxide Removal

Next, the ABCs of CDR principles continue with the same juxtaposition as follows - the outlined principle is matched up with the identified gap.

ABCs principle guidelines	**Annotations for the refinement of the operational guidelines**	**Proposed refined operational guidelines**
a) CDR should be considered in NDCs (Nationally Determined Contribution)	**Relevance: Policymakers**	a) CDR should be considered in NDCs (Nationally Determined Contribution), relevant to policymakers
b) CDR policies should not weaken other mitigation efforts	Operational implications for **research-funders, policy-makers, technology providers and researchers**: at no point should CDR projects be justified as alternatives to emissions reductions efforts. This would include, for instance, when soliciting investment or grant funding, when conducting impact assessments, or in marketing to public or private partners.	b) CDR policies should not weaken other mitigation effort, and not be presented as an alternative

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

ABCs principle guidelines	Annotations for the refinement of the operational guidelines	Proposed refined operational guidelines
c) Resulting CDR efforts should be commensurate with the long-term collective mitigation ambition	TechEthos ethics analysis under D2.2 corroborates and augments this principle: it should be read as a response to the so-called moral hazard concern, especially about delayed mitigation, and in line with the UNFCCC norm of 'common but differentiated responsibilities and respective capabilities'. This is because it is mainly wealthy countries that will fund and develop CDR technologies. **Relevance: Policymakers.**	c) Resulting CDR efforts should be commensurate with the long-term collective mitigation ambition to both avoid delayed mitigation and privileging wealthy countries.
d) Policy mixes should include technology-transfer to help strengthen capacities for CDR	**Technology providers** should cooperate with **policymakers** to implement this principle	d) Policy mixes should foster cooperation between technology providers and policy-makers and include technology-transfer to help strengthen capacities for CDR.
e) Policy mixes should include international cooperation to improve CDR efficiency	Cooperation amongst stakeholders and wider reach to increase desirability of the technology. Also, this could include collaboration of ethics and policy making, for the purpose of the merging of governance and ethics.	e) Policy mixes should include international cooperation to improve CDR efficiency, and should foster the merging of governance and ethics to increase desirability of the technology.

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

ABCs principle guidelines	Annotations for the refinement of the operational guidelines	Proposed refined operational guidelines
f) Policy mixes should include international climate finance transfers to mobilize CDR.		f) Policy mixes should include international climate finance transfers to mobilize CDR.
g) Policies should ensure consistent accounting for CDR results applying conservative baselines and including leakage	Pursue greater international collaboration in relation to CDR to promote the standardisation of removal accounting and the enforcement of such standards (see TechEthos D6.2 Policy Brief on EU Legal Frameworks). **Technology providers** have a responsibility to ensure transparent accounting via cooperation with regulators.	g) Policies should promote international collaboration in order to ensure transparent and consistent accounting for CDR results applying conservative baselines and including leakage.
h) Policies should apply robust MRV methodologies including on leakage		h) Policies should apply robust MRV methodologies including on leakage
i) CDR policies should fulfil principles of inter- and intragenerational equity (e.g., Polluter Pays or Ability to Pay).	In the context of R&I ethics, operationally this principle implies that **researchers** are required to consider whether the allocation of benefits arising from research and innovation will accrue disproportionately to historic emitters. It will be difficult to justify partnerships with entities affiliated with fossil fuel companies if these entities stand to profit from carbon removal. In particular, it is difficult to justify the control of intellectual property associated with carbon removal technologies by fossil fuel interests.	i) CDR policies should fulfil principles of inter- and intragenerational equity (e.g., Polluter Pays or Ability to Pay) in order to avoid the benefits arising from research and innovation accruing disproportionately to historic emitters.

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

ABCs principle guidelines	Annotations for the refinement of the operational guidelines	Proposed refined operational guidelines
j) Efforts should internationally be differentiated per common-but-differentiated responsibilities		j) Efforts should internationally be differentiated per common-but-differentiated responsibilities
k) Efforts should internationally be differentiated by respective capacities and (national) circumstances.		k) Efforts should internationally be differentiated by respective capacities and (national) circumstances.
l) Policies should include a national determination of clear objectives, policies, and metrics for CDR	A transparency requirement could be emphasised to specify the contribution of CDR. Countries would have to specify the extent to which they plan to use CDR technologies to reach their Net Zero target. This would minimise the opportunity for 'ethics washing' which refers 1) to presenting an interest in ethics without substantially tackling ethical issues, and 2) agreeing to a lower level of regulation in order to avoid a more stringent regime of regulation (in D2.2).	l) Policies should include a national determination of clear objectives, policies, and metrics for CDR, including a transparency requirement concerning plans for Net Zero targets.
m) Policies should consider both short- and long-term effectiveness and efficiency	Governance is new and emerging so it is challenging to assess effectiveness. One way would be to include ongoing evaluation of the guidelines, as also suggested in the TEAeM framework (see TechEthos D5.1 - Enhancement of ethical frameworks and outline of detailed ethics framework).	m) Policies should consider both short- and long-term effectiveness and efficiency, and include guidelines for ongoing evaluation.

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

ABCs principle guidelines	Annotations for the refinement of the operational guidelines	Proposed refined operational guidelines
n) Policies should be procedurally just	The notion of procedural justice could be unpacked further to consider: ● A clarification about processes for public deliberation and participatory decision-making in the overall direction of CDR policy, in the context of mitigation and energy policy more broadly (see suggestion made for the Tollgate principle). ● Decisions about where to cite CDR facilities, in particular CCS facilities but also carbon removal activities in themselves.	n) Policies should be procedurally just, taking into account 1) public deliberation and participatory decision-making, 2)location for CDR facilities, 3) role of the public in the formulation of national emissions targets and of plans for reaching them with CDR, 4) the desirability of education for the public involved in the decision-making, 5) the incorporation of ethical results to ensure impact.

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

ABCs principle guidelines	Annotations for the refinement of the operational guidelines	Proposed refined operational guidelines
o) The policy design process should involve public participation and stakeholder involvement	● There could be more clarification about the role of public participation, and kind of methods of engagement in the formulation of national emissions targets, and clarification of the role of CDR in achieving these targets. ● The terms 'public' and 'stakeholder' are very broad and should be clarified. Educating the public in advance of consultation to possibly identify the knowledgeable stakeholders is desirable but not necessary. ● Ethical analysis needs to be more focused, there needs to be a pathway to incorporate the results to ensure impact, and ways of measuring the impact of the results. One approach is gamification which allows the public to get actually involved in consultation and ethics box-ticking becomes more difficult.	
p) Policies should contribute to sustainable development	**Policymakers:** sustainable development in this context should include the protection of fundamental rights. This is especially salient with respect to fundamental rights in relation to land use changes in biofuel supply chains (see TechEthos D6.2, Policy Briefs on EU Legal Frameworks). CDR should be viewed in the context of a suite of potential systemic interventions to counteract the effects of climate change, which straddle mitigation, adaptation and restoration, for example transportation, health, diet, agricultural, information, and food systems interventions. This concern responds to results of TechEthos consultation (see TechEthos D3.5) that stressed the importance of avoiding artificial forced-choice framing which forces us towards	p) Policies should contribute to sustainable development including the protection of fundamental rights.

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

ABCs principle guidelines	Annotations for the refinement of the operational guidelines	Proposed refined operational guidelines
	technological solutions. Solutions suggested as potentially useful to address these challenges include countering misinformation, empowering stakeholders and communities in developing countries, and considering broader ecological concerns as part of CDR and SRM conversations.	
q) Policies should prevent transboundary harm	**Policymakers**: This should involve, for instance, ensuring the adequacy of environmental liability frameworks (see TechEthos D6.2, Policy Briefs on EU Legal Frameworks).	q) Policies should prevent transboundary harm ensuring the adequacy of environmental liability frameworks
r) Policies should prioritize rectifying damage at source		r) Policies should prioritize rectifying damage at source
s) Policy designs should reflect multi-risk trade-offs including policy or technology failure risks as well as countervailing risks of omitting policy steps.	There needs to be a re-elaboration about how to understand risk and uncertainty for operational reasons . In addition to evaluating multi-risk trade-offs, policy designs should reflect norms for responsible decision-making in the face of uncertainty, most notably precautionary reasoning (as suggested in D2.2 - Identification and specification of potential ethical issues and impacts and analysis of ethical issues, in 4.4.3). Also, this may include an understanding of the risk posed by the effect of not doing something i.e. the counter-factual argument (see D5.1 TEAeM framework). This results in a complex risk/risk balance, ie. the risk of continuing to not use SRM vs the possible risks involved with SRM deployment.	s) Policy designs should reflect multi-risk trade-offs including policy or technology failure risks as well as countervailing risks of omitting policy steps, and reflect norms for precautionary reasoning.

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

ABCs principle guidelines	Annotations for the refinement of the operational guidelines	Proposed refined operational guidelines
t) Anticipation of longer-term CDR needs incl. toward net-zero or net-negative emissions targets		t) Anticipation of longer-term CDR needs incl. toward net-zero or net-negative emissions targets
u) Policy mixes should include strategies for preventing over-promise and under-delivery	Ethics washing can be one of the reasons for over-promise and under-delivery hence it should be specifically addressed, for example, by emphasising the importance of transparency (see the refinement proposed for principle L).	u) Policy mixes should include strategies for preventing over-promise and under-delivery and should require transparency to specifically address ethics washing.
v) Policies should include intermittent targets and policy objectives		v) Policies should include intermittent targets and policy objectives
w) Policies should be adapted upon missing intermittent targets and objectives.		w) Policies should be adapted upon missing intermittent targets and objectives.
x) Policies should involve increasingly specific targets for various CDR and emission reduction methods		x) Policies should involve increasingly specific targets for various CDR and emission reduction methods

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

ABCs principle guidelines	Annotations for the refinement of the operational guidelines	Proposed refined operational guidelines
y) Policies should reflect CDR methods' specificities	A requirement for clarity about terminology should be respected, as addressed in (TechEthos D6.2 Policy Brief on EU Legal Frameworks). Also, to ask for some kind of ethical analysis of proposals on whether to use nature-based or engineered-based forms of CDR. It is not actually clear whether nature-based forms of CDR are more ethically acceptable than engineering-based forms of CDR removals methods. The choice of method may have repercussions on biological diversity. However, TechEthos showed that when these kinds of technologies are nature-based, they sound more appealing to people. (Finding from the scenario 'Post-consumer societies and natural climate solutions' in D3.1). This principle can be refined through specificity in the form of specific best practice to be developed specific to CDR.	y) Policies should reflect CDR methods' specificities and best practice, and require clarity about terminology. It should promote ethical analysis of proposals on whether to use nature-based or engineered-based forms of CDR.
z) Policy ensembles should meet the needs of the targeted methods		z) Policy ensembles should meet the needs of the targeted methods

Table 5: Refined ABCs principle guidelines for CDR

The guidance on CDR was informed by expert consultation under TechEthos Deliverable 3.5, which highlighted the concern that a policy-level fixation on technological-fixes would neglect systemic, socially-driven responses to climate change (whether through transit, farming, energy, built-environment, lifestyle or any number of others). These are potent interventions in their own right, and should be viewed as part of a holistic climate response rather than simply as co-benefits cited to justify a given CDR policy.

A number of ethical issues follow from a focus on technological fixes. Using catastrophic forced-choice situations to make policies that push quick-acting, short-term technological fixes, represents a core ethical concern. Such an approach ignores potentially longer-lasting, more efficacious, non-technological and systemic interventions. Finally, approaching CDR through the lens of technological fixes means ignoring serious environmental harms and

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

human exploitation and harm not directly associated with levels of atmospheric carbon pollution—for example various forms of water, air, and land pollution or ocean acidification.

Solutions suggested as potentially useful to address these ethical issues require looking beyond technological fixes to climate change. As noted in the updated guidance, importance was placed on situating climate engineering technologies amidst a broader tapestry of interventions in carbon and pollution mitigation and reduction, and adaptation—for example transportation, health, diet, agricultural, information, and food systems interventions. Actively countering misinformation came up as an important component of this discussion. In addition, discussion revolved around empowering stakeholders and communities in developing countries to build the expertise to have informed and respected seats at decision-making tables. Participants discussed the importance of considering broader ecological concerns as part of CDR conversations; this recommendation also applied to SRM.

With respect to carbon capture and storage, social and ethical concerns arose related to carbon storage siting. These touched on whether vulnerable communities would be included and/or further disadvantaged in decision-making about where to site storage facilities for captured carbon. An additional concern relates to abuse of political economic power; for example, of multinational fossil energy companies potentially standing to profit from removal of the very pollution they profited from emitting into the atmosphere (to say nothing of government economic subsidies enacted to enable such pollution). Again the above principles endeavour to reflect these concerns, and should be read in light of them.

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

3.2 Digital Extended Reality

3.2.1 Introduction

D2.1 in TechEthos indicates that a number of scholars have highlighted a problem concerning ethical guidelines in Extended Reality and described the existing gap in regulation (Birckhead et al., 2019; Spiegel, 2018; Vaidyam et al., 2019; Zhou et al., 2019). For example, Birckhead et al (2019) believes that the state of current clinical VR research is heterogeneous and describes it as a “Wild West'' with a lack of clear guidelines and standards. Thus, the main gap in existing XR guidelines overall is the lack of a generalist approach to dXR ethics. This lack has been at least partially addressed in D2.2 by performing the ethical analysis of dXR. However, the outcomes have not been fully operationalised in the form of guidelines or an ethical framework. Therefore, under the guidance of the intra-consortium expert consultation with a dXR specialist Alexei Grinbaum, it is suggested to employ an existing generalist ethical self-assessment framework for AI regulation from the HLEG called "Assessment List for Trustworthy Artificial Intelligence" (ALTAI) and transform it to bear on XR by using D2.2 outcomes.

In addition, TechEthos approach could be used to identify and amend gaps in NLP-oriented private enterprise ethical guidelines. This would offer an opportunity for an intersectoral analysis that is readily operationalisable. Microsoft's Guidelines for Human-AI Interaction are a compendium of 18 principles that provide pragmatic recommendations for developers and designers of AI systems, focusing on aspects such as user engagement, system behaviour, and error management (note these MS guidelines can be seen as rather techno-centric without sufficient accountability accepted by the producer). The TechEthos approach extends beyond the practicalities of AI system design to include reflections and considerations about the broader impacts of AI on society, such as issues of fairness, privacy, and power asymmetries. Given their distinct yet compatible areas of focus, the Microsoft Guidelines and the TechEthos approach can be united by identifying gaps amending the guidelines.

This integrated approach could also promote a more dynamic conversation between the private and academic sectors in AI ethics, encouraging mutual learning and collaboration. It recognises that both practical and theoretical ethical considerations are vital for responsible AI development and usage.

3.2.2 Selected guidelines for dXR

Below we present the ALTAI Guidelines which we used as a starting point for the refinement of the first guideline for digital extended reality.

Original ALTAI Guidelines (shortened)

(European Commission. Directorate General for Communications Networks, Content and Technology., 2020)

1 Human Agency and Oversight

> AI systems should support human agency and human decision-making, as prescribed by the principle of respect for human autonomy. This requires that AI systems should both: act as enablers for a democratic, flourishing and equitable society by supporting the user’s agency; and uphold fundamental rights, which should be underpinned by human oversight.

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

In this section AI systems are assessed in terms of their respect for human agency and autonomy as well as human oversight.

Human Agency and Autonomy

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

Human Oversight

This … helps to self-assess necessary oversight measures through governance mechanisms such as human-in-the-loop (HITL), human-on-the-loop (HOTL), or human-in- command (HIC) approaches. Human-in-the-loop refers to the capability for human intervention in every decision cycle of the system. Human-on-the-loop refers to the capability for human intervention during the design cycle of the system and monitoring the system's operation. Human-in-command refers to the capability to oversee the overall activity of the AI system (including its broader economic, societal, legal and ethical impact) and the ability to decide when and how to use the AI system in any particular situation. The latter can include the decision not to use an AI system in a particular situation to establish levels of human discretion during the use of the system, or to ensure the ability to override a decision made by an AI system.

2 Technical Robustness and Safety

A crucial requirement for achieving Trustworthy AI systems is their dependability (the ability to deliver services that can justifiably be trusted) and resilience (robustness when facing changes). Technical robustness requires that AI systems are developed with a preventative approach to risks and that they behave reliably and as intended while minimising unintentional and unexpected harm as well as preventing it where possible. This should also apply in the event of potential changes in their operating environment or the presence of other agents (human or artificial) that may interact with the AI system in an adversarial manner. The questions in this section address four main issues: 1) security; 2) safety; 3) accuracy; and 4) reliability, fall-back plans and reproducibility.

3 Privacy and Data Governance

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

Privacy

This … helps to self-assess the impact of the AI system's impact on privacy and data protection, which are fundamental rights that are closely related to each other and to the

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

fundamental right to the integrity of the person, which covers the respect for a person's mental and physical integrity.

Data Governance

This … helps to self-assess the adherence of the AI system ('s use) to various elements concerning data protection.

4 Transparency

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

Traceability

This … helps to self-assess whether the processes of the development of the AI system, i.e. the data and processes that yield the AI system's decisions, is properly documented to allow for traceability, increase transparency and, ultimately, build trust in AI in society.

Explainability

This … helps to self-assess the explainability of the AI system. The questions refer to the ability to explain both the technical processes of the AI system and the reasoning behind the decisions or predictions that the AI system makes. Explainability is crucial for building and maintaining users' trust in AI systems. AI driven decisions – to the extent possible – must be explained to and understood by those directly and indirectly affected, in order to allow for contesting of such decisions. An explanation as to why a model has generated a particular output or decision (and what combination of input factors contributed to that) is not always possible. These cases are referred to as 'blackboxes' and require special attention. In those circumstances, other explainability measures (e.g. traceability, auditability and transparent communication on the AI system's capabilities) may be required, provided that the AI system as a whole respects fundamental rights. The degree to which explainability is needed depends on the context and the severity of the consequences of erroneous or otherwise inaccurate output to human life.

Communication

This … helps to self-assess whether the AI system's capabilities and limitations have been communicated to the users in a manner appropriate to the use case at hand. This could encompass communication of the AI system's level of accuracy as well as its limitations.

5 Diversity, Non-discrimination and Fairness

In order to achieve Trustworthy AI, we must enable inclusion and diversity throughout the entire AI system's life cycle. AI systems (both for training and operation) may suffer from the inclusion of inadvertent historic bias, incompleteness, and bad governance models. The continuation of such biases could lead to unintended (in)direct prejudice and discrimination against certain groups or people, potentially exacerbating prejudice and marginalisation. Harm can also result from the intentional exploitation of (consumer) biases or by engaging in unfair competition, such as the homogenisation of prices by means of collusion or a non-transparent market. Identifiable and discriminatory bias should be removed in the collection phase where possible. AI systems should be user-centric and designed in a way that allows all people to use AI products or services, regardless of their age, gender, abilities or characteristics. Accessibility to this technology for persons with disabilities, which are present in all societal groups, is of particular importance.

Accessibility and Universal Design

Particularly in business-to-consumer domains, AI systems should be user-centric and designed in a way that allows all people to use AI products or services, regardless of their age, gender, abilities or characteristics. Accessibility to this technology for persons with disabilities, which are present in all societal groups, is of particular importance. AI systems should not have a one-size-fits-all approach and should consider Universal Design principles addressing the widest possible range of users, following relevant accessibility standards. This will enable equitable access and active participation of all people in existing and emerging computer-mediated human activities and with regard to assistive technologies.

Stakeholder Participation

In order to develop Trustworthy AI, it is advisable to consult stakeholders who may directly or indirectly be affected by the AI system throughout its life cycle. It is beneficial to solicit regular feedback even after deployment and set up longer term mechanisms for stakeholder participation, for example by ensuring workers information, consultation and participation throughout the whole process of implementing AI systems at organisations.

6 Societal and Environmental Well-being

In line with the principles of fairness and prevention of harm, the broader society, other sentient beings and the environment should be considered as stakeholders throughout the AI system's life cycle. Ubiquitous exposure to social AI systems in all areas of our lives (be it in education, work, care or entertainment) may alter our conception of social agency, or negatively impact our social relationships and attachment. While AI systems can be used to enhance social skills, they can equally contribute to their deterioration. This could equally affect peoples' physical and mental well-being. The effects of AI systems must therefore be carefully monitored and considered. Sustainability and ecological responsibility of AI systems should be encouraged, and research should be fostered into AI solutions addressing areas of global concern, for instance the Sustainable Development Goals.32 Overall, AI should be used to benefit all human beings, including future generations. AI systems should serve to maintain and foster democratic processes and respect the plurality of values and life choices of individuals. AI systems must not undermine democratic processes, human deliberation or democratic voting systems or pose a systemic threat to society at large.

Environmental Well-being

This … helps to self-assess the (potential) positive and negative impacts of the AI system on the environment. AI systems, even if they promise to help tackle some of the most pressing societal concerns, e.g. climate change, must work in the most environmentally friendly way possible. The AI system’s development, deployment and use process, as well as its entire supply chain, should be assessed in this regard (e.g. via a critical examination of the resource usage and energy consumption during training, opting for less net negative choices). Measures to secure the environmental friendliness of an AI system’s entire supply chain should be encouraged.

Impact on Work and Skills

AI systems may fundamentally alter the work sphere. They should support humans in the working environment, and aim for the creation of meaningful work. This subsection helps self-assess the impact of the AI system and its use in a working environment on workers, the relationship between workers and employers, and on skills.

Impact on Society at large or Democracy

This … helps to self-assess the impact of an AI system from a societal perspective, taking into account its effect on institutions, democracy and society at large. The use of AI systems should be given careful consideration, particularly in situations relating to the democratic processes, including not only political decision-making but also electoral contexts (e.g. when AI systems amplify fake news, segregate the electorate, facilitate totalitarian behaviour, etc.).

7 Accountability

The principle of accountability necessitates that mechanisms be put in place to ensure responsibility for the development, deployment and/or use of AI systems. This topic is closely related to risk management, identifying and mitigating risks in a transparent way that can be explained to and audited by third parties. When unjust or adverse impacts occur, accessible mechanisms for accountability should be in place that ensure an adequate possibility of redress.

Auditability

This … helps to self-assess the existing or necessary level that would be required for an evaluation of the AI system by internal and external auditors. The possibility to conduct evaluations as well as to access records on said evaluations can contribute to Trustworthy AI. In applications affecting fundamental rights, including safety-critical applications, AI systems should be able to be independently audited. This does not necessarily imply that information about business models and intellectual property related to the AI system must always be openly available.

Risk Management

Both the ability to report on actions or decisions that contribute to the AI system's outcome, and to respond to the consequences of such an outcome, must be ensured. Identifying, assessing, documenting and minimising the potential negative impacts of AI systems is especially crucial for those (in)directly affected. Due protection must be available for whistle-blowers, NGOs, trade unions or other entities when reporting legitimate concerns about an AI system.

When implementing the above requirements, tensions may arise between them, which may lead to inevitable trade-offs. Such trade-offs should be addressed in a rational and methodological manner within the state of the art. This entails that relevant interests and values implicated by the AI system should be identified and that, if conflict arises, trade-offs should be explicitly acknowledged and evaluated in terms of their risk to safety and ethical principles, including fundamental rights. Any decision about which trade-off to

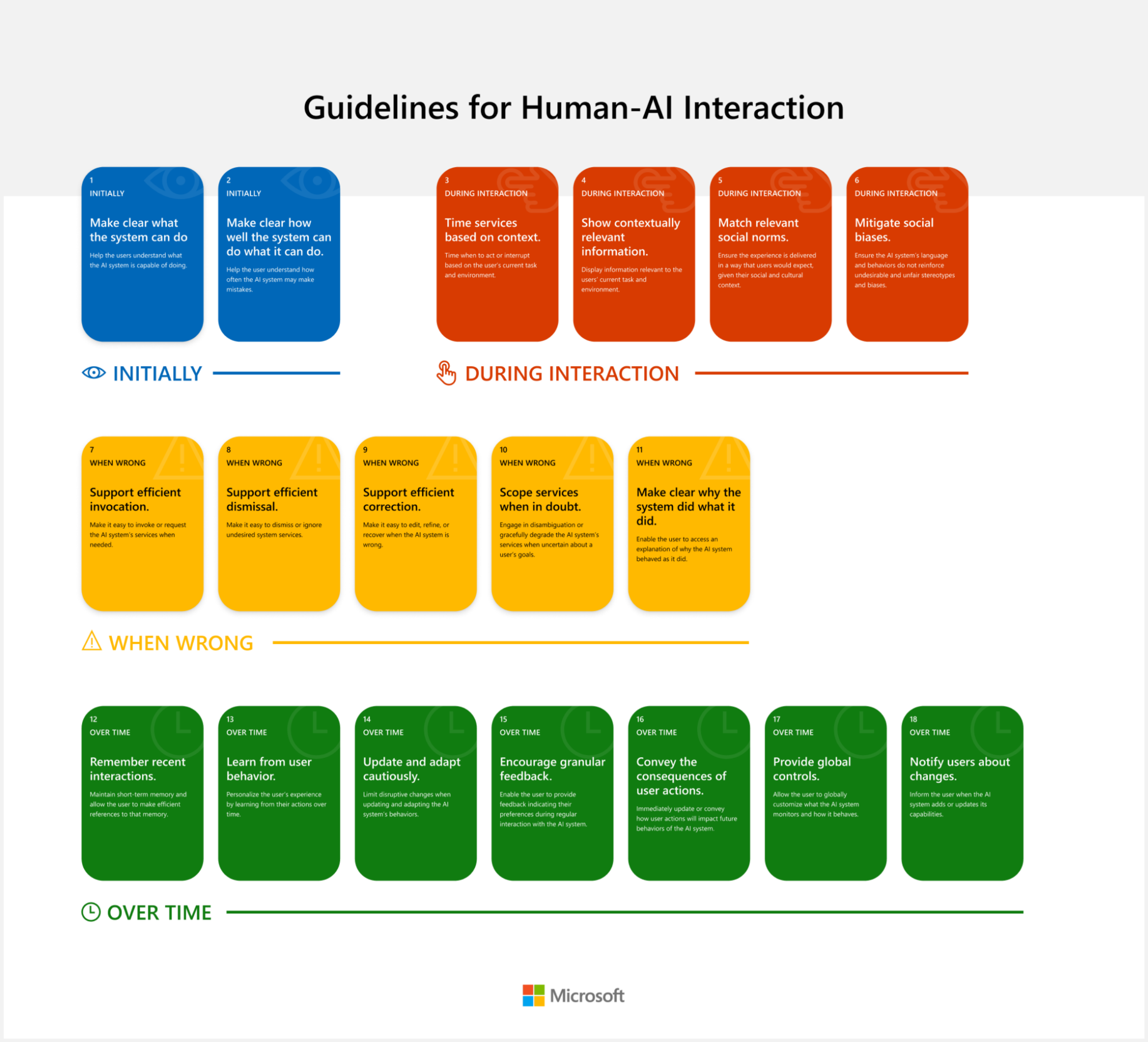
This project has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement No.101006249.

make should be well reasoned and properly documented. When adverse impact occurs, accessible mechanisms should be foreseen that ensure adequate redress.

Original Microsoft Guidelines for Human-AI Interaction

(https://www.microsoft.com/en-us/research/project/guidelines-for-human-ai-interaction/)

Below we present the second starting set of guidelines for digital extended reality, using the Microsoft Guidelines for Human-AI Interaction.



Guidelines for Human-AI Interaction

1 INITIALLY

Make clear what the system can do

Help the users understand what the AI system is capable of doing.

2 INITIALLY

Make clear how well the system can do what it can do.

Help the user understand how often the AI system may make mistakes.

3 DURING INTERACTION

Time services based on context.

Time when to act or interrupt based on the user's current task and environment.

4 DURING INTERACTION

Show contextually relevant information.

Display information relevant to the users' current task and environment.

5 DURING INTERACTION

Match relevant social norms.

Ensure the experience is delivered in a way that users would expect, given their social and cultural context.

6 DURING INTERACTION

Mitigate social biases.

Ensure the AI system's language and behaviors do not reinforce undesirable and unfair stereotypes and biases.

INITIALLY

DURING INTERACTION

7 WHEN WRONG

Support efficient invocation.

Make it easy to invoke or request the AI system's services when needed.

8 WHEN WRONG

Support efficient dismissal.

Make it easy to dismiss or ignore undesired system services.

9 WHEN WRONG

Support efficient correction.

Make it easy to edit, refine, or recover when the AI system is wrong.

10 WHEN WRONG

Scope services when in doubt.

Engage in disambiguation or gracefully degrade the AI system's services when uncertain about a user's goals.

11 WHEN WRONG

Make clear why the system did what it did.

Enable the user to access an explanation of why the AI system behaved as it did.

WHEN WRONG

12 OVER TIME

Remember recent interactions.

Maintain short-term memory and allow the user to make efficient references to that memory.

13 OVER TIME

Learn from user behavior.

Personalize the user's experience by learning from their actions over time.

14 OVER TIME

Update and adapt cautiously.

Limit disruptive changes when updating and adapting the AI system's behaviors.

15 OVER TIME

Encourage granular feedback.

Enable the user to provide feedback indicating their preferences during regular interaction with the AI system.

16 OVER TIME

Convey the consequences of user actions.

Immediately update or convey how user actions will impact future behaviors of the AI system.

17 OVER TIME

Provide global controls.

Allow the user to globally customize what the AI system monitors and how it behaves.

18 OVER TIME

Notify users about changes.

Inform the user when the AI system adds or updates its capabilities.

OVER TIME

Microsoft

Figure 6 Microsoft Guidelines for Human-AI Interaction (https://www.microsoft.com/en-us/research/project/guidelines-for-human-ai-interaction/)

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

AI Design Guidelines		Example Applications of Guidelines
1	**Make clear what the system can do.** Help the user understand what the AI system is capable of doing.	[Activity Trackers, Product #1] "Displays all the metrics that it tracks and explains how. Metrics include movement metrics such as steps, distance travelled, length of time exercised, and all-day calorie burn, for a day."
2	**Make clear how well the system can do what it can do.** Help the user understand how often the AI system may make mistakes.	[Music Recommenders, Product #1] "A little bit of hedging language: 'we think you'll like'."
3	**Time services based on context.** Time when to act or interrupt based on the user's current task and environment.	[Navigation, Product #1] "In my experience using the app, it seems to provide timely route guidance. Because the map up- dates regularly with your actual location, the guidance is timely."
4	**Show contextually relevant information.** Display information relevant to the user's current task and environment.	[Web Search, Product #2] "Searching a movie title returns show times in near my location for today's date"
5	**Match relevant social norms.** Ensure the experience is delivered in a way that users would expect, given their social and cultural context.	[Voice Assistants, Product #1] "[The assistant] uses a semi-formal voice to talk to you - spells out "okay" and asks further questions."
6	**Mitigate social biases.** Ensure the AI system's language and behaviors do not reinforce undesirable and unfair stereotypes and biases.	[Autocomplete, Product #2] "The autocomplete feature clearly suggests both genders [him, her] without any bias while suggesting the text to complete."
7	**Support efficient invocation.** Make it easy to invoke or request the AI system's services when needed.	[Voice Assistants, Product #1] "I can say [wake command] to initiate."

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

AI Design Guidelines		Example Applications of Guidelines
8	**Support efficient dismissal.** Make it easy to dismiss or ignore undesired AI system services.	[E-commerce, Product #2] "Feature is unobtrusive, below the fold, and easy to scroll past...Easy to ignore."
9	**Support efficient correction.** Make it easy to edit, refine, or recover when the AI system is wrong.	[Voice Assistants, Product #2] "Once my request for a reminder was processed I saw the ability to edit my reminder in the UI that was displayed. Small text underneath stated 'Tap to Edit' with a chevron indicating something would happen if I selected this text."
10	**Scope services when in doubt.** Engage in disambiguation or gracefully degrade the AI system's services when uncertain about a user's goals.	[Autocomplete, Product #1] "It usually provides 3-4 suggestions instead of directly auto completing it for you"
11	**Make clear why the system did what it did.** Enable the user to access an explanation of why the AI system behaved as it did.	[Navigation, Product #2] "The route chosen by the app was made based on the Fastest Route, which is shown in the subtext."
12	**Remember recent interactions.** Maintain short term memory and allow the user to make efficient references to that memory.	[Web Search, Product #1] "[The search engine] remembers the context of certain queries, with certain phrasing, so that it can continue the thread of the search (e.g., 'who is he married to' after a search that surfaces Benjamin Bratt)"
13	**Learn from user behaviour.** Personalise the user's experience by learning from their actions over time.	[Music Recommenders, Product #2] "I think this is applied because every action to add a song to the list triggers new recommendations."
14	**Update and adapt cautiously.** Limit disruptive changes when updating and adapting the AI system's behaviours.	[Music Recommenders, Product #2] "Once we select a song they update the immediate song list below but keeps the above one constant."

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

AI Design Guidelines		Example Applications of Guidelines
15	**Encourage granular feedback.** Enable the user to provide feedback indicating their preferences during regular interaction with the AI system.	[Email, Product #1] "The user can directly mark something as important, when the AI hadn't marked it as that previously."
16	**Convey the consequences of user actions.** Immediately update or convey how user actions will impact future behaviours of the AI system.	[Social Networks, Product #2] "[The product] communicates that hiding an Ad will adjust the relevance of future ads."
17	**Provide global controls.** Allow the user to globally customize what the AI system monitors and how it behaves.	[Photo Organizers, Product #1] "[The product] allows users to turn on your location history so the AI can group photos by where you have been."
18	**Notify users about changes.** Inform the user when the AI system adds or updates its capabilities.	[Navigation, Product #2] "[The product] does provide small in-app teaching callouts for important new features. New features that require my explicit attention are pop-ups."

Table 6: AI Design Guidelines and their applications

3.2.3 dXR guidelines' gaps

The main gap in existing XR guidelines overall is the lack of a generalist approach to dXR ethics. This lack has been at least partially addressed in D2.2 by performing the ethical analysis of dXR. Thus we used TechEthos analysis of ethical principles (D2.2) to propose improvements to the guidelines in that regard.

The potential uses of XR for influencing behaviour raise ethical concerns around autonomy, privacy, and consent (Adomaitis et al., 2022). For example, XR can influence human behaviour, thought and belief due to its immersive nature, interactivity, and ability to simulate real-world experiences. This can include altering perception, inducing emotions, prompting users to take certain actions, or shaping beliefs and attitudes. XR can also be used to create persuasive environments that nudge users towards certain behaviours or choices, such as encouraging healthy habits or environmental conservation.

In the ethical analysis (Adomaitis et al., 2022), TechEthos underlined the importance of physical safety. Users can harm themselves, other users, or the system they are interacting with. The last case consists in performing adversarial attacks or other malicious actions that lower or distort the

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

functioning of the system. Examples of such actions include evasion attacks, data poisoning attacks, model replication, and penetration (backdoor) attacks.

With regards to digital safety, TechEthos showed that dXR raises critical privacy concerns: 1) Eye tracking is used to adjust the XR system to match the user's gaze and provide a more immersive experience. However, eye tracking also raises privacy concerns, as it can reveal their immediate reactions, preferences, and desires. 2) dXR uses sensors and cameras to scan a user's home environment and create its digital replica. Home environment scanning could potentially reveal sensitive information about a user's home life, such as their living situation, income level, or family relationships. It could also reveal sensitive information about the user's possessions, such as valuables or confidential documents. Referring to individualised learned information can lead to nudging.

Furthermore, there are important 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.

In using dXR systems, unfair bias can arise due to the assumptions of the developers, the data used to train the algorithms, or the user interactions with dXR. dXR can contribute to biases by perpetuating stereotypes, reinforcing existing power imbalances, or excluding certain groups of people.

Due to the immersive and interactive nature of dXR, users can have a heightened emotional response to the experiences they encounter. dXR can also be used to perpetuate toxic or harmful behaviours, such as harassment or discrimination, particularly when anonymity or perceived social distance is present. We also show that it is important to distinguish cognitive and emotional effects on individuals. For example, knowing that one is conversing with a machine does not stop emotional projection (Adomaitis et al., 2022). People can project competences to avatars, which is ungrounded. Thus, it is important to establish the limits of what virtual avatars can and cannot do.

A concern for environmental well-being is fundamental. dXR technologies require computing equipment and servers, which can have a significant carbon footprint and contribute to climate change.

Accountability is also a key ethical concern. In situations where dXR technologies are used for critical applications such as healthcare or transportation, it may be difficult to establish who is legally responsible in case of accidents or malfunctions.

Lastly, in the Microsoft Guidelines for Human-AI Interaction we noted that different areas of concern were grouped under principle 7) Societal and Environmental Well-being, so we split the principle in two areas separating Societal concerns from Environmental Well-being, by adding a new guideline concerning societal impact, specifically relating to training and skills-transfer from dXR to physical reality.

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

3.2.4 Proposed improvements to guidelines (XR)

Refined ALTAI Guidelines

ALTAI Guidelines do not focus on XR in particular. The following TechEthos content can be used to suggest improvements to the current operational guidelines in that regard:

ALTAI Guidelines	Annotations for the refinement of the operational guidelines	Proposed refined operational guidelines
1. Human Agency and Oversight.	The potential uses of dXR for influencing behaviour raise specific ethical concerns around autonomy and non manipulation. There is a need for limiting nudging and subliminal manipulation in the metaverse.	Human Agency and Oversight to raise specific ethical concerns around autonomy and non manipulation.
2. Technical Robustness and Safety.	dXR systems must be designed with user safety in mind, and include appropriate safeguards to prevent injury or harm to users. This can include physical safety measures, such as ensuring that users do not trip or fall while using dXR systems, as well as health/psychological safety, such as ensuring that users are not exposed to harmful stimuli or content. Since dXR are IoT devices, cybersecurity concerns and measures in place for IoT should also be observed for dXR.	Technical Robustness and Safety - with user safety and include appropriate safeguards including physical safety measures.
3. Privacy and Data Governance.	In dXR systems, Privacy and Data Governance are different from AI because since dXR is a form of mixed reality, data collection and storage cannot just be dealt with with one-time consent click. Furthermore, since dXR devices still collect types of data that the user is not aware s/he is generating and that cannot be consciously controlled, privacy and responsible data governance structures specific to dXR should be in place.	Privacy and Data Governance - data collection and storage cannot just be dealt with with one-time consent click. Privacy and responsible data governance structures specific to dXR should be in place.
4. Transparency.	dXR users should know whether they are interacting with a human or machine avatar, and which parts of the environment are digitally augmented to limit risky attributions of emotions and competence.	Transparency - users should know whether they are interacting with a human or machine avatar

ALTAI Guidelines	Annotations for the refinement of the operational guidelines	Proposed refined operational guidelines
5. Diversity, Non-discrimination and Fairness.	dXR should not or must not contribute to biases by perpetuating stereotypes, reinforcing existing power imbalances, or excluding certain groups of people. dXR must be flexible enough to accommodate a wide range of human users and ensure accessibility. This is poignant due to the possibility of exclusion through automatisation of certain tasks in dXR. Responsible use and accountability (as multiple actors involved) must be ensured to minimise power imbalance & biases. Clarification is needed for who is accountable for undesirable social impacts that the technology might create, such as bias/stereotypes. The user must know and understand the objective and purpose of the dXR and the data sets used for its training.	Diversity, Non-discrimination and Fairness - dXR should not or must not contribute to biases by perpetuating stereotypes. The system should be flexible enough to accommodate a wide range of human users and ensure accessibility. Responsible use and accountability (as multiple actors involved) must be ensured to minimise power imbalance & biases.
6. Societal impact.	Societal impact of dXR involves training and skills-transfer. The process of training with dXR assumes that skills acquired via virtual experience are equivalent or transferable to material conditions. (from D2.2) Therefore there is a need for evaluating the relevance of dXR training for developing real-world skills.	Societal impact - there is a need for evaluating the relevance of dXR training for developing real-world skills.
7. Environmental Well-being.	XR technologies require computing equipment and servers, which may have a significant carbon footprint and contribute to climate change. There are currently no reliable, standard measures to quantify the environmental impact of dXR and evaluate its carbon footprint, therefore metrics need to be developed. dXR systems should also be designed to be as environmentally friendly as possible at the stage of production. Where resources (eg. rare earth metals) are necessary, they must not contribute to human suffering.	Environmental Well-being dXR systems should also be designed to be as environmentally friendly as possible at the stage of production. Where resources (eg. rare earth metals) are necessary, they must not contribute to human suffering.

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

ALTAI Guidelines	Annotations for the refinement of the operational guidelines	Proposed refined operational guidelines
8. Accountability.	Where dXR technologies are used for critical applications it may be difficult to establish who is responsible in case of accidents or malfunctions. Accountability should emphasise the importance of a) traceability - tracing the pathway of events to the system provider, and b) sharing responsibility between different types of actors.	Accountability should emphasise the importance of a) traceability - tracing the pathway of events to the system provider, and b) sharing responsibility between different types of actors.

Table 7: Refined ALTAI guidelines for XR

Refined Microsoft Guidelines for Human-AI Interaction

Microsoft Guidelines for Human-AI Interaction can benefit from enhancement from TechEthos analysis in the following areas:

Microsoft Guidelines for Human-AI Interaction	Annotations for the refinement of the operational guidelines	Proposed refined operational guidelines
1. **Make clear what the system can do.** Help the user understand what the AI system is capable of doing.		Provide clear instructions to users about a system's capabilities.
2. **Make clear how well the system can do what it can do.** Help the user understand how often the AI system may make mistakes.	To recognise the computationally-constructed anthropomorphisation in dXR and reduce the potential for manipulation, The provider must apply specific control and awareness mechanisms and establish the limits of what virtual avatars can and cannot do.	Use language to develop clear instructions that inform the user of the system avoiding anthropomorphisation, hyperbole and exaggeration of its capacities.

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

Microsoft Guidelines for Human-AI Interaction	Annotations for the refinement of the operational guidelines	Proposed refined operational guidelines
3. **Time services based on context.** Time when to act or interrupt based on the user's current task and environment.	dXR developers must remind users that they are in an dXR environment to ensure that they are aware of the boundaries between the virtual and real world, and to prevent any potential confusion or disorientation. This reminder can be particularly important in situations where the dXR environment is highly realistic or immersive, such as in medical simulations or training scenarios.	Provide alerts to users to support occupational health.
4. **Show contextually relevant information.** Display information relevant to the user's current task and environment.	Minimise potential for nudging arising from the timing of presentation of contextually-relevant information.	Giver user's greater agency of features of a system allowing personalisation to best suit intended tasks.
5. **Match relevant social norms.** Ensure the experience is delivered in a way that users would expect, given their social and cultural context.	Matching to relevant social norms can be further unpacked in terms of a) cultural relevance to particular groups and communities in which dXR is deployed; and b) alignment of the system with values and norms of human behaviour. Social and cultural context should be grounded in human rights legislation so that values and norms of human behaviour take into account the diversity of all populations.	Representational aspects of the system may rely on shared socio-cultural meanings. Give a range of options to allow user control over assets in a system.
6. **Mitigate social biases.** Ensure the AI system's language and behaviours do not reinforce undesirable and unfair stereotypes and biases.	dXR should not or must not contribute to biases by perpetuating stereotypes, reinforcing existing power imbalances, or excluding certain groups of people. dXR must be flexible enough to accommodate a wide range of human users and ensure accessibility. This is poignant due to the possibility of exclusion through automatisation of certain tasks in dXR.	Ensure that socio-cultural information in a system is not drawn from a narrow substratum of society. Where possible consult with a range of peoples, and err towards shared values of a society.

Microsoft Guidelines for Human-AI Interaction	Annotations for the refinement of the operational guidelines	Proposed refined operational guidelines
	Responsible use and accountability (as multiple actors involved) must be ensured to minimise power imbalance & biases. Clarification is needed for who is accountable for undesirable social impacts that the technology might create, such as bias/stereotypes. To avoid perpetuating stereotypes and biases, dXR creators should take steps to ensure that their content is diverse, inclusive, and reflective of a variety of perspectives.	
7. **Support efficient invocation.** Make it easy to invoke or request the AI system's services when needed.		**7. Support efficient invocation.** Make it easy to invoke or request the AI system's services when needed.
8. **Support efficient dismissal.** Make it easy to dismiss or ignore undesired AI system services.	There needs to be a clear and accessible "red button" option for the users of dXR for terminating tasks a) selectively, so that there is an option to terminate some, and not just all tasks, and b) temporarily, with the option of resuming them later.	Develop multiple closing down options. Save and delete options for users. Data, if requested by the user, should be permanently deleted.
9. **Support efficient correction.** Make it easy to edit, refine, or recover when the AI system is wrong.		**9. Support efficient correction.** Make it easy to edit, refine, or recover when the AI system is wrong.
10. **Scope services when in doubt.** Engage in disambiguation or gracefully degrade the AI system's services when uncertain about a user's goals.	All providers should be aware of similar cases of misuse and they should offer a clear policy for response, and take measures for prevention and controls.	Users should be informed of potential misuse of data and/or system.

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

Microsoft Guidelines for Human-AI Interaction	Annotations for the refinement of the operational guidelines	Proposed refined operational guidelines
11. **Make clear why the system did what it did.** Enable the user to access an explanation of why the AI system behaved as it did.		**11. Make clear why the system did what it did.** Enable the user to access an explanation of why the AI system behaved as it did.
12. **Remember recent interactions.** Maintain short term memory and allow the user to make efficient references to that memory.	Make users aware that nudging can occur when referring to individualised learned information.	Make explicit data processes, saving, sharing or deleting of user activity.
13. **Learn from user behavior.** Personalize the user's experience by learning from their actions over time.		**13. Learn from user behavior.** Personalize the user's experience by learning from their actions over time.
14. **Update and adapt cautiously.** Limit disruptive changes when updating and adapting the AI system's behaviors.	Before releasing AI systems and dXR software using ML components, designers should carry out comprehensive studies of emergent behaviour of such systems. The results of these tests should inform the design of control mechanisms of dXR systems.	Minimise regular updates.
15. **Encourage granular feedback.** Enable the user to provide feedback indicating their preferences during regular interaction with the AI system.	A specific method of evaluating granular feedback in applications for under represented, vulnerable or disabled groups is needed. An XR system should explain the captured feedback to the user and ask for their confirmation.	Provide relevant communication initiatives across user groups.

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

Microsoft Guidelines for Human-AI Interaction	Annotations for the refinement of the operational guidelines	Proposed refined operational guidelines
16. **Convey the consequences of user actions.** Immediately update or convey how user actions will impact future behaviors of the AI system.		**16. Convey the consequences of user actions.** Immediately update or convey how user actions will impact future behaviors of the AI system.
17. **Provide global controls.** Allow the user to globally customise what the AI system monitors and how it behaves.		**17. Provide global controls.** Allow the user to globally customise what the AI system monitors and how it behaves.
18. **Notify users about changes.** Inform the user when the AI system adds or updates its capabilities.		**18. Notify users about changes.** Inform the user when the AI system adds or updates its capabilities.

Table 8: Refined Microsoft Guidelines for Human-AI Interaction, for XR

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

3.3 Neurotechnology

3.3.1 Introduction

Neurotechnologies have rapidly developed in recent years and their potential to transform society is significant. However, the ethical implications of these emerging technologies are complex and require careful consideration. The OECD Recommendation on Responsible Innovation in Neurotechnologies (2019) and the Neuroethics Guiding Principles for the NIH BRAIN Initiative (2018) have been developed to address these ethical challenges. The OECD Recommendation highlights the importance of responsible innovation and ethical considerations in the development and deployment of neurotechnologies. The Neuroethics Guiding Principles for the NIH BRAIN Initiative provides a framework for ensuring the ethical use of neurotechnologies in research and clinical applications. As with climate engineering, the ethical considerations surrounding neurotechnologies are urgent and require thoughtful engagement from a broad range of stakeholders. The TechEthos project evaluated 26 various ethics frameworks and determined that the OECD Recommendation on Responsible Innovation in Neurotechnologies and Neuroethics Guiding Principles for the NIH BRAIN Initiative were both the most recent and relevant to neurotechnologies in terms of their comprehensiveness.

As with the two previous technology families, a mind-mapping exercise was conducted to identify relevant guidelines for research and development in neurotechnologies. The literature was categorised into "codes," "frameworks," and "guidelines/standards." Using expert consultation, we determined which of these documents best approximated current best practices. The process involved two stages. Firstly, we referred to the expert consultation exercise conducted under Task 3.4, which aimed to elicit responses to scenarios representing possible futures in the context of imagined research and innovation pathways. The objective of this exercise was to refine the scenarios to ensure they interrogate the most salient ethical intuitions as precisely as possible, and to identify concerns (in TechEthos Deliverable D3.6.) to be addressed through the improvement of existing operational guidelines.

Priority was given in the selection of recommendations for neurotechnologies to those that specifically address the special ethical issues raised by this area as opposed to those that address related issues in other fields. This strategy assisted in focusing the available regulations to those that are most pertinent to the unique difficulties of neurotechnologies. Although some of the documents found throughout the search may have some relevance to the larger discussion of technology and ethics, it's possible that they do not address the specifics and complexity of neurotechnologies. We were able to more effectively ensure that the subsequent work promotes engagement with the particular ethical considerations unique to NT while still being able to take a comprehensive overview of their broader ethical context by concentrating on guidelines that are explicitly relevant to neurotechnologies.

In the second stage of the shortlisting process for neurotechnologies, the TechEthos team consulted with internal experts to guide their selection. In light of the unique challenges presented by neurotechnologies, the team also considered the OECD Recommendation on Responsible Innovation in Neurotechnologies (2019) and the Neuroethics Guiding Principles for the NIH BRAIN Initiative (2018). These documents were specifically developed for the field of neurotechnologies and provide a more comprehensive framework for ethical considerations in this area. The expert consultation process ensured that the selected guidelines were relevant and appropriate to the unique ethical challenges posed by neurotechnologies.

The decision to start with the OECD Guideline on Responsible Innovation in Neurotechnologies was made for a number of reasons. First, the Organization for Economic Co-operation and Development (OECD), a reputable and globally acknowledged authority for establishing rules and recommendations for many industries, produced the guidelines. In a similar vein, the 2019 publication date of the guidelines indicates that they reflect contemporary developments in the field of neurotechnologies. The rules are also particular to neurotechnologies rather than a similar topic, which makes them more useful and relevant. Overall, the OECD Guideline on Responsible Innovation in Neurotechnologies provides a framework for responsible innovation that can direct additional conversations and advancements in the field, making it a solid place to start when considering improvement of guidelines for the field of neurotechnologies.

3.3.2 Selected Guidelines for Neurotechnologies

Below we present the OECD Recommendation on Responsible Innovation in Neurotechnologies which will be used as a starting point for the first guideline for neurotechnologies.

Original OECD Recommendation on Responsible Innovation in Neurotechnologies

(OECD, 2019):

1. Promoting responsible innovation to address health challenges
    a. First and foremost, promote beneficial applications of neurotechnologies.
    b. Integrate ethical considerations and take into account public values and concerns at the planning stage and design phase of technological development.
    c. Foster alignment of public support and economic incentives for neurotechnology innovation with the greatest health needs.
    d. Avoid harm, and show due regard for human rights and societal values, especially privacy, cognitive liberty, and autonomy of individuals.
    e. Prevent neurotechnology innovation that seeks to affect freedom and self-determination, particularly where this would foster or exacerbate bias for discrimination or exclusion.
    f. Encourage greater awareness of existing systems of oversight and, where appropriate, evaluate and work towards adapting existing laws and regulations for medical practice and research for application to activities involving neurotechnology.
2. Prioritising safety assessment
    a. Engage in communication among researchers, research participants, health professionals, patients, members of the public, private stakeholders, and government stakeholders to incorporate concepts of autonomy, harm reduction, safety into research prioritisation.
    b. Encourage early consideration of potential unforeseen side effects in the research and development of neurotechnologies.
    c. Promote market entrance based on sufficient evidence as to the safety, quality, and efficacy of new products and procedures as defined by relevant authorities.
    d. Establish mechanisms for both short-term and long-term oversight, monitoring, and reporting of product safety and security, including the implementation of rigorous safety and security standards.
3. Promoting inclusivity

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

   a. Strive to ensure neurotechnology is both developed for and available to those in need.
   b. Promote an enabling policy environment that advances the inclusion of underrepresented populations including, inter alia, social and economic populations, as well as sex- and age-specific groups, in neurotechnology research and development.
   c. Take into account the diversity of cultures and strive to minimise inequalities with respect to, inter alia, socio-economic, cultural norms, in the development and use of neurotechnology.

4. Fostering scientific collaboration
   a. Promote interdisciplinary research and development where communities of scientists and engineers interact closely with the social sciences and humanities communities as well as with user and other relevant groups.
   b. Foster pre-competitive consortia of collaborative research across public research institutions, private non-profit organisations, private sector entities, and patient communities.
   c. Support the development of standards and best practices for the technical as well as ethical, legal, and social aspects of innovation in neurotechnology.
   d. Support an international culture of "open science" by creating joint infrastructures and environments for sharing, aggregating, auditing, and archiving data relating to neurotechnology as appropriate.
5. Enabling societal deliberation
   a. Promote open communication across expert communities and with the public to promote neurotechnology literacy and the exchange of information and knowledge.
   b. Engage in multi-stakeholder dialogues and deliberation to ensure diverse inputs into decision making processes, public policy and governance
   c. Ensure that the results of formal dialogues are considered and taken into account in decision making wherever possible.
   d. Ensure processes for engaging stakeholders are fair, transparent, and predictable.
   e. Encourage transparent processes of technology appraisal to deepen and inform public debate about the longer-term trajectory of neurotechnology.
6. Enabling capacity of oversight and advisory bodies
   a. Encourage regulatory agencies, funding bodies, research institutions and/or private actors to respond to opportunities and ethical, legal and social issues raised by advances in brain research and neurotechnology.
   b. Encourage research into the ethical, legal and social dimensions of neurotechnology
   c. Promote the further development of ethical guidance and best practices including rigor and reproducibility.
   d. Ensure that oversight and advisory bodies possess appropriate multi-disciplinary expertise for constructive technology assessment, horizon scanning, scenario planning, and review of research.

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

e. Develop institutional capacity and mechanisms of technology appraisal and/or foresight to anticipate and evaluate potential neurotechnology outcomes and pathways.

7. Safeguarding personal brain data and other information
   a. Provide clear information to the public and research participants about the collection, storage, processing, and potential use of personal brain data collected for health purposes.
   b. Ensure that means of obtaining consent adequate to protect the autonomy of individuals are in place, including consideration of special cases of limited decision-making capacity.
   c. Promote opportunities for individuals to choose how their data are used and shared, including options for accessing, amending and deleting personal data.
   d. Promote policies that protect personal brain data from being used to discriminate against or to inappropriately exclude certain persons or populations, especially for commercial purposes or in the context of legal processes, employment, or insurance.
   e. Protect information gained through the application of neurotechnology from unauthorised use, including through the use of data access agreements when appropriate.
   f. Promote confidentiality and privacy and mitigate security breaches, including through the implementation of rigorous security standards.
   g. Ensure not only traceability of data collected and processed but also of medical acts in which neurotechnology is used.
8. Promoting cultures of stewardship and trust across the public and private sector
   a. Encourage development of best practices and business conduct that promote accountability, transparency, integrity, trustworthiness, responsiveness, and safety.
   b. Support innovative approaches to social responsibility through the development of accountability mechanisms.
   c. Foster communication in the public sphere that avoids hype, overstatement, and unfounded conclusions, both positive and negative, and that discloses interests in a transparent manner.
   d. Identify any issues, gaps, and challenges within systems of governance and explore possible solutions through dialogue among regulators, the private sector, and the public.
   e. Promote trust and trustworthiness through norms, and practices of responsible business conduct.
9. Anticipating and monitoring potential unintended use and/or misuse.
   a. Promote mechanisms to anticipate, and prevent, potentially harmful, short and long-term unintended uses and impacts before neurotechnologies are deployed.
   b. Implement safeguards and consider mechanisms to support integrity, autonomy, protection of private life, non-discrimination and dignity of the individual or of groups in the short and/or long term.
   c. Anticipate and prevent activities that seek to influence decision processes of individuals or groups by purposely affecting freedom and self-determination

through, for example, intrusive surveillance, unconsented assessment, manipulation of brain states and/or behaviour.

d. Where possible, take active steps to protect against potential misuse of neurotechnology.

Original Neuroethics Guiding Principles for the NIH BRAIN Initiative

(Greely et al., 2018)

Below we present the second starting set of guidelines for neurotechnologies, using the Neuroethics Guiding Principles for the NIH BRAIN Initiative.

This guideline applies to the process of developing the technology and making it as scientifically rigorous as possible. It complements the OECD approach which focuses on producing the best possible technological solution.

1. Make assessing safety paramount : Human subjects protections place the highest priority on research participant safety, including physical, psychological, and emotional consequences of research participation, in the short, intermediate, and long term. This is particularly important in neuroscience research because the complexity of the human brain lends unpredictability to outcomes of intervention and may heighten the likelihood and potential severity of unexpected consequences, including those emerging at later times because of the brain's plasticity. Safety also is crucial when implementing interventions for widespread clinical use in treating brain diseases and disorders. Safety can never be guaranteed, but risks must be rigorously assessed and carefully weighed against likely benefits in both research and treatment. The development of safe interventions depends on robust experimental design throughout the research pipeline, including adherence to the highest standards for rigour and reproducibility. Early-stage research with nonhuman model systems must be carefully designed to identify potential limitations during translational phases of research. For example, new methods of neuromodulation (invasive or otherwise) may create unanticipated interactions and reverberating consequences. New gene-editing technologies such as CRISPR/CAS, while offering hope for mitigating or eliminating brain disorders, are still in their infancy and carry potential for off-target effects. It is essential to attend to safety data from preclinical studies and to monitor safety throughout research when evaluating such innovative approaches for potential efficacy in humans. Research participants must be thoroughly informed of potential risks and benefits, as well as the possibility of unexpected safety issues.

2. Anticipate special issues related to capacity, autonomy, and agency: Contemporary neuroscience research may enable greater understanding of brain disorders associated with impaired, fluctuating, or diminished decision-making capacities and diminished agency (our ability to choose our actions) and autonomy (our ability to freely make informed choices). Some of these disorders may present in children, in whom these characteristics are also limited. Responsible BRAIN-funded research must study, not only “competent” and autonomous adults, but also people with diminished or developing autonomy and decision-making capacity. The challenges of a fair consent process that allows participation of those with limited, “different,” or fluctuating capacity to consent are not new but require constant attention. For example, in research with patients with Alzheimer's dementia, routine assessment of how well participants receive and process information

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

and their decision-making ability is crucial. This may prove especially challenging for patient participants with advanced forms of the disease or when research involves innovative techniques that may perturb capacity in ways unfamiliar to participants. Some interventions may lead to unanticipated changes in preferences and agency, as in reported personality changes after deep-brain stimulation for movement disorders (Lewis et al., 2015). In contrast, patients with neuropsychiatric conditions may actively seek such alterations to enhance their agency or restore capacities. Researchers may find themselves in the paradoxical position of seeking informed consent from participants while at the same time manipulating neural processes necessary for consent capacity and autonomous choice. For example, brain stimulation paradigms may target circuits involved in reward processing and motivation. Given our limited understanding of whether excessive stimulation might undermine patient participants' future decision making, how much control regarding stimulation parameters should go to participants in alignment with their autonomy interests rather than to researchers? Researchers should be particularly cautious to preserve and monitor research participants' ability to consent, including consent to continued participation in research. Providing participants with accurate, easy to understand, and evidence-based information about potential risks and benefits will promote well informed decisions about participation in neuroscience research (https://www.accessdata.fda.gov/scripts/cdrh/cfdocs/cfCFR/CFRSearch.cfm?fr=50.25). Care must be taken to avoid overpromising to possible participants, who may be desperate for a helpful intervention, and to discourage them from believing that personal benefits are more likely than they are.

3. Protect the privacy and confidentiality of neural data: Research participants have a reasonable expectation of privacy regarding their neural data and that data's interpretation, which might include perceptions, emotions, memories, and thoughts. NIH BRAIN Initiative research is developing better methods to measure brain structure and activity. These data will be stored for analysis and shared often with other researchers with appropriate privacy protections to advance efforts to understand the brain. Neural data should be treated as private, sensitive information; its collection, transmission, and storage should adhere to best practices for security and encryption. Conflicts may exist between privacy/confidentiality and data sharing. For example, large, shared databases containing brain imaging data may be extremely useful for researchers studying both healthy and atypical brain functioning, but every brain is unique and someday a brain MRI might be as identifying as a fingerprint. A person with access to a shared database, as well as an individual's MRI, might be able to determine whether the individual was in the database and, if so, obtain personal information about him or her from the "de-identified" database. It is important that researchers and policymakers find ways to manage these problems. Research participants' confidentiality cannot be guaranteed both because of the risks of unauthorised release of identified data through hacking and the possibilities of re-identification. Research participants must be given clear information about these issues and an honest chance to decide whether to accept the risks.

4. Attend possible malign uses of neuroscience tools and neurotechnologies: Novel tools and technologies, including neurotechnologies, can be used both for good ends and bad. Researchers should be mindful of possible misuses that might range from intrusive surveillance of brain states to efforts to incapacitate or impermissibly alter a person's behaviour. Researchers have a responsibility to try to predict plausible misuses and ensure that foreseeable risks are understood, as appropriate, by research participants, IRBs,

ethicists, and government officials. When possible, misuse should be prevented, for example, through design of the technology, such as ensuring secure wireless device connections.

5. Move neuroscience tools and neurotechnologies into medical or nonmedical uses with caution: BRAIN Initiative research includes cutting-edge science and first-in-human applications of novel neurotechnologies. Accordingly, the likelihood of individual benefit may be low and uncertain and risks could be significant. Researchers must thoroughly identify and minimise potential research risks. A thoughtful justification of risks based on the potential benefits is essential. Hopes for neuroscience extend beyond research into exciting prospects for novel therapeutics. In addition to safety, researchers should consider questions of efficacy and equity before novel neurotechnologies become widely available. Researchers and others involved in the NIH BRAIN Initiative should discourage the premature widespread use or inappropriate adoption of new technologies, especially those that may be offered directly to consumers or in non-health-care settings, such as in the legal system. For example, researchers looking for neural markers of deception or pain should be aware that segments of society may be eager to use such tools for non-health-related ends. Premature adoption of such tools before accuracy is known is not appropriate.

6. Identify and address specific concerns of the public about the brain: People care deeply about their minds and brains and have concerns that researchers may not sufficiently recognize. Even scientifically unjustified fears can have important consequences for public response to neuroscience work. Although sensitivity about brain-related issues varies between cultures, three examples follow. Fear of mental invasion reaches far back into human history, as does the idea of cognitive liberty—that the freedom and privacy of one's mind (and thus brain) is sacrosanct. Some might have concerns that a beneficial improvement in ability to control the dysfunctional mind (e.g., from memory loss or seizures) also may have detrimental outcomes and potentially threatens cognitive liberty (Ienca and Andorno, 2017). Second, many people perceive their identity as being within their brain. Novel neurointerventions might disrupt that identity; for example, brain implants might alter a persons' sense of self or change their behavior in unexpected or unwelcome ways (Gilbert et al., 2017). Researchers should be aware of these justified concerns that research could "make a person someone else." Last, many consider the human mind and brain to be distinguishing, perhaps definitive, features of being human. Research with human/nonhuman brain chimeras, neural organoids, and ex vivo human brain tissue can provoke intellectual, visceral, and moral concerns, including concerns about the potential development of morally significant features in these tissues. Researchers, funders, and others should try to identify issues arising from their research that the public might find sensitive, taking into account the possibility of sensationalised media reports. Both the public and researchers will benefit if the latter consider public concerns when planning, implementing, and discussing research, as described in the next principle.

7. Encourage public education and dialogue: Public trust in science is a precious commodity. To the greatest extent possible, researchers should build—and retain—that trust by keeping the public informed. Public dialogue should be bidirectional, where researchers stay abreast of the public's desires, concerns, and degree of knowledge. Some conflicts between informing the public about research as it proceeds and researchers' appropriate desires to delay sharing preliminary findings before appropriate review may be

unavoidable. Nevertheless, transparency is crucial, particularly with potentially controversial research, to avoid unduly concerning the public. Being a scientist today requires not only good work, but also good communication about that work. Modern society offers scientists a wide array of ways to communicate beyond the traditional peer-reviewed paper and academic conference talk. Good ethical stewardship of one's work calls on scientists to find methods that best suit them, whether through public talks, online scholarship, creating social media content, giving interviews, or other paths. Researchers have an obligation to share knowledge both about the brain and about where we continue to be ignorant about the brain's workings, along with possible benefits and risks of research. University and government communications offices also have a critical role to play in promoting transparency. Hyperbole is in part driven by the imaginations of scientists, the public, and neuroethicists and because hype about the next great breakthrough is widely used to hold attention. Researchers, science journalists, communications offices, and others—including neuroethicists—have essential roles to play in promoting appropriate understanding, avoiding hyperbole, and correcting overly optimistic interpretations.

8. Behave justly and share the benefits of neuroscience research and resulting technologies: The former Presidential Commission for the Study of Bioethical Issues wrote "… a fundamental principle of fairness suggests that society should seek to assure that the benefits and burdens of new technologies are shared" (https://bioethicsarchive.georgetown.edu/pcsbi/sites/default/files/PCSBI-Synthetic-Biology-Report-12.16.10_0.pdf). Early BRAIN Initiative studies are likely to be small and fairness in selection of research participants is critical because more people may want to participate than can be included given finite opportunities and participants with few options for treatment may be more open to untested options. For example, experiments testing visual prostheses may be very appealing to persons severely affected by vision loss. Similarly, the possible appeal of brain–machine interface experiments for those suffering from tetraplegia warrants careful processes for selecting early trial participants. As technologies are found to be safe and effective and enter clinical use, attention to widespread sharing of the benefits of those technologies and interventions will become a priority. Limited access to safe and effective neural technologies should not exacerbate existing health disparities or inequalities, but neither should the burdens of research fall disproportionately on those who lack access to established interventions.

3.3.3 Neurotechnologies guidelines' gaps

The predetermined template guidelines serve as high-level policy recommendations for neurotechnologies. There are currently no operational rules for R&I in these sectors that are properly developed. Therefore, the guidelines chosen represent best practice in terms of high-level guidance. Some of the listed principles have direct implications for the operational guidance of research and development, while for others, additional work is needed to operationalize the guidance for the R&I context in particular, differentiating this function from legislative and policy guidance.

There are various gaps and restrictions in the OECD Guideline on Responsible Innovation in Neurotechnologies, despite the fact that it is a thorough and significant collection of guidelines. For instance, the recommendations provide room for ambiguity and confusion over who should be held accountable for the use and misuse of neurotechnologies. In particular, the

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

principles of promoting responsible innovation to address health challenges is also articulated in a very general and high-level manner, which may make it challenging to put into reality. The rules only apply to the sphere of health, which is another drawback because ethical innovation for neurotechnologies goes beyond just solving health problems. It is essential to have a wider and more comprehensive perspective on responsible innovation that takes into account the different social, cultural, and ethical aspects involved in using neurotechnologies as they are employed in a variety of fields, including entertainment, defence, and education (Garden & Winickoff, 2018).

Concerning the prioritisation of safety assessment, such assessments are conducted as part of the anticipatory technology ethics (ATE) that the TechEthos project has adopted overall, and that has since been refined (Umbrello et al., 2023). In particular, the TechEthos project has highlighted the distinction between physical safety and digital safety (D2.2). The OECD Guideline, in particular the second guideline on prioritising safety assessments does not specify the kind of safety to be considered. More specifically, there is a focus only on product safety, which is *de facto* physical safety, rather than any explicit focus on digital safety.

The exercises on ethical considerations for neurotechnologies surfaced a number of relevant concerns. One key issue is the limited scope of extant principles, which tend to focus primarily on the physical risks associated with the use of these technologies, while not giving sufficient attention to the potential digital risks. Additionally, these principles often limit their scope of risks to the domain of health, rather than acknowledging the potential risks in other domains, such as entertainment or defence. Experts note that this exclusive focus on physical health risks neglects the broader social and ethical implications of neurotechnologies. As such, there is a need for a more holistic approach to the ethical considerations of these technologies that recognizes the potential for digital harms and considers the broader societal impact beyond just the domain of health.

The TechEthos cross-cutting principle of irreversibility is a crucial factor to take into account while developing and implementing neurotechnologies. Neurotechnology treatments are irreversible, especially when invasive techniques are involved. This has ethical ramifications for protecting individual brain data. Although the two recommendations offer crucial factors for moral neurotechnologies, they might not properly address the risk of permanent harm to people's privacy and autonomy. It is critical to take into account the potential negative effects of neurotechnology interventions and to set up measures to preserve personal data. Therefore, it is crucial to include the principle of irreversibility in the ethical standards for neurotechnologies, especially when it comes to protecting individual brain data and other information, to ensure that the creation and application of these technologies are consistent with moral principles and ethical values.

As mentioned above, irreversibility, one of the TechEthos cross-cutting principles, is one notable lacuna of the OECD guidelines. Irreversibility highlights the necessity to take into account the long-term effects of neurotechnological interventions and the possibility of irreparable harm. The OECD recommendations may not adequately address the dangers connected to neurotechnological procedures, especially those that are irreversible, due to the absence of this principle. The addition of the principle of irreversibility can address this gap by offering a principle on how to balance the risks of irreparable injury with the potential benefits of neurotechnological therapies.

The OECD Guidelines likewise do not explicitly address the distinction between the risk of doing something and the risk of not doing something, which is a potential gap in the guidelines. While the principle of safeguarding personal brain data and other information is critical, it only addresses the risk of intervention, such as unauthorised access, use, or disclosure of personal brain data. However, the risk of inaction, i.e., not using neurotechnologies in certain contexts, such as in medical treatment, could also be significant. For example, not using neurotechnology to diagnose or treat a neurological disorder could lead to a worsening of the patient's condition. Therefore, it is important for the guidelines to consider both the risks of intervention and non-intervention, and to provide guidance on how to balance these risks in different contexts.

Although the Neuroethics Guiding Principles for the NIH BRAIN Initiative offer a useful foundation for ethical considerations in the development of neurotechnology, they also present some coverage gaps. The lack of attention given to digital safety is one of the biggest gaps. Nevertheless, there is no mention of the possible risks associated with digital data and information, such as cybersecurity threats, data breaches, and privacy violations. The guidelines do stress the necessity of physical safety in the use of neurotechnologies.

Another gap in the Neuroethics Guiding Principles is the relative lack of nuance in the principle of safety. While safety is a critical consideration in the development and use of neurotechnologies, the guidelines do not provide much guidance on how to balance safety concerns with other ethical considerations, such as autonomy, privacy, and justice. The principle of safety is also relatively narrow in its focus, primarily addressing the safety of human subjects involved in neuroscience research and the safety of the broader public. The guidelines do not address safety concerns related to other potential impacts of neurotechnologies, such as social and cultural impacts or the potential for unintended consequences (Lynch, 2004).

3.3.4 Proposed improvements to Guidelines (NT)

We expanded upon the Neuroethics Guiding Principles for the NIH BRAIN Project and the OECD Guidelines on Responsible Innovation in Neurotechnologies (both applicable to neurotechnologies). Thus, using the consortium members' in-depth knowledge as well as the results of WP2 and WP3, we conducted an online workshop (10th July 2023) where we identified lacunae in these concepts. We juxtaposed the concepts included in the two current ethical guidelines for neurotechnologies with the gaps found through expert consultation. During the workshop we noted that to enhance the specificity of general principles, the refined guidelines should consider using a higher level approach followed by mid-level and then implementation tool kit ideas.

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

Revised OECD Guidelines on Responsible Innovation in Neurotechnologies

OECD Recommendation on Responsible Innovation in Neurotechnologies’ principle guidelines	Annotations for the refinement of the operational guidelines	Proposed refined operational guidelines
1) Promoting responsible innovation d) Avoid harm, and show due regard for human rights and societal values, especially privacy, cognitive liberty, and autonomy of individuals e) Prevent neurotechnology innovation that seeks to affect freedom and self-determination, particularly where this would foster or exacerbate bias for discrimination or exclusion	Currently there is no practical way to implement these principles into operational guidelines. However, the use of an ethics by design approach to translate this first principle into a more practical principle is helpful.	1) Promoting responsible innovation, using ethics by design d) Avoid harm, and show due regard for human rights and societal values, especially privacy, cognitive liberty, and autonomy of individuals e) Prevent neurotechnology innovation that seeks to affect freedom and self-determination, particularly where this would foster or exacerbate bias for discrimination or exclusion
2) c) Foster alignment of public support and economic incentives for neurotechnology innovation with the greatest health needs	Context is fundamental to the operational guidelines and therefore the principles would always need to be adapted depending on the particular emerging technology in question. Context for this principle should be specified in terms of actors and groups to clarify the criteria to evaluate significant and urgent health needs. Outcomes beyond therapeutic uses of the technology into the realm of	2) c) Foster alignment of public support and economic incentives for neurotechnology innovation with the greatest health needs, based around criteria for the evaluation of significant and urgent health needs and beyond, and in both the short and longer terms

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

OECD Recommendation on Responsible Innovation in Neurotechnologies' principle guidelines	**Annotations for the refinement of the operational guidelines**	**Proposed refined operational guidelines**
	enhancement should be made accessible across populations. Developers should work with socio-technology planners to consider the long-term viability of the technology and its maintenance. (specific case in D3.5, p. 18).	
3) Prioritising safety assessment d) Establish mechanisms for both short-term and long-term oversight, monitoring, and reporting of product safety and security, including the implementation of rigorous safety and security standards	This refined guideline needs to include distinction between physical risk and digital risk. In addition to the use of safety assessment to evaluate the physical risk to each of the actors involved, there should be a way to evaluate the extent of digital risk too.	3) Prioritising safety assessment d) Establish mechanisms for both short-term and long-term oversight, monitoring, and reporting of product physical and digital risk, safety and security, including the implementation of rigorous safety and security standards
4) Promoting inclusivity	The refined guideline should include the use of consultation exercises to promote stakeholder engagement with citizens and under represented groups.	4) Promoting inclusivity across all strata of society

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

OECD Recommendation on Responsible Innovation in Neurotechnologies' principle guidelines	Annotations for the refinement of the operational guidelines	Proposed refined operational guidelines
5) Fostering scientific collaboration a) Promote interdisciplinary research and development where communities of scientists and engineers interact closely with the social sciences and humanities communities as well as with user and other relevant groups	This principle needs to focus on co-design at an early stage of interaction. The refined guideline should consider the underlying role the public has on science and hence its collaborative role in creating research. This is an example of ethics by design whereby communities can help to identify issues early during the development stage of technologies The guideline should consider a 'Citizens science' approach where there is a dispersed form of knowledge-making involved. For example, amateur interest in science acts as a source of contribution to science. The refined guidelines can contain descriptive language and description of use cases to make the principle meaningful, for example the addition of pictures to words to simplify abstract language.	5) Fostering scientific collaboration, including co-design and using ethics by design throughout the process a) Promote interdisciplinary research and development where communities of scientists and engineers interact closely with the social sciences and humanities communities as well as with user and other relevant groups
6) Enabling societal deliberation b) Engage in multi-stakeholder dialogues and deliberation to ensure diverse inputs into decision making processes, public policy and governance	The refined guideline is about communicative action and the need to reflect on raising society's awareness around emerging technologies. Alongside consultation, capacity-building would be useful - so that participants have sufficient knowledge and understanding to inform and give meaningful contributions to the consultation process. One such way is gamification, as used in TechEthos (D3.2) to engage with citizens from different backgrounds.	6) Enabling societal deliberation to raise society's awareness around emerging technologies, using techniques such as gamification to illustrate ethical issues b) Engage in multi-stakeholder dialogues and deliberation to ensure diverse inputs into decision making

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

OECD Recommendation on Responsible Innovation in Neurotechnologies' principle guidelines	Annotations for the refinement of the operational guidelines	Proposed refined operational guidelines
	Ethical analysis needs to be more focused, there needs to be a pathway to incorporate the results to ensure impact, and ways of measuring the impact of the results. One approach is gamification which allows the public to get actually involved in consultation and ethics box-ticking becomes more difficult.	processes, public policy and governance
7) Enabling capacity of oversight and advisory bodies	This principle should include an element of informed consent that can be specifically used as a means for societal deliberation.	7) Enabling capacity of oversight and advisory bodies, including informed consent in the context of societal deliberation
8) Safeguarding personal brain data and other information	The guideline should include a specification of irreversibility. Irreversibility describes physical integrity for the short term, but it is difficult to define the effects of irreversibility for mental integrity and in the long term. There is a need to specify a useful timeframe for irreversibility because with NT, long term assessment is needed. In addition to specifying the irreversibility impact on individuals, there is a need to specify how irreversibility impacts on society in general. For example, societal impact can be in the risk of bias being programmed into the technology by developers, much like it has happened for AI.	8) Safeguarding personal brain data and other information, especially with respect to irreversibility in all its forms

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

OECD Recommendation on Responsible Innovation in Neurotechnologies' principle guidelines	Annotations for the refinement of the operational guidelines	Proposed refined operational guidelines
	Also, understanding of irreversibility in this guideline may include an understanding of the risk posed by the effect of not doing something i.e. the counter-factual argument (see D5.1 TEAeM framework).	
9) Promoting cultures of stewardship and trust across the public and private sector	Awareness needs to be given to the diverse aspects of populations, in terms of bodily, cognitive and neurodiversity.	9) Promoting cultures of stewardship and trust across the public and private sector, and across all populations in terms of bodily, cognitive and neurodiversity
10) Anticipating and monitoring potential unintended use and/or misuse. a) Promote mechanisms to anticipate, and prevent, potentially harmful, short and long-term unintended uses and impacts before neurotechnologies are deployed	There needs to be assessment methods for short term and long term impacts of physical as well as mental integrity.	10) Anticipating and monitoring potential unintended use and/or misuse, and assessment of impacts for mental and physical impacts in the short and longer terms a) Promote mechanisms to anticipate, and prevent, potentially harmful, short and long-term unintended uses and impacts before neurotechnologies are deployed

Table 9: Refined OECD Guidelines on Responsible Innovation in Neurotechnologies

Comparison of the 2 guidelines - they are complementary, OECD greater interest around ethics that help steer emerging technologies (NIH focused around researchers). NIH focuses around virtuous governance amongst different communities harmonised to do good governance work.

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

Revised Neuroethics Guiding Principles for the NIH BRAIN Project

Neuroethics Guiding Principles for the NIH BRAIN Initiative principle guidelines	**Annotations for the refinement of the operational guidelines**	**Proposed refined operational guidelines**
1) Make assessing safety paramount	The guideline could integrate participant safety (physical, psychological and emotion) with notions of digital safety. Also, the guideline could contemplate measures for the safety for non-human subjects, mentioned as 'non-human model systems' in this principle	1) Make assessing physical, psychological and emotion and digital safety for all subjects paramount
2) Anticipate special issues related to capacity, autonomy, and agency		2) Anticipate special issues related to capacity, autonomy, and agency
3) Protect the privacy and confidentiality of neural data		3) Protect the privacy and confidentiality of neural data
4) Attend to possible malign uses of neuroscience tools and neurotechnologies		4) Attend to possible malign uses of neuroscience tools and neurotechnologies
5) Move neuroscience tools and neurotechnologies into medical or nonmedical uses with caution	This principle can be applied potentially to other areas. So it would be good to clarify the points in which some of the principles become useful to other sectors as well. Even if they are meant for health and they approach other sectors, a recognition of trans-applicability can be made.	5) Move neuroscience tools and neurotechnologies into medical or nonmedical uses with caution and with respect to the specific ethical issues in the proposed other applied contexts

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

Neuroethics Guiding Principles for the NIH BRAIN Initiative principle guidelines	Annotations for the refinement of the operational guidelines	Proposed refined operational guidelines
6) Identify and address specific concerns of the public about the brain	Also, include a range of tools and techniques that could be used to address and alleviate the public concerns with ethical issues in NT research (social readiness tool in D5.6). Generic concepts of public & relevant stakeholders (e.g. patients) need to be refined	6) Identify and address specific concerns of the public about the brain, using appropriate tools and techniques
7) Encourage public education and dialogue	Dialogue between researchers and the public should be organised as a meaningful, respectful two way process, characterised by active listening.	7) Encourage public education and dialogue, via a meaningful two way process
8) Behave justly and share the benefits of neuroscience research and resulting technologies		8) Behave justly and share the benefits of neuroscience research and resulting technologies

Table 10: Refined Neuroethics Guiding Principles for the NIH BRAIN Project

4 Discussion

The task (5.2) was essentially to reflect on existing guidelines and make suggestions for improving these operational guidelines with greater ethical input, for our three selected technology families: 1) Climate Engineering 2) Digital Extended Reality and 3) Neurotechnologies.

Since the mid-2010s there has been a proliferation of ethical guidelines generated as a result of concerns about technological development and concerns to mitigate risks. There has been a rise in European Commission's funding to support 'ethics by design' practices (for example, a typical requirement on an EU funded project is an ethics work package (WP), and business initiatives in these areas. Ethical guidelines are typically produced by academics on funded research projects, but are also created in other contexts, by think tanks, charities and businesses (see for example Microsoft's Human-Computer Interaction guidelines examined in this proposal). This raises the question of legitimacy and status of guidelines. Who is producing the guidelines? What is the basis for legitimacy and recognition of particular guidelines? Why choose one guideline over another? As

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

TechEthos expert highlighted, The Tollgate guidelines, for instance, were produced by two scholars, and while they have been cited in academic circles they have no regulatory or statutory basis. Alternatively, an autonomous academic approach is crucial to provide an alternative to ethical codes produced by the vested interests of big business. That said, 'ethics washing' can be carried out by both academics and businesses.

There are also issues about the ontological status of guidelines. What ethical models underpin them? Ethically speaking there are a vast range of ethical positions, philosophies and viewpoints, and rather than regard them as fixed across time, they are variable, changing with different priorities, equality between men and women being one feature of transformative ethical politics. Subsequently, ethicists anchor their guidelines in core political values, some of which are codified in law (the right to privacy), while others constitute ethical values established as a result of shifting political priorities (response to global warming), or societal and personal effects (digital reality technologies).

TechEthos has drawn attention to the problem of 'ethics washing' or using ethics as a compliance tool. In TechEthos Deliverable D2.2 (2022) the TechEthos consortium noted the decision to treat substantive ethical values as unavoidable, which highlights the tendency of approaches which purport to be neutral but "rationalize the status quo" (Adomaitis et al., 2022, p. 22). Ethicists, as well as stakeholders could be held accountable here, as ethics is not a professional practice with certification and registration like the medical profession. A higher degree (PhD typically) in social science, including but not limited to, philosophy underpins the qualifications, begging the question 'who is watching the watchers?' Moreover, policy makers, with briefs to institutionalise ethics in technologies may prefer the formulaic approach to ethics, rather than its contradictory and sometimes conflicting approach to questions of morality, power, and status (see Boddington, 2023) and which see 'applied ethics' as that which conforms with ideals already pre-set and preformulated. This is particularly relevant for our selected technologies: when the cause *and the* response to the problem is technological - e.g. in climate engineering (CE); where the global tentacles of communication technologies are integrated into daily lives before ethical regulations are developed to curb their harms, e.g. Digital Extended Reality (dXR), where ethics may conflict with the the goal of big business, or the bio-medical-industrial complex driving research in Neurotechnologies (NT).

In preparing this report a number of methodological decisions were taken regarding the (i) the type of guidelines that must address the TechEthos technology families remit, (ii) could be produced by any relevant stakeholder family and (iii) had to focus on applied ethics. Reminder, the focus was on the ethical aspect of the guidelines, hence for some ethical issues already there was no further elaboration of ethical aspects if they were not needed.

Building on a body of state of the art literature in the area, the ethics team drew on existing guidelines, and then in consultation with the wider consortium, experts and under-represented groups, tried to identify those principles in the guidelines that were regarded as (i) meaningful and well developed and/or (ii) broad enough to encompass the wider family concerns. For example, Microsoft's Human-Computer Interaction guidelines provided a good basis to reflect on Digital Extended Reality, but did not sufficiently cover the wide range of issues identified in Deliverable D2.2, subsequently deliberation was required to fill in the gaps, where possible. Some guidelines were sufficiently developed and subsequently did not require further elaboration (see ABCs). At other times, the guidelines were too general to easily translate into applied actionable steps, for example the Tollgate Principles. In this case, the consultation process increased the detail of the proposals to create elaborated and definable principles that could be put to use by others.

Where there are a number of cross-cutting ethical issues, there are concerns unique to a particular technology that ethics guidelines must accommodate. For example, during the reflection on the guidelines for neurotechnologies a number of specific concerns were identified. A key issue is the limited scope of extant principles, which tend to focus primarily on the physical risks associated with the use of these technologies, while not giving sufficient attention to the potential digital risks. Additionally, principles often limit their scope of risks in further ways with health domains prioritised over potential risks in other domains, such as entertainment or defence. There is a need for a more holistic approach to the ethical considerations of these technologies that recognizes the potential for digital harms and considers the broader societal impact.

Moreover, ethical guidelines are not often measured in terms of their efficacy, therefore there is a need for more studies to examine how guidelines reshape particular technological arenas. Do they fully comply with the remit to produce 'ethics by design' technologies? Or are such guidelines unworkable in real-world situations? We are aware that by taking one set of decisions on the development and regulation of technologies, overt issues may be mitigated, but others might come to the fore.

These issues are beyond the scope of this deliverable and the TechEthos project, but urgently requires more research. Our approach to mitigate this in TechEthos was by starting the process of piloting/testing our guidelines with an organisation per each of the technology families. However, the complex set of activities required to identify and suggest improvements to the existing guidelines did not allow sufficient time for the final part of the task requiring the testing of the operational guidelines. We concluded that this part of the task requires time outside the deliverable timeline. However, we are in the process of contacting organisations in order to gain feedback on the usefulness of the proposals to improve the operational guidelines. To date, we have contacted partners within the TechEthos consortium to identify possible or potential collaborating organisations; we have considered partners from cluster projects such as Hybrida[4]; we also followed up on organisations identified during the digital ethnographies; and considered leads from personal contacts[5]. We will disseminate the feedback through the final report by the end of the project.

As of now, a novel methodology was developed that utilised and built on existing knowledge, categorising each principle and using it as the basis for wider consultation. The TechEthos methodology is transferable and can be used for refining guidelines for other kinds of emerging technologies. The proposed improvements to the guidelines present in this document are an outcome of that process. Moreover, while guidelines have a particular remit, and struggle to be useful in other contexts, or be too general for translating a principle into a specific action, we compensate for this by elaborating each principle with sufficient information. We tried to strengthen the transferability of the principles to other emerging technologies.

There were also grey areas worthy of note. Ethical guidelines are supposed to give clear guidance on the benefits and risks of a particular pathway, but this is not always the case. Ethical values can shift across time and context. We noted this in the area of neurotechnologies whereby there was no clear division between clinical and cosmetic enhancements. What might be today's cosmetic

[4] See https://hybrida-project.eu/

[5] See for example https://makesunsets.com/

enhancements may become a typical procedure, even clinically necessary. The medical industries have established sufficient gate-keeping to ensure that ethics of the person is valued at all times, but this may differ according to cultural or business context. For example, in the United States certain experimental procedures are more acceptable, while in China, experimental processes (including animal experimentation) that are banned in Europe, are carried out to develop novel neurotechnologies.

Moreover, experimental practices developed by business may be devoid of the responsibilities for long-term care. There is a need to ensure long-term viability of the technology, a) with conscious decision-making taking into account patients' circumstances, and b) including financial viability to avoid businesses being solely responsible for making sensitive decisions about patients (i.e. by starting treatment and then interrupting it based on a business case not on individual patient's need (for example, a specific case is detailed in TechEthos Deliverable D3.5, p. 18). The process described here comes into conflict with the business project of maximising profit and developing innovative products, always dependent on the former if produced in the private sector.

A general feature in all the guidelines expressed a commitment to equality. Through the process of proposing improvements to the guidelines we underlined how equality is an economic, political, social and ethical issue central to the development of new and emerging technologies and can be compromised by new technologies when their tools reinforce biases, and/or only become accessible to the wealthiest populations where they are available. Cultural differences may also shape how technologies are incorporated into daily practices, and may contribute to new kinds of hierarchies. Moreover, the three technologies produce vast amounts of data not recorded in human societies. This produces new modes of exploitation, value, but also comes with risks to do with privacy and security. Take for instance the growth of large language models (LLM) such as ChatGPT which work by scraping information on the web. There is little concern for copyrighted data. Moreover, the majority of the data that is scraped is a) in the English language and b) comes from Western countries (esp. North America), and is biassed towards this model of life and geared towards the middle classes.

Moving away from compliance to ethics by design will require co-creation with developers during crucial stages of the technologies' development and potentially deployment. The suggested improvements to the existing operational guidelines constitute the roadmap towards responsible innovation when developing new and emerging technologies. The uniqueness of the proposed improved guidelines is encapsulated in the diversity and inclusion of a wide range of voices including developers, policy-makers, academics and users from under-represented groups. Agility, flexibility and dialogue form the basis for how these guidelines should be used by different groups, to ensure their participation in incorporating the values into their working and living practices.

While there is no universal ethical guidance across the three TechEthos technology families and beyond, we have synthesised a set of key recommendations that can be used for proposed improvements to guidelines:

- **Bespoke governance/institutional infrastructures** - relevant administrative bodies to ensure the guidelines are properly applied, training and support in how to interpret and use the guidelines
- **Diverse stakeholder participation** - enable engagement with broadest range of stakeholders, including co-creation, co-decision making
- **Impact** - testing the efficacy of the outcomes, from use of the guidelines, with real-world examples

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

- **Inter-sector skills and knowledge exchange** - institutionalise cooperation between technology providers and policy makers
- **Responsibility to the future** - responsible forecasting, ethical defensibility, sustainability
- **Social and communicative awareness** - enable the developers and technologists to be socially aware, for example in terms of making language more accessible and gaining feedback

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

5 References

Adomaitis, L., Grinbaum, A., & Lenzi, D. (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. TechEthos Project Deliverable.* www.techethos.eu

AGU Climate Intervention Engagement: Leading the Development of an Ethical Framework. (2022). AGU. https://www.agu.org/-/media/Files/Learn-About-AGU/AGU-Climate-Intervention-Ethical-Framework.pdf

Biermann, F., Oomen, J., Gupta, A., Ali, S. H., Conca, K., Hajer, M. A., Kashwan, P., Kotzé, L. J., Leach, M., Messner, D., Okereke, C., Persson, Å., Potočnik, J., Schlosberg, D., Scobie, M., & VanDeveer, S. D. (2022). Solar geoengineering: The case for an international non-use agreement. *WIREs Climate Change*, *13*(3), e754. https://doi.org/10.1002/wcc.754

Birckhead, B., Khalil, C., Liu, X., Conovitz, S., Rizzo, A., Danovitch, I., Bullock, K., & Spiegel, B. (2019). Recommendations for Methodology of Virtual Reality Clinical Trials in Health Care by an International Working Group: Iterative Study. *JMIR Mental Health*, *6*(1), e11973. https://doi.org/10.2196/11973

Boddington, P. (2023). *AI Ethics: A Textbook*. Springer Nature Singapore. https://doi.org/10.1007/978-981-19-9382-4

Bodle, R., & Oberthür, S. (2014). *Options and Proposals for the International Governance of Geoengineering*. Federal Environment Agency (Germany).

Cannizzaro, S., Brooks, L., Richardson, K., Umbrello, S., Bernstein, M., & Adomaitis, L. (2021). *TechEthos D2.1 Methodology for ethical analysis, scan results of existing ethical codes and guidelines*. www.techethos.eu

Cox, E. M., Pidgeon, N., Spence, E., & Thomas, G. (2018). Blurred Lines: The Ethics and Policy of Greenhouse Gas Removal at Scale. *Frontiers in Environmental Science, 6*. https://www.frontiersin.org/articles/10.3389/fenvs.2018.00038

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

European Commission. Directorate General for Communications Networks, Content and Technology. (2020). *The Assessment List for Trustworthy Artificial Intelligence (ALTAI) for self assessment*. Publications Office. https://data.europa.eu/doi/10.2759/002360

Garden, H., & Winickoff, D. (2018). *Issues in neurotechnology governance*.

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), 143–174. https://doi.org/10.1080/21550085.2018.1509472

Greely, H. T., Grady, C., Ramos, K. M., Chiong, W., Eberwine, J., Farahany, N. A., Johnson, L. S. M., Hyman, B. T., Hyman, S. E., & Rommelfanger, K. S. (2018). Neuroethics guiding principles for the NIH BRAIN initiative. *The Journal of Neuroscience*, *38*(50), 10586.

Hermansson, H., & Hansson, S. O. (2007). A three-party model tool for ethical risk analysis. *Risk Management*, 9(3), 129–144.

Heyward, C. (2013). Situating and Abandoning Geoengineering: A Typology of Five Responses to Dangerous Climate Change. *PS: Political Science & Politics*, *46*(1), Article 1. https://doi.org/10.1017/S1049096512001436

Heyward, C., Rayner, S., & Savulescu, J. (2017). Early Geoengineering Governance: The Oxford Principles. In D. M. Kaplan (Ed.), *Philosophy, Technology, and the Environment* (p. 0). The MIT Press. https://doi.org/10.7551/mitpress/9780262035668.003.0007

Honegger, M., Baatz, C., Eberenz, S., Holland-Cunz, A., Michaelowa, A., Pokorny, B., Poralla, M., & Winkler, M. (2022). The ABC of Governance Principles for Carbon Dioxide Removal Policy. *Frontiers in Climate*, *4*. https://www.frontiersin.org/articles/10.3389/fclim.2022.884163

Honegger, M., Burns, W., & Morrow, D. R. (2021). Is carbon dioxide removal 'mitigation of climate change'? *Review of European, Comparative & International Environmental Law*, *30*(3), Article 3. https://doi.org/10.1111/reel.12401

Hubert, A.-M. (2021). A Code of Conduct for Responsible Geoengineering Research. *Global Policy*, *12*(S1), 82–96. https://doi.org/10.1111/1758-5899.12845

Hubert, A.-M., & Reichwein, D. (2015). *An Exploration of a Code of Conduct for Responsible Scientific*

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

Research involving Geoengineering. 96.

IPCC 2022. (2022). *Climate Change 2022: Impacts, Adaptation and Vulnerability. Contribution of Working Group II to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change* (p. 3056). Cambridge University Press. 10.1017/9781009325844

ISO/TC 265—Carbon dioxide capture, transportation, and geological storage. (2023, March 7). ISO. https://www.iso.org/committee/648607.html

Lenferna, G. A., Russotto, R. D., Tan, A., Gardiner, S. M., & Ackerman, T. P. (2017). Relevant climate response tests for stratospheric aerosol injection: A combined ethical and scientific analysis. *Earth's Future*, *5*(6), 577–591.

Loomis, R., Cooley, S. R., Collins, J. R., Engler, S., & Suatoni, L. (2022). A Code of Conduct Is Imperative for Ocean Carbon Dioxide Removal Research. *Frontiers in Marine Science*, *9*. https://www.frontiersin.org/articles/10.3389/fmars.2022.872800

Lynch, Z. (2004). Neurotechnology and society (2010–2060). *Annals of the New York Academy of Sciences*, *1013*(1), 229–233.

Microsoft. (2017). *Microsoft AI principles. Microsoft*. https://www.microsoft.com/en-us/ai/our-approach-to-ai

Mills, C. W. (2005). 'Ideal Theory' as Ideology. *Hypatia*, *20*(3), Article 3. https://doi.org/10.1353/hyp.2005.0107

Morrow, D. R. (2018). Putting the Tollgate Principles into Practice. *Ethics, Policy & Environment*, *21*(2), 175–177.

Morrow, D. R., Kopp, R. E., & Oppenheimer, M. (2013). Political legitimacy in decisions about experiments in solar radiation management. In W. C. G. Burns & A. Strauss (Eds.), *Climate Change Geoengineering: Philosophical Perspectives, Legal Issues, and Governance Frameworks*. Cambridge University Press.

Nericcio, L. (2018). Examining the Implications of the Tollgate Principles for the Governance of Geoengineering. *Ethics, Policy & Environment*, *21*(2), 184–186.

OECD. (2019). *Recommendation of the Council on Responsible Innovation in Neurotechnology*. OECD.

Rayner, S., Heyward, C., Kruger, T., Pidgeon, N., Redgwell, C., & Savulescu, J. (2013). The Oxford

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

Principles. *Climatic Change*, *121*(3), Article 3. https://doi.org/10.1007/s10584-012-0675-2

Reynolds, J. L. (2019). Solar geoengineering to reduce climate change: A review of governance proposals. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences*, *475*(2229), Article 2229. https://doi.org/10.1098/rspa.2019.0255

Rothenberger, L., Fabian, B., & Arunov, E. (2019). RELEVANCE OF ETHICAL GUIDELINES FOR ARTIFICIAL INTELLIGENCE – A SURVEY AND EVALUATION. *Research-in-Progress Papers*. https://aisel.aisnet.org/ecis2019_rip/26

Shepherd, J. G. (2009). *Geoengineering the climate: Science, governance and uncertainty*. Royal Society. https://doi.org/10/29)

Spiegel, J. S. (2018). The Ethics of Virtual Reality Technology: Social Hazards and Public Policy Recommendations. *Science and Engineering Ethics*, *24*(5), 1537–1550. https://doi.org/10.1007/s11948-017-9979-y

Umbrello, S., Bernstein, M. J., Vermaas, P. E., Resseguier, A., Gonzalez, G., Porcari, A., Grinbaum, A., & Adomaitis, L. (2023). From speculation to reality: Enhancing anticipatory ethics for emerging technologies (ATE) in practice. *Technology in Society*, *74*, 102325. https://doi.org/10.1016/j.techsoc.2023.102325

Vaidyam, A. N., Wisniewski, H., Halamka, J. D., Kashavan, M. S., & Torous, J. B. (2019). Chatbots and Conversational Agents in Mental Health: A Review of the Psychiatric Landscape. *The Canadian Journal of Psychiatry*, *64*(7), 456–464. https://doi.org/10.1177/0706743719828977

von Schomberg, R. (2012). The Precautionary Principle: Its Use Within Hard and Soft Law. *European Journal of Risk Regulation*, *3*(2), 147–156.

Wolff, J. (2020). Fighting risk with risk: solar radiation management, regulatory drift, and minimal justice. *Critical Review of International Social and Political Philosophy* 23(5), 564-583.

Wolff, J. (1996). Integration, justice, and exclusion. In U. Bernitz & P. Hallstrom (Eds.), Principlesof justice and the European Union (pp. 15–26). Stockholm: Juristforlaget.

Zhou, L., Gao, J., Li, D., & Shum, H.-Y. (2019). The Design and Implementation of XiaoIce, an Empathetic Social Chatbot. *arXiv:1812.08989 [Cs]*. http://arxiv.org/abs/1812.08989

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

TECHETHOS

FUTURE ○ TECHNOLOGY ○ ETHICS

Coordinated by

Partners

Linked Third Parties

www.techethos.eu info@techethos.eu

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