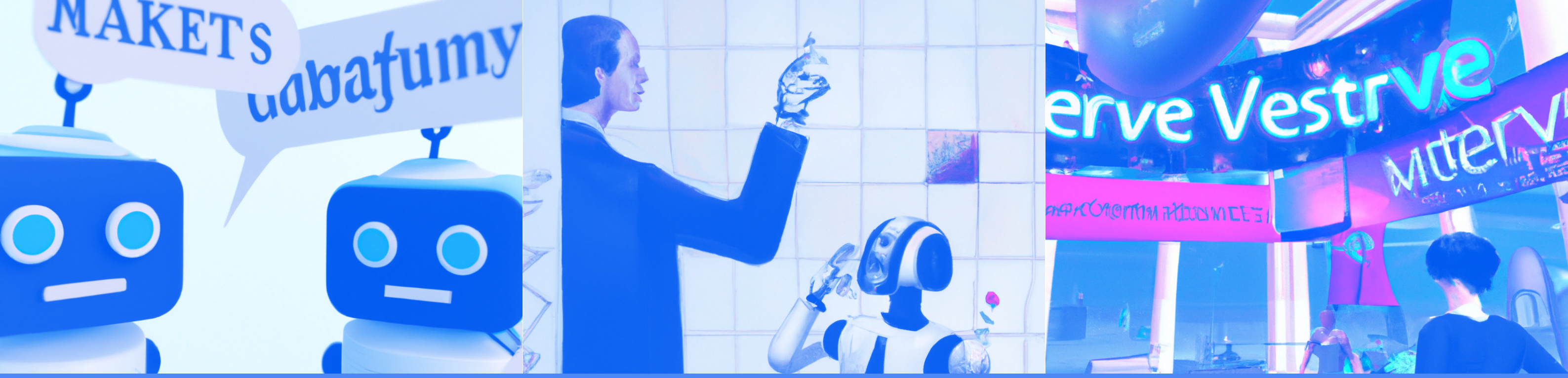


Policy Brief

XR and General Purpose AI: from values and principles to norms and standards

TECHETHOS

FUTURE ○ TECHNOLOGY ○ ETHICS

Highlights

Highlights

The TechEthos project focused its ethical analysis on eXtended Reality and Natural Language Processing (NLP) within the larger context General Purpose Artificial Intelligence (AI). AI has already been broadly implemented for a variety of purposes, applications and services, from simple data collection and analysis to sophisticated, human-like operations. The types of use open completely different scenarios in terms of ethical risks. We focus here on two specific aspects:

- An AI system can be outfitted with language capabilities and an avatar representation, both of which raise a problem of indistinguishability between human likeness/language and machine simulation thereof;
- Personal data and biometric data collected via XR devices is used for training next-generation general-purpose AI, such as emotional AI systems or chatbots that efficiently nudge people toward desired behaviour.

Who is this for?

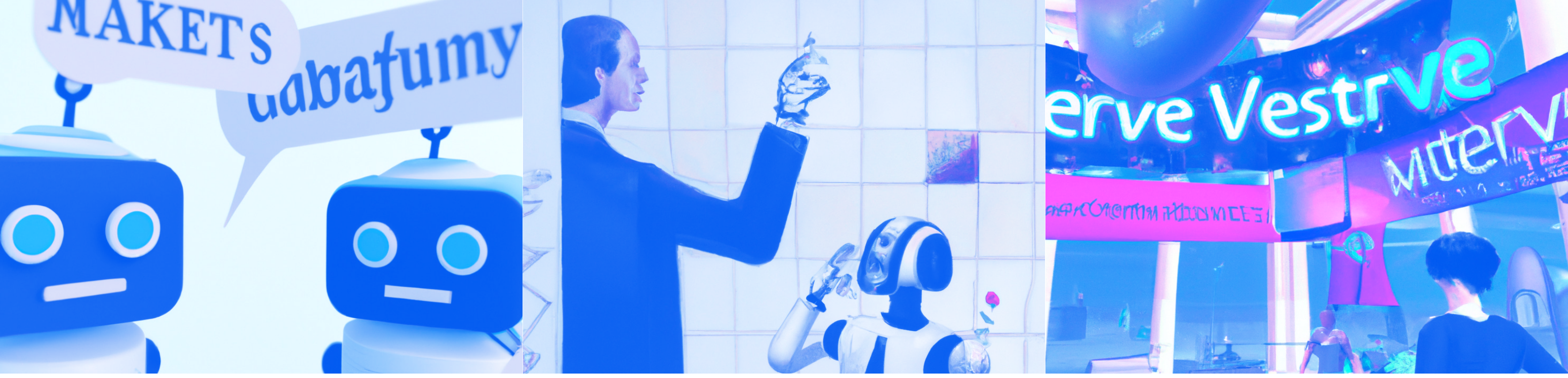
This policy brief seeks to inform those involved in the governance and development of XR technologies and general purpose AI, and is primarily aimed at **EU policymakers** and **tech developers.**

Background

Values and high-level principles are not enough for AI regulation

Ethical issues of AI systems are usually formulated through the lens of values and principles. However, European policy makers should go beyond merely listing such values and principles, because manufacturers may not immediately understand how to implement them in the design of AI systems. For the proposed EU regulation to be effective, we offer an operationalization of the values and principles in the form of suggested norms and standards. Here, we list new and emerging issues to supplement, enhance and update the Assessment List for Trustworthy Artificial Intelligence (ALTAI) developed by the High-Level Expert Group on AI.

TechEthos receives funding from the EU H2020 research and innovation programme under Grant Agreement No 101006249. This output reflects the views of the authors. The Research Executive Agency and the European Commission are not responsible for any use which might be made of the information contained herein.



Key Messages

Promote transparency in XR

Transparency in XR refers to the awareness of the human user as to the nature of entities or objects they encounter or interact with. By nature, these entities can be digital, material, or mixed. Merely knowing the nature of an object is insufficient for preventing effects on the user. Even a well-informed user spontaneously projects knowledge, emotions, intentions, or cognitive states on the AI system.

- There is the need to have clear information in plain language about the nature of the environment and of the entities or objects that the user interacts with;
- This information needs to be presented at key moments and intervals during the user's interaction in the virtual environment, such as the beginning of a conversation with a chatbot;
- A European norm should specify a standard protocol to determine the user's subjective understanding of this information;
- The manufacturer should present immediate log-out options to the user who wishes to leave the virtual environment.

Address risks of harmful manipulation

Manipulation refers to the ability of AI systems to manipulate users in order to achieve a hidden goal, both in virtual and material environments. Unsupervised or self-supervised AI systems can demonstrably develop manipulative techniques (for example, lying or emotional nudging) without explicit intent of the manufacturer.

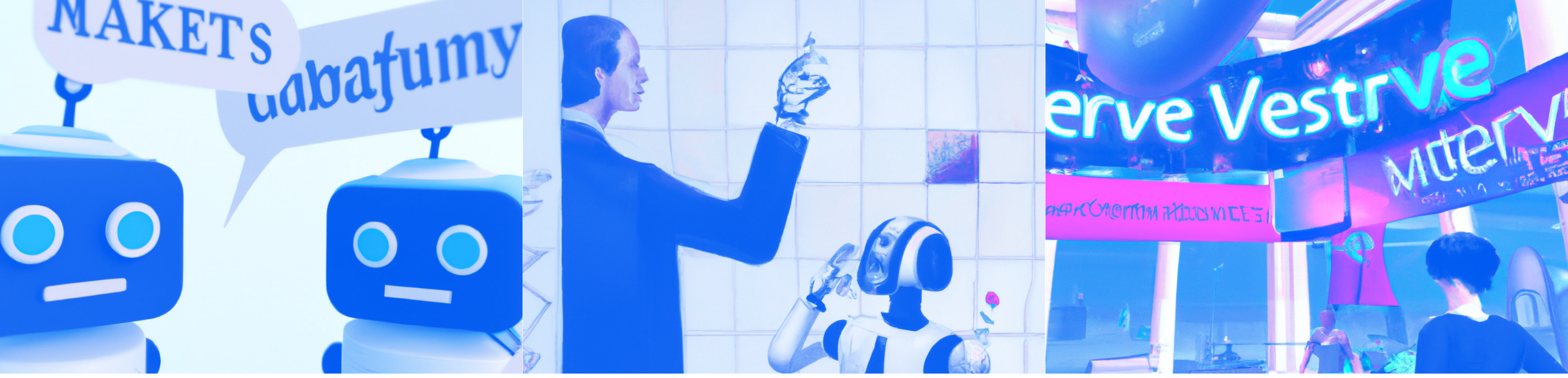
- AI systems performing self-supervised learning or reinforcement learning solely based on awards for achieving predefined goals can lead to undesired consequences. Such AI systems require special provisions to prohibit deception and to prevent "the ends justify the means" strategies;
- Nudging or manipulation to the sole benefit of the manufacturer or the operator should be prohibited, while nudging to the benefit of the user should be evaluated on a case-by-case basis depending on context;
- In some adversarial scenarios, deception becomes a goal (for example, spreading misinformation for political gain), showing the need to rigorously enforce human-machine distinction;
- Machine-generated language should be watermarked in order to maintain the human-machine distinction on sufficiently large textual outputs;
- Watermarks should be present in all outputs produced by Generative AI, including text, images, audio, and video. Watermarks should be easily verifiable by human users.

Protect the dignity of users

Dignity refers to the due respect in a virtual environment with regard to digital representations of real humans, especially of deceased individuals or well-known figures. This problem is further emphasized by Generative AI being able to create new original content (for example, non-plagiarized language) for an avatar pretending to be a real person.

- Avatars pretending to be a real individual, combined with generative AI that produces highly similar but original outputs, may constitute an infringement of human

TechEthos receives funding from the EU H2020 research and innovation programme under Grant Agreement No 101006249. This output reflects the views of the authors. The Research Executive Agency and the European Commission are not responsible for any use which might be made of the information contained herein.



dignity, unless they are covered by informed consent of the impersonated subject. Since the outputs of AI systems cannot be predicted with certainty, dignity cannot be absolutely protected but should be checked via a set of controls and benchmarks;

- To ensure respect of their dignity, human subjects must have a say in what will happen with their personal data posthumously. Currently, this topic is insufficiently covered by the General Data Protection Regulation (GDPR).

Clarify who is responsible

The most important concern with regard to **responsibility** refers to the **identification of agents behind avatars** in a shared virtual environment.

- Virtual actions in the metaverse can lead to psychological and material effects on human users. Therefore, even a virtual action implies an ethical responsibility;
- Without avatar identification, this ethical responsibility remains a virtuality. Only virtual types of retribution (loss of digital goods or status, digital prison, banning, reduced access, etc.) can be envisaged;
- With avatar identification, real-world responsibility will apply to the actions of human-driven avatars. This includes liability of human agents for damage in a virtual environment;
- Agents who bear responsibility for virtual actions include the developer, the "trainer" (overseeing the selection of training data), the manufacturer, and the user. In each case, the sharing of responsibility should be determined depending on context.

Determine appropriate levels of autonomy

Autonomy in Generative AI refers to the projection of moral and cognitive traits from the user onto the interlocutor, especially when the latter is a machine.

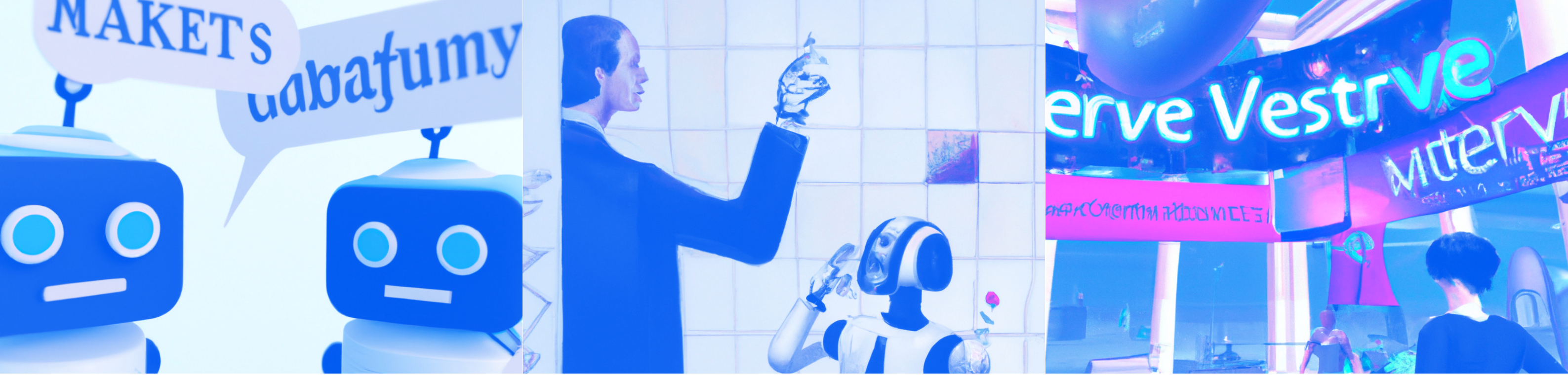
- Projection of moral traits on text-generating AI systems should be artificially limited because such systems (for example, chatbots) do not bear responsibility for their outputs;
- The names of chatbots, especially endowed with an avatar, should not be freely chosen by the users in order to avoid reinforcing the projection of subjecthood on chatbots and endowing users with excessive power;
- In controlled environments, for example in education, psychiatry, childcare or geriatric care, strong personalization can be allowed if the functionality of a chatbot relies on projecting trust onto the machine.

Ensure equitable labour conditions

In many respects, **virtual labour** is equivalent to material labour and needs to be compensated fairly. Hence, equitable labour conditions in XR need to be ensured.

- If artefacts in a virtual environment are bought or sold, there needs to be a transparent mechanism to split profits and to compensate the workers;
- Surveillance capabilities in virtual work environments should be limited and regulated in light of privacy and autonomy concerns.

TechEthos receives funding from the EU H2020 research and innovation programme under Grant Agreement No 101006249. This output reflects the views of the authors. The Research Executive Agency and the European Commission are not responsible for any use which might be made of the information contained herein.



Uphold decency in generative AI

Decency in Generative AI refers to the possibility of an offensive or harmful interaction between the user and the AI system.

- Harms, including types of toxic language, should be labelled at the training stage and processed accordingly during machine learning. Filters for potentially harmful outputs should be put in place;
- Manufacturers should define and implement a policy specifying how the AI system will respond to toxic inputs from the user;
- Manufacturers should design and implement mitigation techniques against unfair bias, particularly on gender, sensitive and protected data, as well as mitigation techniques against cultural stereotyping.

Evaluate the environmental impacts of XR and generative AI

General Purpose AI system might be highly demanding in terms of energy consumption, raising concerns regarding their **environmental footprint**. Environmental issues raised by General Purpose AI systems are caused by the resources used for training and by the amount of computation required to execute each prompt. Currently, manufacturers are scaling up their computational resources to address the demand, however the volume is set to increase rapidly if and when prompts begin to be produced by other AI systems.

- Manufacturers should filter inputs to allow only human-generated prompts;
- The infrastructure for General Purpose AI should prioritise edge computing outsourcing the computational load to end-user devices.

Address privacy and security concerns

Privacy and security issues refer to the trade-off between the right to privacy and the right to physical safety and security. A dedicated TechEthos policy note provides a detailed legal analysis of these concerns.

Further reading

- 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. zenodo.7619852
- Glaese A., et al. (2022) Improving alignment of dialogue agents via targeted human judgements. arXiv:2209.14375
- Hacker, P., Engel, A., Mauer, M., (2023). Regulating ChatGPT and other Large Generative AI Models. arXiv.2302.02337
- Kirchenbauer J., et al. (2023). A Watermark for Large Language Models. arXiv:2301.10226
- Weidinger L., et al. (2021) Ethical and social risks of harm from Language Models. arXiv:2112.04359

Authors

Policy brief prepared by Laurynas Adomaitis and Alexei Grinbaum, CEA (Paris, France).

Illustrations used in this brief were generated by the software DALL E using prompts related to chatbots

Keep in touch

@ www.techethos.eu
info@techethos.eu
@TechEthosEU
TechEthosEU

TechEthos receives funding from the EU H2020 research and innovation programme under Grant Agreement No 101006249. This output reflects the views of the authors. The Research Executive Agency and the European Commission are not responsible for any use which might be made of the information contained herein.