These projects have received funding from The European Union's research and innovation programme under GA No. 951911 (AI4Media), 951847 (ELISE), 101070617 (ELSA), 101070596 (euROBIN), 761758 (HumaneAI Net), 952215 (TAILOR) and 952070 (VISION).

# Report on the key findings from the Theme Development Workshop "Trusted AI: The Future of Creating Ethical & Responsible AI Systems"

– September 2023 –

## Executive Summary

The 2nd cross-cutting Theme Development Workshop (TDW) on "Trusted AI: The Future of Creating Ethical and Responsible AI Systems"[1], jointly organised by AI4Media, ELISE, ELSA, euRobin, HumanE-AI-Net, CLAIRE, TAILOR and VISION[2], took place on 13 September 2023, with the aim of developing and identifying the most promising and emerging themes related to the overarching concept of Trustworthy AI. In this one-day workshop, experts from academia, industry, and policy jointly developed initial impetus for the European Artificial Intelligence (AI) Research and Innovation Roadmap. Stimulated by the introductory talks and presentations by selected experts, participants actively discussed a variety of topics in the breakout sessions and shared their key findings in the subsequent plenary presentations. In addition, some initial ideas for follow-up activities and further collaborations were identified.

This report provides a summary of the outcomes of the Theme Development Workshop, "Trusted AI: The Future of Creating Ethical and Responsible AI Systems". In order to make the results available to a wider audience and in particular to the European AI community, this report is being made public via the organiser's web pages.

[1] Further information via the VISION website: https://www.vision4ai.eu/tdw-trusted-ai/ (as of 12 September 2023)

[2] In alphabetical order.

These projects have received funding from The European Union's research and innovation programme under GA No. 951911 (AI4Media), 951847 (ELISE), 101070617 (ELSA), 101070596 (euROBIN), 761758 (HumaneAI Net), 952215 (TAILOR) and 952070 (VISION).

Authors of this report:
(in alphabetical order)


- Janina Hoppstädter, German Research Centre for Artificial Intelligence (DFKI) & CLAIRE
- Kyra Kiefer, German Research Centre for Artificial Intelligence (DFKI) & CLAIRE
- Lidia Dutkiewicz, KU Leuven Centre for IT & IP Law – imec
- Naoise Holohan, IBM Research Europe - Ireland
- Noémie Krack, KU Leuven Centre for IT & IP Law – imec
- Maura Pintor, University of Cagliari
- Vasileios Mezaris, Information Technologies Institute (ITI), Centre for Research and Technology Hellas (CERTH)
- Wico Mulder, TNO
- Cees Snoek, University of Amsterdam
- Plamen Angelov, Lancaster University, UK
- Schutera, Mark, ZF Friedrichshafen AG
- Marlies Thönnissen, German Research Centre for Artificial Intelligence (DFKI) & CLAIRE
- André Meyer-Vitali, German Research Centre for Artificial Intelligence (DFKI)
- Giulio Mecacci, Donders Institute for Brain, Cognition and Behaviour - The Netherlands

These projects have received funding from The European Union's research and innovation programme under GA No. 951911 (AI4Media), 951847 (ELISE), 101070617 (ELSA), 101070596 (euROBIN), 761758 (HumaneAI Net), 952215 (TAILOR) and 952070 (VISION).

# Table of Contents

These projects have received funding from The European Union's research and innovation programme under GA No. 951911 (AI4Media), 951847 (ELISE), 101070617 (ELSA), 101070596 (euROBIN), 761758 (HumaneAI Net), 952215 (TAILOR) and 952070 (VISION).

# Introduction

In September 2020, four new AI networks were established by the European Commission via the call "Towards a vibrant European network of AI excellence centres" (ICT-48-2020). The aim of these networks is to foster the collaboration between the best research teams in Europe, and to address the major scientific and technological challenges in the field of AI. These four networks are coordinated and supported by the VISION project to foster activities that reach critical mass and enable the creation of a world-class AI ecosystem in Europe.

One of these activities is the so-called Theme Development Workshops (TDWs), an innovative format that brings together key stakeholders from industry, academia, and policy to jointly identify the most important AI research topics and challenges in a given field or for a given industry sector. In December 2020, it was agreed between the respective coordinators and leadership teams of TAILOR, VISION, HumanE-AI-Net, and CLAIRE to plan and conduct a series of joint (co-organized) Theme Development Workshops starting in 2021. Five workshops in the joint series focused on the Public Sector, Mobility, Healthcare, Manufacturing, and Energy sectors. The results from these workshops highlighted issues not only relevant to their respective industry sectors, but also beyond. These cross-cutting issues are addressed in the so-called cross-cutting TDWs. To this end, the first cross-cutting TDW on "AI: Mitigating Bias and Disinformation" was held on May 18. The workshop on Trusted AI is now the second cross-cutting workshop and was organised by the six Networks of Excellence for AI (AI4Media, ELISE, ELSA, euRobin, HumanE-AI-Net, TAILOR) together with CLAIRE under the lead of the CSA VISION in the context of WP4 Joint Forces of Academia and Industry. This report is the result of the seventh joint TDW organised and conducted as part of this workshop series. The main purpose of this joint cross-cutting TDW was to discuss upcoming overarching topics in AI, generate input for a European AI Trend Radar and prepare the ground for follow-up activities and collaboration.

*Trustworthy AI in Europe is dedicated to the development and adoption of AI systems that are reliable, responsible, and transparent in Europe. We recognize that trust is essential for fostering long-term sustainable innovation, economic growth, and societal well-being in the digital era. Our mission is to ensure that AI technologies deployed in Europe adhere to ethical principles, respect fundamental rights, and address the unique challenges and values of the European context. By promoting transparency, fairness, accountability, and the protection of personal data, we aim to build trust among individuals, businesses, and public institutions. Through collaboration with stakeholders, policymakers, and researchers, Trustworthy AI Europe seeks to shape robust regulatory frameworks, establish standards, and foster a culture of transparency and trust in AI. Our mission is to leverage AI's potential to benefit European society, contribute to sustainable development, and empower citizens, while safeguarding their privacy, security, and fundamental rights. Together, we strive to create a European AI landscape that upholds trust, ethical values, and the public interest.*

These projects have received funding from The European Union's research and innovation programme under GA No. 951911 (AI4Media), 951847 (ELISE), 101070617 (ELSA), 101070596 (euROBIN), 761758 (HumaneAI Net), 952215 (TAILOR) and 952070 (VISION).

# Keynotes and introductory presentations

The TDW was opened by co-chairs Philipp Slusallek, German Research Centre for Artificial Intelligence (DFKI), Filareti Tsalakanidou, CERTH-ITI, and Lorraine Wolter (CISPA) on behalf of the Organizing Committee (OC), which includes other representatives from Aalto University, CEA, CISPA, DFKI, CERTH-ITI, and Umeå University. The co-chairs presented the TDW objectives, agenda and program, and introduced the invited keynote speakers to the participants.

The inspiring keynote speeches were delivered by high-level experts from academia, public sector and industry. These introductory presentations highlighted different perspectives of the importance of trustworthiness in AI and served as a basis for the discussions on the need for trustworthiness in Artificial Intelligence for the future. The keynotes also encouraged the in-depth discussions in the following breakout sessions.

## Introductory presentations

### *Principles of Trusted AI*

**About the keynote speaker**
Dr. André Meyer-Vitali is a computer scientist who got his Ph.D. in software engineering, ubiquitous computing and distributed AI from the University of Zürich. He worked on many applied research projects on Ambient Intelligence and multi-agent systems at Philips Research and TNO (The Netherlands) and contributed to AgentLink. He also worked at the European Patent Office. Currently, he is a senior researcher at DFKI (Germany) focused on engineering and promoting Trusted AI and is active in the AI networks TAILOR and CLAIRE. His research interests include Software und Knowledge Engineering, Design Patterns, Neuro-Symbolic AI, Causality, and Agent-based Social Simulation (ABSS) with the aim to create Trust by Design.

André Meyer-Vitali delved into the fundamental aspects of Trusted AI, emphasising the need for trust not just in individuals' daily interactions but also in the tools and technologies they use, especially AI systems. This trust is particularly crucial in critical applications and infrastructures where the implications of malfunction or misuse can be severe.

He discussed the motivation behind Trusted AI, emphasising the alignment with industrial standards and regulations. By complying with these standards, AI systems can be deemed safe, reliable, and secure. The European AI Act was highlighted as a pivotal document, employing a risk-based approach that categorises AI applications into different risk levels. High-risk applications, which pose significant societal and individual risks, are the focal point for building trust and ensuring safety.

To create Trusted AI, four key aspects were outlined that form the foundation of their approach:

These projects have received funding from The European Union's research and innovation programme under GA No. 951911 (AI4Media), 951847 (ELISE), 101070617 (ELSA), 101070596 (euROBIN), 761758 (HumaneAI Net), 952215 (TAILOR) and 952070 (VISION).

1. **Models and Explanations:** Using combinations of explicit knowledge models and methods (neuro-explicit AI) allows for reliable predictions about a system's behaviour, which enables transparent, insightful, and plausible explanations and simulations with generalised models from knowledge and training.

2. **Causality and Grounding:** Understanding cause-and-effect relationships within AI systems is essential. Beyond mere correlations, understanding the reasons behind the system's decisions is critical. Grounding AI concepts in the real world involves attaching meaning to labels and ensuring that the AI system understands the context in which it operates.

3. **Modularity and Structure:** Breaking down complex AI systems into modular components enhances understanding and control. By decomposing large systems into smaller, manageable parts, experts can comprehend the system as a whole. This approach also allows for more effective monitoring and adaptation to changing circumstances.

4. **Human Agency and Oversight:** While AI systems can perform complex tasks, human oversight remains essential. Ensuring that humans understand the system's functioning and can intervene when necessary is crucial. This involves mutual awareness, where AI systems understand human intentions and vice versa, fostering a collaborative and accountable relationship.

Additionally, André Meyer-Vitali introduced the concept of the Centre for European Research in Trusted AI (CERTAIN[3]). This initiative aims to bring together researchers, experts, and stakeholders to advance the understanding and implementation of Trusted AI. By fostering collaboration on both local and European scales, the centre intends to accelerate research and development in these critical areas.

In summary, the speaker's detailed discussion outlined the multifaceted nature of Trusted AI, emphasising the need for a comprehensive approach that encompasses functionality, causality, modularity, and human oversight. The introduction of the Centre for European Research in Trusted AI further underscores the commitment to advancing these principles through collaborative research and knowledge sharing.

[3] https://www.certain-trust.eu/

These projects have received funding from The European Union's research and innovation programme under GA No. 951911 (AI4Media), 951847 (ELISE), 101070617 (ELSA), 101070596 (euROBIN), 761758 (HumaneAI Net), 952215 (TAILOR) and 952070 (VISION).

## *Role of the EU and orientation of EU policy making in relation to trustworthy, responsible and ethical AI*

**About the keynote speaker**

Antoine-Alexandre has a multidisciplinary background in political science, applied economics and international relations. For the past two and a half years, he has been working for the Commission's unit in charge of AI policy development and coordination. More specifically, he is closely following the interinstitutional negotiations on the AI Act proposal and is responsible for the AI standardisation strategy that will support the implementation of the future legislation.

The keynote presentation offered a comprehensive insight into the European Commission's initiatives in the field of Artificial Intelligence (AI) and provided an overview of the forthcoming AI Act. It also delved into the Commission's strategies to ensure the effective implementation of AI legislation.

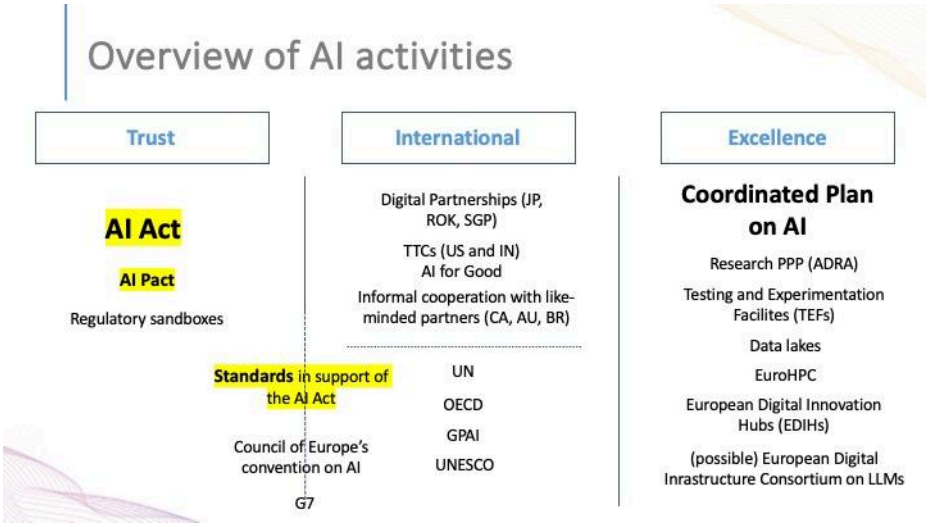


*Figure 1: Overview of AI activities*
*(taken from presentation from Antoine-Alexandre André, European Commission)*

**Current Commission Activities on AI:** The European Commission's approach to AI was built upon two foundational pillars: the creation of an ecosystem of trust and an ecosystem of excellence. These two pillars served as the main building blocks of their AI strategy (see figure 1):

- **Ecosystem of Trust (Left):** This aspect focused on legislation designed to establish rules for safeguarding the safety, health, and fundamental rights of European citizens. It aimed to provide a legal framework that ensured responsible and ethical AI development and deployment.
- **Ecosystem of Excellence (Right):** This part of the strategy was based on a coordinated plan for AI adoption. It aimed to incentivize and accelerate the adoption of AI solutions

These projects have received funding from The European Union's research and innovation programme under GA No. 951911 (AI4Media), 951847 (ELISE), 101070617 (ELSA), 101070596 (euROBIN), 761758 (HumaneAI Net), 952215 (TAILOR) and 952070 (VISION).

while aligning efforts at the European, national, and regional levels. The goal here was to foster excellence in AI innovation and application.

**Overview of Commission's International AI Activities.** While the presentation touched on various international AI-related activities, it primarily focused on the creation of an ecosystem of trust, aligning with the theme of the workshop on Trusted AI.

**Ecosystem of Trust.** The presentation delved into the Commission's proposed legislation, known as the AI Act. This legislation played a pivotal role in establishing trust in AI technologies. The presentation explained the fundamental concepts behind the AI Act and emphasised the crucial role of standardisation in implementing the Act's requirements. The central aim here was to promote the adoption of AI systems within the EU that were not only technologically advanced but also trustworthy and compliant with fundamental rights.

**Why Regulate AI.** The question arose: Why did the European Commission seek to regulate AI systems? The answer was rooted in the unique characteristics of AI technology. While AI applications were pervasive across various domains, they had been partially covered by different pieces of legislation at both national and European levels. The scale and impact of AI systems created significant challenges in applying existing rules effectively, potentially jeopardising the safety and fundamental rights of European citizens. Therefore, the Commission initiated efforts to develop horizontal regulation, allowing for a coherent approach to AI without stifling European industry innovation.

**Proposed Legislation.** The presentation highlighted four essential elements to understand the essence of the proposed AI Act:

1. **Horizontal Legislation:** The AI Act introduced uniform rules governing the market placement of AI systems considered as products. It covered the entire AI lifecycle, addressed risks to safety, health, and fundamental rights, and sought to create a single market for trustworthy AI. Importantly, it was designed to complement existing EU and national laws.
2. **Innovation-Friendly:** The proposed regulation was intended to be innovation-friendly. It aimed to provide legal certainty for operators while instilling trust in the AI market.
3. **Level Playing Field:** The AI Act was designed to create a level playing field for all players, regardless of their origin—EU or non-EU.
4. **Risk-Based Approach:** The Act adopted a risk-based approach, classifying AI systems into categories based on their potential risks to safety and fundamental rights. This categorization ranged from high-risk systems to those that posed minimal risks (see figure 2).

These projects have received funding from The European Union's research and innovation programme under GA No. 951911 (AI4Media), 951847 (ELISE), 101070617 (ELSA), 101070596 (euROBIN), 761758 (HumaneAI Net), 952215 (TAILOR) and 952070 (VISION).

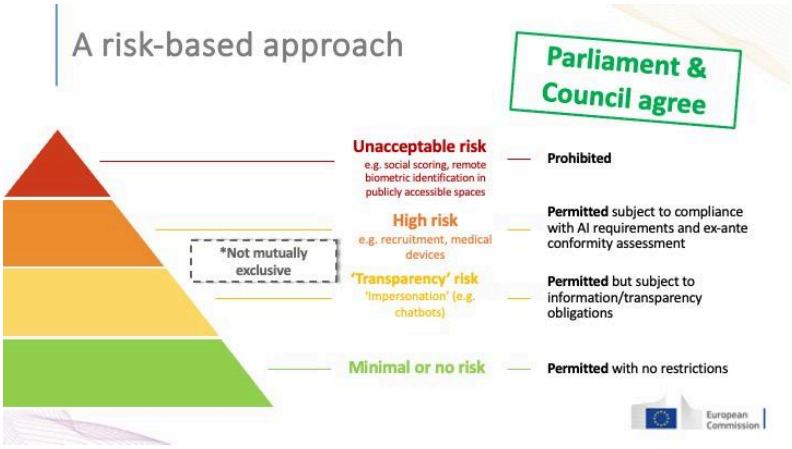


*Figure 2: A risk-based approach*
*(taken from presentation from Antoine-Alexandre André, European Commission)*

**State of Play in Negotiations.** The presentation provided an update on the progress of negotiations with co-legislators, the Council, and the European Parliament. The Council adopted its position in December 2022, largely preserving the overall architecture of the Commission's proposal. Meanwhile, the European Parliament engaged in extensive deliberations before adopting the AI Act in June 2023. Trilogue discussions began in June 2023, and there was a commitment to conclude the negotiations before the European Parliament's mandate ended in spring 2024. The key points of discussion included general-purpose AI, foundation models, high-risk use cases, and prohibitions.

**Next Steps.** Trilogues were ongoing, with a strong commitment from co-legislators to adopt the AI Act by spring 2024. The Commission hoped to reach a political agreement on the text even before the end of the year. Additionally, a transition period of two or three years was envisioned between the formal adoption of the AI Act and its full implementation. During this period, an AI pact with industry was planned to facilitate early compliance and preparation.

**The AI Pact.** Commissioner Breton introduced the AI pact to encourage industry players to voluntarily commit to implementing the AI Act's requirements ahead of the legal deadline. This pact aimed to create a framework for companies to demonstrate their commitment to AI objectives, share their preparations for compliance, and ensure the trustworthy design, development, and use of AI.

**Timeline for Industry Pledges.** The presentation included a timeline for industry pledges, emphasising that the timeline was subject to change based on ongoing discussions at the EU and international levels. The activities under the AI pact would align closely with the outcomes of the AI Act negotiations and the standardisation work intended to support the Act's implementation.

**The Role of Standards.** An essential aspect of the proposed AI Act was its alignment with the Product Safety New Legislative Framework (NLF). This approach involved setting essential requirements in the main legal act while detailing how these requirements were met through European harmonised standards. Harmonised standards, if adopted by a company, signified

These projects have received funding from The European Union's research and innovation programme under GA No. 951911 (AI4Media), 951847 (ELISE), 101070617 (ELSA), 101070596 (euROBIN), 761758 (HumaneAI Net), 952215 (TAILOR) and 952070 (VISION).

compliance with the AI Act's requirements. The European Standardization Organization played a critical role in developing these standards.

**Areas of Standardisation Work.** The presentation outlined various areas where harmonised standards would be developed to operationalize the AI Act's requirements. These areas encompassed cybersecurity, transparency, robustness, accuracy, and more. The Commission aimed to have a substantial number of harmonised standards available three to six months before the AI Act's application, highlighting the significance of these standards in practical implementation.

**Commission's Activities Related to AI Standardization.** The Commission actively engaged in monitoring and supporting standardisation activities at the EU and international levels. These activities involved mapping relevant standardisation efforts, participating in strategic standardisation activities, and providing legal, political, and operational support to the European Standardization Organization. Furthermore, there was a push to involve civil society representatives, researchers, industry experts, and stakeholders in AI standardisation activities.

**Encouragement for Future Engagement.** In conclusion, the presentation urged all stakeholders not to view the AI Act's adoption as the endpoint but rather as the beginning of a journey towards practical implementation. Active engagement in standardisation activities was encouraged to ensure the AI Act's high-level requirements translated effectively into practice. The ultimate goal was to establish an ecosystem of trust and promote the widespread adoption of trustworthy AI across Europe.
This detailed overview emphasised the European Commission's commitment to fostering responsible and trustworthy AI while ensuring the competitiveness of European industries in the global AI landscape.

To get engaged in standardisation activities to help the European Commission ensure deployment of trustworthy and responsible AI in Europe, please reach out to Antoine-Alexandre André via antoine-alexandre.andre@ec.europa.eu.

These projects have received funding from The European Union's research and innovation programme under GA No. 951911 (AI4Media), 951847 (ELISE), 101070617 (ELSA), 101070596 (euROBIN), 761758 (HumaneAI Net), 952215 (TAILOR) and 952070 (VISION).

## *Ethical AI*

**About the keynote speaker**

Meeri is the CEO and Founder of Saidot, a Finnish start-up providing technology and services to help enterprises build and deploy responsible AI. Saidot's SaaS platform for AI Governance and Transparency is used by major public and private organizations to apply systematic AI governance and to communicate transparently about their AI. Meeri was the chair of the ethics working group in Finland's national AI program that submitted its final report in March 2019. Meeri is also the Chair of IEEE's AI Impact Use Cases Initiative and an alumnus of the Berkman Klein Center for Internet & Society at Harvard University. Meeri has been a member of Snap Inc.'s Safety Advisory Board since June 2023.

Meeri Haataja, a distinguished speaker from Saidot, delivered a keynote on the expansive and crucial topic of "Ethical AI." Drawing from her practical experiences working closely with private companies and governments, she emphasised the growing challenges in ensuring the ethical alignment of AI systems as their capabilities advance. In her insightful presentation, she navigated through three key statements, driving home the urgent need to integrate ethics seamlessly into the development and deployment of AI systems.

Opening with the fundamental question, "How is AI ethics?" Haataja encouraged a nuanced understanding by defining ethical AI as systems aligned with the ethical norms and values of both operators and the broader community they serve. This alignment, she explained, involves considering the complexity of ethical norms and values and recognizing the influence of fundamental rights, particularly within the European perspective.

Building upon the European context, Haataja highlighted the ethical principles and guidelines that have been established over the years. These principles revolve around human autonomy, emphasising an individual's capacity to make decisions based on their beliefs and values. This principle becomes particularly relevant in the context of generative AI systems, where concerns about election manipulation and other potential misuse arise.

The second principle discussed was the prevention of harm, which Haataja emphasised in the current context of chatbots and large language models potentially producing harmful content that could lead to physical or emotional harm for users. The principle of fairness, recognizing equality and avoiding discrimination in AI systems, was also deemed crucial but acknowledged as easier said than done. Haataja pointed out the challenges in achieving fairness, especially with the rapid deployment of large language models across various use cases.

The final principle discussed was explainability, the capability to understand and interpret models and ideas. Haataja identified this as one of the significant challenges associated with large language models, emphasising the importance of addressing this issue in the current AI landscape.

These projects have received funding from The European Union's research and innovation programme under GA No. 951911 (AI4Media), 951847 (ELISE), 101070617 (ELSA), 101070596 (euROBIN), 761758 (HumaneAI Net), 952215 (TAILOR) and 952070 (VISION).

In conclusion, Meeri Haataja's keynote shed light on the pressing need to weave ethical considerations into the fabric of AI development. Her comprehensive exploration of ethical norms, principles, and the practical challenges posed by evolving AI technologies underscored the critical importance of aligning AI advancements with ethical values to foster responsible and trustworthy AI systems.

These projects have received funding from The European Union's research and innovation programme under GA No. 951911 (AI4Media), 951847 (ELISE), 101070617 (ELSA), 101070596 (euROBIN), 761758 (HumaneAI Net), 952215 (TAILOR) and 952070 (VISION).

## *Responsible AI in the industry*

**About the keynote speaker**

Dr. Marc Steen works as a senior research scientist at TNO, a leading research and technology organization in The Netherlands. He earned MSc, PDEng and PhD degrees in Industrial Design Engineering at Delft University of Technology. He is an expert in Human-Centred Design, Value-Sensitive Design, Responsible Innovation, and Applied Ethics of Technology. Marc will talk about "ethics for people who work in tech" (also the title of his recently published book); he will present a practical approach to integrating ethics in the development and deployment of AI systems.

Dr. Marc Steen, Senior Research Scientist at TNO, delivered a concise yet enlightening keynote on "Responsible AI in Industry." Within his keynote, he provided a glimpse into TNO's role, his perspective on responsible AI, and a practical method for integrating ethics into the development and deployment of AI systems.

Starting with a brief overview, Dr. Steen outlined TNO as an organisation with 3000 dedicated individuals conducting applied scientific research, strategically positioned between academia and industry. Over the past decade, Dr. Steen has specialised in applied ethics, particularly focusing on data, algorithms, and what is currently referred to as AI.

He recounted a shift in perception within TNO, as colleagues sought his approval for projects, treating ethics as a potential barrier to be overcome. Dr. Steen, an engineer by profession, proposed an alternative viewpoint, presenting ethics as a steering wheel rather than a barrier. In this analogy, projects become vehicles, and ethics functions as the tool to navigate and steer the project in the right direction, avoiding collisions and wrong turns.

Dr. Steen underlined viewing ethics as a process of ethical reflection, inquiry, and deliberation.

He presented a method he developed, named "Rapid Ethical Deliberation." He explained that it was specifically designed to integrate into Agile or Scrum frameworks, aligning with the iterative nature of these methodologies. The method comprises three iterative steps—identification of issues, hosting dialogues with project and external stakeholders, and making decisions based on critical reflections—and four ethical perspectives: consequenstialisme, which looks at pros and cons; duty ethics, which deals with duties and rights; relational ethics, which looks at interactions and power; and virtue ethics, which views technologies as tools to cultivate relevant virtues. There is a Canvas available for this: https://ethicsforpeoplewhoworkintech.com/.

Highlighting the iterative nature of the process, Dr. Steen emphasised that it is not a linear sequence but a continuous loop. He acknowledged the challenge of obtaining quick answers in just one hour but noted that the process leads to more precise and specific questions that can be addressed over time. The method, in essence, serves as an ongoing experiment, allowing teams to modify their projects and be accountable for the ethical considerations throughout the development and deployment phases.

These projects have received funding from The European Union's research and innovation programme under GA No. 951911 (AI4Media), 951847 (ELISE), 101070617 (ELSA), 101070596 (euROBIN), 761758 (HumaneAI Net), 952215 (TAILOR) and 952070 (VISION).

In conclusion, Dr. Marc Steen's keynote provided a valuable insight into embedding responsible AI into industrial practices. His emphasis on ethics as a steering wheel, coupled with a practical and iterative ethical deliberation method, offered a thoughtful approach for navigating the evolving landscape of AI with responsibility and accountability.

These projects have received funding from The European Union's research and innovation programme under GA No. 951911 (AI4Media), 951847 (ELISE), 101070617 (ELSA), 101070596 (euROBIN), 761758 (HumaneAI Net), 952215 (TAILOR) and 952070 (VISION).

# Key results from the Breakout Sessions

**Session 1: AI explainability for vision tasks**

This session discussed the current capabilities of AI explainability methods for visual data classifiers and other vision tasks; how explanations can be presented to the data scientist / end-user; what we can expect to understand from the provided explanations; and, the next steps towards advanced AI explainability.

The discussion made clear that using AI to address vision tasks is widely applicable in various sectors; and, the need for introducing and exploiting AI explainability in addressing such vision tasks is equally widespread. Need for and potential usage of AI explainability includes but is not limited to i) applications in academia, such as supporting AI researchers working on developing visual content classifiers to identify biases and other deficiencies in their classifiers; i) applications in the infotainment and creative industries sector, e.g. in visual content recommendation, to help the users understand better how the recommendation system works; iii) medical-domain applications, e.g. in relation to AI-based medical image analysis, for supporting physicians in interpreting and trusting AI; iv) applications in manufacturing / production line monitoring / Industry 4.0 / predictive maintenance; v) Safety and security applications involving visual information, e.g. the analysis of video streams.

The **key challenges** that emerged from the discussion are:

- The need for advanced explainability methods to generate explanations that are accurate and informative (e.g. in the case of an image classification model, explanations that are not limited to highlighting where in the image the model "looks", but also what exactly is "sees")
- How to communicate explanations to the end-users, and how to maximise their usefulness to them
- The inherent diversity of the explainability problem: different classes of AI methods and different applications require different explainability methods and different forms of explanations
- How to handle the potential trade off between the visual classifier's accuracy and its level of explainability
- The possibly significant computational cost of deriving the explanations
- The challenge of assessing the goodness of the explainability methods and of their explanations
- The existence of expertise and time barriers for introducing the outcomes of AI explainability research in industry

In response to these challenges, the session participants formulated the following five recommendations:

1. **Intensify research on developing explainability methods** that are accurate and informative, without compromising the accuracy of the classification / analysis method

These projects have received funding from The European Union's research and innovation programme under GA No. 951911 (AI4Media), 951847 (ELISE), 101070617 (ELSA), 101070596 (euROBIN), 761758 (HumaneAI Net), 952215 (TAILOR) and 952070 (VISION).

whose results we want to explain (i.e. avoid introducing an accuracy-explainability trade off), and without introducing significant computational overhead (i.e. we should focus on Green-XAI)

2. **Embrace the diversity of the explainability problem**: acknowledge that we will need explainability methods designed / adapted to different AI models and applications
3. **Understand what forms of explanations** are most useful for each target user group, and direct research efforts accordingly
4. **Develop reliable evaluation protocols** for assessing the goodness of the explanations / explainability methods
5. **Lower the expertise / time barriers** for introducing explainability in industry; emphasise research and development of industry solutions where explainability is embedded in the overall AI solution, rather than leave it to the end-user industry players to introduce XAI research outcomes in their existing AI infrastructure

**Session 2: Ethical considerations and new challenges of Generative AI**

This session aimed to explore the risks and challenges raised by generative AI from an interdisciplinary perspective (legal, ethical, societal, technical, cybersecurity). Lately, ChatGPT, DALL-E and others have been massively used by the public since their release. Many scholars and civil society representatives expressed concerns about the short, mid and long term effects of the use of generative AI. After the ban of ChatGPT by the Italian DPA and the call to pause the training of AI systems more powerful than GPT-4, the sessions aimed to reflect on the ethical, legal and technical challenges and what safeguards are necessary to address those.

The discussion delved into the meaning generated by generative AI and the question of responsibility, concluding that humans should bear responsibility for AI's output. The importance of ethics in technology design was emphasised, as design choices can lead to unethical outcomes, even if technology itself is "neutral". Addressing bias was discussed, and it was suggested that while bias can be mitigated, it may never be entirely eliminated. The role of users in judging the quality of AI output and the need for checks and balances were also explored. Regarding innovation, the debate touched on whether slowing down innovation is feasible, with considerations about its impact on big players and the potential favouring of established companies.

The discussion identified the following **challenges**:

1. **AI Anthropomorphism & Manipulation:** The challenge here lies in the tendency of users to attribute human-like qualities to AI systems. This can lead to the belief that these systems possess ethical thinking or moral considerations, which they do not. Additionally, generative AI's potential for emotional manipulation, such as using persuasive language or manipulating user emotions, raises ethical concerns.
2. **The Scale of Generative AI Use and Development:** With the widespread adoption and development of generative AI, the scale of its use poses challenges. Ensuring responsible

These projects have received funding from The European Union's research and innovation programme under GA No. 951911 (AI4Media), 951847 (ELISE), 101070617 (ELSA), 101070596 (euROBIN), 761758 (HumaneAI Net), 952215 (TAILOR) and 952070 (VISION).

development, oversight, and ethical use become more complex as generative AI systems proliferate across various industries and applications.

3. **Addressing Responsibilities:** Determining who bears responsibility in the generative AI ecosystem is a multifaceted challenge. In addition, different liability questions arise (civil, criminal, for instance when someone commits suicide following discussion and encouragement by a generative AI based chatbot.
4. **Ethics by Design**: This challenge emphasises the importance of incorporating ethical considerations into the design phase of generative AI systems. Design choices can have significant ethical implications, and proactively integrating ethical principles into the design process is essential to mitigate potential issues.
5. **Solving the Bias:** While it's acknowledged that biases in generative AI are inevitable to some extent due to training data, the challenge is determining when biases are acceptable and when they are not and who should decide about this crucial aspect. Moral compasses are needed to achieve this but the unacceptable bias is when the bias diminishes individuals based on irrelevant aspects.
6. **Imbalance of Power:** There is an inherent power imbalance between generative AI providers, society at large, and end-users. Providers have significant control over the technology and its impact. Balancing this power dynamic and ensuring that the interests and values of society and individual users are considered is a complex challenge that needs to be addressed.
7. **Balancing Individual Micro-decisions and societal interests:** Striking the right balance between individual choices and broader societal interests when it comes to generative AI is a delicate ethical challenge.

The **following recommendations** can be elaborated from the discussion:

- Allocate clear responsibilities to humans involved in generative AI lifecycle.
  Developers, providers, users, and end-users all play distinct roles, and delineating their responsibilities, particularly in the context of ethical use and potential consequences, is a critical task.

- Enhance Transparency and Accountability:
  For the output of generative AI, ensure greater transparency by incorporating labels, detailed explanations regarding data sources and traces, and clear information about the parameters used in generating responses. Transparency builds trust and understanding among users. Transparency and awareness about generative AI can also be improved through education, through obligations for providers and developers, through better interface and parameters for end-users.

- Foster User Engagement and Feedback:
  Go beyond simplistic feedback mechanisms like like and dislike buttons for generative AI system outputs. Implement user engagement features that allow users to provide contextual feedback, report issues, and offer suggestions for improvement. Creating more robust feedback channels enhances the performance and accountability of the AI system.

These projects have received funding from The European Union's research and innovation programme under GA No. 951911 (AI4Media), 951847 (ELISE), 101070617 (ELSA), 101070596 (euROBIN), 761758 (HumaneAI Net), 952215 (TAILOR) and 952070 (VISION).

- Prioritise Ethics and Interdisciplinary Collaboration in innovation
  Embed ethical considerations and encourage interdisciplinary collaboration throughout the whole generative AI lifecycle and from the design stage. This approach promotes the alignment of AI with ethical principles and ensures that diverse perspectives are taken into account during development. There is a pressing need for more in-depth conversation about innovation, ethics and the future of society.

- Mitigate AI Anthropomorphism, especially in Chatbots:
  In the design of generative AI systems, particularly chatbots, make a deliberate effort to prevent AI anthropomorphism. Clearly convey to users that these systems lack human-like qualities and ethical reasoning. This helps users maintain realistic expectations and responsible interactions with AI.

- Conduct research about inevitable biases
  Research should be conducted about what can be considered as acceptable bias when it appears inevitable even after mitigation measures. In addition, how and who should assess when bias is diminishing people and what are the limits society can accept.

The challenges surrounding generative AI encompass a wide array of ethical, societal, and technical considerations. Addressing these challenges requires collaboration among various stakeholders, a commitment to ethical design, and ongoing efforts to ensure the responsible and equitable use of generative AI technology.

**Session 3: Rigorous vs empirical AI privacy: Where is the middle ground for defining and evaluating privacy in complex algorithms?**

This session discussed the tradeoffs between using empirical evaluations of privacy leakage against relying on rigorous definitions of privacy protocols in releasing data and AI models. While definitions like differential privacy (DP) provide a robust, mathematical guarantee of privacy, the choice of privacy budget is an important consideration. The current best practice is to use DP with a large privacy budget, and rely on empirical evaluations of privacy attacks to prove privacy. Given that this strategy has failed for other protocols in the past, is it an appropriate one for DP?

The discussion made clear that empirical privacy evaluations are here to stay and are a necessary requirement to justify the need for Privacy-Enhancing Technologies (PETs) and DP in machine learning and AI systems, but also to demonstrate their impact in preventing privacy leakage. In particular, membership inference and database reconstruction attacks were discussed. The **key challenges** that were identified were as follows:

1. Differential privacy a proven method to achieve privacy in AI models, but no standardisation in privacy loss values (epsilon) to use
2. Complex algorithm require larger privacy loss
3. “Accuracy first” method common for choosing epsilon
4. Epsilon is abstract and unintuitive

These projects have received funding from The European Union's research and innovation programme under GA No. 951911 (AI4Media), 951847 (ELISE), 101070617 (ELSA), 101070596 (euROBIN), 761758 (HumaneAI Net), 952215 (TAILOR) and 952070 (VISION).

5. Privacy attacks (membership inference, database reconstruction) used to demonstrate/benchmark effectiveness of PETs - but give no proof of general privacy

In response to these challenges, the following recommendations were adopted by the session:

1. **Libraries of privacy attacks to justify and validate PETs:** An open source library of attacks that can be easily executed on ML and AI systems would provide a concrete baseline for comparison of the privacy leakage of models.
2. **Standardisation of epsilons for various tasks may not be practical:** Although desirable, the practicality of a standardised epsilon (privacy loss) for various tasks would be difficult to implement, given the variety of tasks and data being used in ML and AI applications.
3. **Improvements sought on "accuracy-first" approach:** The current state-of-the-art is to choose the privacy loss (epsilon) parameter via accuracy-first (choose an epsilon that achieves a desired accuracy), but doing so goes against the universality of epsilon as a privacy parameter. Alternative approaches are desired.
4. **More dynamic metric beyond epsilon (confidence interval, chart/graph):** When publishing AI models with DP guarantees, a multi-dimensional approach to reporting the privacy loss is proposed. This could include a graph or confidence interval around the chosen privacy loss, for example.
5. **Increase the awareness of privacy in the broader data community:** There are privacy risks associated with the use of any personal or sensitive data, and the use of such data is typically restricted because of this privacy risk. However, many of these risks can be ameliorated by the use of appropriate PETs, and increasing the awareness of these tools could allow for greater sharing of data and results in a safe manner.
6. **Increase the accessibility of tools and technologies (and perhaps theory) such that non-experts can also have a strong grasp of the field:** DP and PETs are necessarily complex tools, and it is the duty of researchers in the field to provide simple explanations and demos of their tools to promote wider adoption by non-experts.

**Session 4: Monitoring progress in interpretable AI**

It is well-studied how to measure the accuracy of machine learning predictors; it is less trivial to monitor progress in developing models interpretable by humans. We brought together an interdisciplinary group of participants (legal, regulational, technical aspects) to outline the requirements for such monitoring and possible ways to approach this problem.

Within the breakout session on monitoring processes in interpretable AI, the collaborative efforts between the participants yielded key insights and results that shed light on the intricacies of ensuring transparency and progress in the development of AI models interpretable by humans. Concerning that, they following key results have been identified:

1. **Necessity of Common Benchmarks and Interpretability Challenges**
   the discussions underscored the imperative need for standardized benchmarks and challenges to evaluate progress in interpretable AI. Establishing a common ground for assessment is vital to ensure a cohesive understanding of advancements in the field.

These projects have received funding from The European Union's research and innovation programme under GA No. 951911 (AI4Media), 951847 (ELISE), 101070617 (ELSA), 101070596 (euROBIN), 761758 (HumaneAI Net), 952215 (TAILOR) and 952070 (VISION).

2. **Appreciation of Domain Difficulties**
   The participants collectively acknowledged the inherent challenges within the domain, recognizing the multifaceted nature of interpretability. By appreciating the complexities, we can tailor our approaches and solutions to address the specific hurdles encountered in the quest for interpretable AI.
3. **Exploration of Disentanglement, Attribution, and Causal Discovery**
   A significant part of our dialogue revolved around exploring key concepts such as disentanglement, attribution, and causal discovery. These concepts emerged as pivotal elements in monitoring progress toward interpretable AI. Understanding how these factors interplay contributes to a more nuanced and comprehensive assessment.
4. **Interdisciplinary Collaboration for Progress**
   The discussions also emphasized the need for a concerted effort involving experts from diverse fields, including AI, legal, ethical, and regulatory scholarship. The collaborative exchange of ideas and perspectives is essential to navigate the evolving landscape of interpretable AI and overcome its associated challenges.

By delving into these key results, the breakout session laid the foundation for a more informed and collaborative approach to advancing interpretable AI. The recognition of challenges, exploration of fundamental concepts, and the call for interdisciplinary collaboration collectively contribute to a roadmap for the continued development and monitoring of interpretable AI systems.

**Session 5: Causality and Trust**

Causal models can improve the trustworthiness of AI systems (Causality for Trust, C4T). Besides precision and accuracy, which are fundamental to trustworthiness in AI, they are transparent, reproducible, fair, robust, privacy-aware, safe and accountable.

After a motivating and inspiring introduction by Prof. Nejdl (Leibniz University of Hannover), there appeared to be a general consensus that causality can contribute to trust and trustworthiness. However, there was a debate with diverging ideas about how this can be achieved, which left open whether causality should be considered as a requirement to create trust. The omission of causality as a requirement for trust in the AI Act was debated. Should it be required or mentioned as a desirable feature for enabling several ethical characteristics?

The following challenges were identified:

- **How to come up with causal models?** They are not always obvious to define. For example, we discussed Simpson's Paradox. This is a well-documented, but often neglected phenomenon in statistics that cannot be addressed without a causal model. It results in different outcomes depending on how one looks at the same model. How to find the confounding variables that are actually explaining why these effects occur?
- **How to teach causality to Large Language Models (LLMs)?** Can reinforcement learning be used as random controlled experiments? In a first approach, causal models could be used to modify and improve the LLM. Causal reasoning could explain certain statements that are made by the systems and verify their plausibility. Another approach would include

These projects have received funding from The European Union's research and innovation programme under GA No. 951911 (AI4Media), 951847 (ELISE), 101070617 (ELSA), 101070596 (euROBIN), 761758 (HumaneAI Net), 952215 (TAILOR) and 952070 (VISION).

having certain tools that are add-ons. An example is Toolformer, which creates a hybrid or neuro-symbolic system augmenting the output of the LLM. Alternatively, one could use language models to create causal models by extracting the implicit knowledge with reinforcement learning to perform experiments to create causal models.

- **How can causal models improve explanations?** It was unanimously seen as beneficial to have a causal explanation of why certain things happen, why decisions are taken, or why classifications are made. On the other hand, is it necessary that every explanation is based on a causal model or are there other ways that can also be convincing and trustworthy? Similarly, we discussed accountability and responsibility, with accountability as an actionable tool to enforce rules and requirements for trustworthy systems and in how far causality could contribute to such accountability assessments.

Industrial Applications

- Time series analysis and forecasting.
- Monitoring of causal relationships of machines and processes.
- Anomaly detection, root cause analysis.
- Verification of legal requirements.

In general, it's advantageous to know the reason why something goes wrong.

The consensus was that causal explanations will increase trust in AI. It's not always trivial how to find those models, but if we can, we should invest in appropriate research in the relationship between existing models and causal models and how they can improve each other.

**Session 6: Robustness/Verification**

This session looked at technologies to strengthen the secure use of AI technologies. There has been a general introduction to the topic of Machine Learning robustness, including recent relevant technologies developed by the research community. The discussion included some aspects of certifiable robustness, resilience and recovery, and uncertainty and safety in decision-making. Finally, the relationships between robustness and privacy, explainability, and fairness has been discussed to complete the overview of requirements devised to achieve trustworthy ML.

The participants identified the following 5 **key challenges**:

1. AI systems should cover strong assurance requirements (robustness, fairness, accountability, transparency, privacy), and there is not one method that covers all aspects or applications, defining a suitable set of methods is challenging;
2. Human validation is often not feasible, there are many possible metrics but they are often specifically designed for each paper / case and almost never validated. Research is needed to define such metrics and enforce standardised evaluations and benchmarks.
3. The formulation of these requirements is often application-specific;
4. There is a lack of benchmarks depicting industrial applications; and

These projects have received funding from The European Union's research and innovation programme under GA No. 951911 (AI4Media), 951847 (ELISE), 101070617 (ELSA), 101070596 (euROBIN), 761758 (HumaneAI Net), 952215 (TAILOR) and 952070 (VISION).

5. Application-specific risks should be defined via specific objectives and bridged with suitable means of compliance, thus the knowledge of domain experts is required to achieve the overall coverage of all risks.

The identified key challenge and the robustness requirements in general are important in many industrial applications and have different specifications depending on the domain. However, specific applications should consider these requirements more carefully as the decisions involve higher risks and the consequences of taking wrong actions might cause great harm. Cybersecurity, Finance, Healthcare, Transportation (Railway, Aerospace, Autonomous driving, etc ), Logistics are fields in which robustness and trustworthiness should be paramount.

Then, the participants gathered the following **key recommendations**:

1. **Consider the context where your model is going to work on.** Some scenarios have specific requirements that should be considered. For example, cybersecurity is a field in which attackers are implicitly present, so robustness of ML should also envision the possibility of adversarial attacks. In other cases, depending on the available data and the quality of data, a thorough analysis should be performed to avoid biases and reduce the influence of spurious correlations in the data. If the operational domain of the model can be affected by noise or unexpected inputs, it is advised to add redundant and complementary systems to ensure that the model is considered from different angles (security, efficiency, data drifts, …).
2. **Privacy preservation.** The real information and needs of the users have to be considered, and the appropriate ways for conveying the right info should be implemented to ensure full understandability (not only transparency). ML should be specifically protected if sensitive data are used, as attackers can exploit weaknesses of the storage systems (with traditional cyberattacks) or weaknesses specific to the ML system that is used (with novel privacy ML attacks, for example membership inference or model stealing attack).
3. Start from the risk and build the rest of the requirements
4. **Leverage debugging techniques to ensure the model is learning strong causal relations.** These techniques include the use of explainability techniques and analysis of the single errors. Explainability techniques can be used on top of existing models (post-hoc methods) to give insights on what the model is learning. The use of these techniques can reveal if the model is relying on patterns that should not be considered as strong features (either for fairness reasons, leveraging biases in the data, or for robustness reasons, when the model relies strongly on features that are easier for an attacker to modify).

The **key results** of the session highlighted that ML robustness not only finds weaknesses of ML models, but is also useful to understand the limits of these technologies, and can be helpful to design models that are more aligned with human decisions and understanding. Knowing when to trust automated decisions, especially in high-risk contexts, is extremely important to really make use of ML in the best possible way.

These projects have received funding from The European Union's research and innovation programme under GA No. 951911 (AI4Media), 951847 (ELISE), 101070617 (ELSA), 101070596 (euROBIN), 761758 (HumaneAI Net), 952215 (TAILOR) and 952070 (VISION).

**Session 7: AI/ML Benchmarking**

This session reflects on ways and methodologies for evaluating AI/ML solutions in real-world conditions. The discussion will touch upon the best practices for defining meaningful benchmarks, the present and future of AI/ML benchmarking, reproducibility and specific ways and challenges of measuring aspects of systems' trustworthiness on the road towards Creating Ethical & Responsible AI Systems.

Within this breakout session, the participants explored the complexities of assessing AI/ML solutions in real-world scenarios, underscoring the importance of establishing meaningful benchmarks as we progress toward the development of ethical and responsible AI systems.

5 Key Challenges:

1. **Benchmark Definition**
   Clearly describing what a benchmark is measuring emerged as a primary challenge, addressing the need for precision in task design to tackle scientific questions effectively.
2. **Inventor-Evaluator Bias**
   The persistent presence of bias in inventor-evaluator dynamics, particularly in tasks related to AI/ML benchmarking, posed a significant challenge.
3. **Measuring "Ill-Defined" Properties**
   Practical challenges were identified in measuring the performance of "ill-defined" or qualitative properties of AI/ML systems, including issues related to reproducibility, the reality gap, and quantifying notions such as fairness.
4. **Human Element in Human-in-the-Loop Schemes**
   Dealing with the human element in human-in-the-loop schemes raised questions about participant selection, the number needed, and whether experts or non-experts should be involved. The consideration of compensating individuals for evaluating AI systems added another layer of complexity.
5. **Transparency in Evaluation**
   Establishing a transparent evaluation procedure with clearly defined criteria posed a challenge, emphasizing the need for clarity in the assessment process.

5 Key Recommendations:

1. **Transparent Evaluation Procedure**
   Adopting a transparent evaluation procedure with clearly defined criteria was recommended as a foundational step to address challenges related to benchmarking.
2. **Maximizing Quantity of Tests**
   Maximizing the quantity of tests was identified as a key recommendation to ensure more significant evaluations, providing a robust understanding of AI/ML system performance.
3. **Crowdsourced Benchmarks**
   Advocating for crowdsourced benchmarks was highlighted as a strategy to mitigate inventor-evaluator bias, enhance experiment replicability, and diversify the methods under comparison.

These projects have received funding from The European Union's research and innovation programme under GA No. 951911 (AI4Media), 951847 (ELISE), 101070617 (ELSA), 101070596 (euROBIN), 761758 (HumaneAI Net), 952215 (TAILOR) and 952070 (VISION).

## Session 8: AI Ethics: from principles to practice. Putting "ethical" and"responsible" AI into action

The session started with referring to the controversial statement by L. Munn about the flood of AI guidelines and codes of ethics which contain "meaningless principles which are contested or incoherent, making them difficult to apply; (...) isolated principles situated in an industry and education system which largely ignores ethics; and (...) toothless principles which lack consequences and adhere to corporate agendas" (Munn 2023). The participants have indicated that such broad generalisations do not help the cause and pointed out that despite drawbacks, the AI ethics is an important field of (applied) research. On the other hand there is an element of truth in those bold statements.

The **key challenges** identified during the discussion concern:

1) **Proliferation of AI ethics guidelines**, a lack of actual impact the AI ethics guidelines have, a lack of enforcement mechanisms in case of non-compliance and no robust regulatory mechanism to govern ethical AI.
   The proliferation of the AI ethics guidelines is not only the matter of industry-capture, but it is also due to the academics: many have shifted to AI ethics research because of funding opportunities. Whereas this is not problematic *per se*, what is alarming is a misalignment between academic interest in AI ethics as a research field and what is needed in practice. It was also suggested that having many different ethical AI frameworks may be beneficial because of the variety of orientations they apply to. However, to be meaningful, they should be industry and/or use-case specific.

2) **Imbalance of funding between private sector and public sector.**
   The participants indicated that in the last decade we observe a massive imbalance in resources and talent between private and public sector, aggregated by the fact that currently, 70% of individuals with PhDs in AI find employment in the private sector. To this end, it is a private sector-centred logic that drives what we, as a society, focus on. More funding is needed to develop technology which prioritises public, and not private, values. The approach to AI ethics should move from a reactive approach to a more anticipatory approach. Much more reflection is needed on the question why we need to adapt societal values to fast-pace technology and not the other way around; and on the topic of techno-moral change, i.a. how technology changes our values.

3) **Principle-based approaches to AI Ethics have (to some degree) failed.**
   An argument was made that the principle-based approach to AI ethics has failed. That is because it is unclear how to evaluate and balance values against each other, how to implement them in technical systems, and how to enforce them in practice. However, the participants pointed out the role which principles can play in regulation, namely they can be a good starting point for discussion.

These projects have received funding from The European Union's research and innovation programme under GA No. 951911 (AI4Media), 951847 (ELISE), 101070617 (ELSA), 101070596 (euROBIN), 761758 (HumaneAI Net), 952215 (TAILOR) and 952070 (VISION).

4) **A need for a novel set of interdisciplinary skills and on-going governance required to embed ethics in the entire cycle of AI development**: from concept development to evaluation.
Responsible development of technology requires groundwork, implementation of the processes, documentation, multi-disciplinary collaboration, stakeholder convening, a skills set different from what most academics, ethicists and philosophers traditionally do.

5) **Challenges of compliance with legal obligations of risks identification and mitigation and conducting Fundamental Rights Impact Assessments for SMEs and not compliance-oriented organisations.**
The participants also discussed a regulatory approach to AI ethics through the lens of the AI Act proposal. It was pointed out that the AI Act proposal has two main aims when it comes to AI ethics: i) harmonisation of the vocabulary; ii) making principles enforceable. Experts pointed out that the AI Act does not contain a specific list of ethical principles, but rather requirements which are based on ethical principles. To illustrate, a human agency and oversight principle translates into auditing and impact assessments requirements. Similarly, a transparency principle translates into a requirement of the disclosure of the datasets for the foundation models. It was also mentioned that the AI is still not very clear which impact assessments will be mandatory for each AI application. In the discussion in the Council, there is a proposal to introduce an Environmental and Fundamental Rights Impact Assessments as a legal obligation. Recently, more than 100 university professors from all over Europe and beyond called on the European institutions to include a mandatory Fundamental Rights Impact Assessment (FRAIA) for both public and private institutions deploying artificial intelligence (AI) technologies in the future regulation on artificial intelligence. However, critics point out a high compliance cost which such mandatory assessments may cause. Overall, the participants supported the idea that ethical impact assessment and risk identification and mitigation obligations are likely to be legally binding obligations. Yet, the attention was brought to the fact that the private organisations may assess risks in accordance with their own risk assessment methodologies.

The **key recommendations** identified during the discussion concern:

1. **Meaningful involvement of affected stakeholders from the phase of question articulation/problem definition to avoid techno-solutionism.** The participants pointed out a need for an interdisciplinary stakeholder engagement in the question articulation sessions. The starting point of any ethical AI considerations should be a reflection about the problems which people in the particular setting are facing, and whether a technology is even necessary to tackle the problem.

2. **Embedded-ethics approach**. There is a clear need for an embedded-ethics approach which incorporates reflections on potential consequences of AI development throughout the whole process. An embedded-ethics approach is an interdisciplinary process which differs from a 'normal' ethicists' process. It consists of: i) involving stakeholders from early on; ii) prototyping solutions with stakeholders and mapping potential consequences ; (iii)

These projects have received funding from The European Union's research and innovation programme under GA No. 951911 (AI4Media), 951847 (ELISE), 101070617 (ELSA), 101070596 (euROBIN), 761758 (HumaneAI Net), 952215 (TAILOR) and 952070 (VISION).

understanding of the values which are important in the sector. Embedding ethical and societal considerations - in fact prioritising them - requires on-going efforts instead of one-off assessments

3. **Values that Matter approach, iterative prototyping-based approach.** The participants also pointed out other value-sensitive design-like approaches: Values that Matter approach from the University of Twente (Smits, M, Bredie, B, van Goor, H & Verbeek, P-P 2019) which consists of translating required values into specifications and then empirically checking them and refining the model and prototyping-based approach. It was pointed out that the approach to AI ethics should be a positive one, ambition and aim oriented, and not a negative one pointing out what is *not* supposed to happen (e.g. do not harm).

4. **A need to re-focus the conversation from high-level principles to AI justice.** The participants concluded that while AI ethics in corporate settings can maintain the existing power relations, justice takes a broader perspective to challenge the status quo. AI justice offers a new lens to look at technology in practice.

5. **Principled-based approach in the AI Act can be useful, but it should be contextualised.** The participants agreed that there is a need for a use-case centred approach to ethics guidelines. The discussion has also centred around the tools for ethics assessments. Certain best practices models have been identified: i) in the context of the EU funded DARLENE project an empirical research was conducted to understand the values which are important for the stakeholders; ii) EU funded ALIGNER project has developed metrics how to measure fundamental rights risks: Fundamental Rights System Assessment and AI System Governance Impact; (iii) Fundamental Rights Impact Assessments (FRAIAs) are used by Dutch municipalities; (iv) the AI ethics maturity model which offers a holistic maturity framework (Krijger, J., Thuis, T., de Ruiter, M. et al. 2023). The participants noted that, sadly, there is not much awareness about these models and not much interest from the developers. Encouragement in this field would be welcomed. When discussing the AI Act, the participants also opted to leave the research out of the scope of the Act. Instead, guidelines for the Ethics Boards at the academic institutions may prove beneficial.

6. **Voluntary only AI ethics are not sufficient.** The participants agreed that enforcement of ethical AI principles and requirements is needed. There is a need for an enforcement agency with sufficient resources, knowledge, and skills. It was pointed out that the AI Act is not clear on the fact who will be responsible for this enforcement and there is a risk of lack of harmonisation if every Member State will do it on its own.

### Session 9: Meaningful Human Shared Control

Over the past years, we have seen a number of guidelines promoting 'human-in/on/out of-the-loop' approaches to ensure human control and oversight over AI systems. However, mere human presence alone is not sufficient to ensure such control. Instead, there is a need for a dynamic allocation of tasks between AI systems and human operators/overseers who are not just static observers but rather fully understand their own responsibilities. This topic explores the interplay

These projects have received funding from The European Union's research and innovation programme under GA No. 951911 (AI4Media), 951847 (ELISE), 101070617 (ELSA), 101070596 (euROBIN), 761758 (HumaneAI Net), 952215 (TAILOR) and 952070 (VISION).

between the dynamic transfer of tasks and ensuring the long-term control over the socio-technical system.

Participants identified the following **key challenges**:

1. Defining and effectively prioritising stakeholders and values that populate a certain sociotechnical system or context of technological development. This is a challenge because of the inherent unclarity of the loci of control in the interaction between intelligent machines and human agents, and because of the sheer number of stakeholders that can be identified for any given case scenario (ranging from developers to policy makers to final users and more)
2. Defining the most effective control strategies, which may be context dependent and there might be no solution fitting all cases. For instance, traded control (where a human agent completely relinquishes control at some point in time) might offer advantages in certain cases, while a symbiotic, dynamic interaction (where the amount of contribution may e.g. dynamically and continuously vary) might be recommendable in other cases. Establishing what we want to achieve (i.e. the values at stake) in which cases is again essential to design effective control strategies.
3. Defining effective mechanisms of responsibility attribution through forms of control that can grant a meaningful (self-)attribution of responsibility across the different controllers and agents that populate a sociotechnical system. This is a challenge due to many factors affecting human AI interaction, such as opacity, unpredictability, delusions of agency and so on.
4. Defining societal desirability (vis-à-vis technical feasibility) of very high degrees of (or even full) autonomy. This challenge is about determining where, when and for which reasons, certain contexts may require that, in principle, the use of automated systems is limited. Examples of contexts where this is crucial, also from the literature, may be autonomous warfare and judicial decision making.
5. Achieving conceptual clarity. It is insufficiently clear what control means in different disciplines, and even pragmatic definitions tend to change significantly across different disciplines, e.g. engineering and law.

Participants identified the following **actionable insights and policy recommendations:**

1. The controlled environment should be assessed case by case and different strategies and degrees of autonomy should be defined avoiding one solution fits all strategies where possible.
2. Tasks and functions should be automatised in a step by step fashion, gradually replacing atomic tasks while continuously auditing for unwanted consequences.
3. Broad-level, multi-stakeholder discussion fora should be formed to assess the societal desirability of AI in different scenarios. This is a normative dimension that inherently concerns intersubjective societal values and needs, and explicitly abstracts from technical feasibility.
4. A common language between the multiple stakeholders should be strived for, and this could be done through philosophical tools and techniques, e.g. hermeneutic analysis or

These projects have received funding from The European Union's research and innovation programme under GA No. 951911 (AI4Media), 951847 (ELISE), 101070617 (ELSA), 101070596 (euROBIN), 761758 (HumaneAI Net), 952215 (TAILOR) and 952070 (VISION).

conceptual engineering. This includes conceptual clarification of the taxonomies of control and responsibility, especially in the light of the particularly opaque relations established in the context of human-AI interaction.

**Session 10: Human Oversight and Explainability for AI**

This session analysed the architectures, mechanisms and methods capable of generating meaningful and evidence based assurance which is necessary to secure and maintain the safety and security dimensions of AI systems. Through interdisciplinary investigation between technical experts and legal, ethics and governance experts it evaluated the existing and emerging methods and mechanisms. Further, new approaches were discussed, for establishing and certifying safe and secure AI to deliver meaningful and effective AI assurance including meaningful and effective human oversight.

Participants identified the following **key challenges**:

1. Legal, regulatory (including standardisation) frameworks of such technologies is a challenge in itself partially due to differences in taxonomy, interpretation, quantification, etc.
2. Metrics, benchmarks, measure/evaluate if and to what extent an application is explainable
3. Who is the user does influence the explainability (experience, information need, context of the task)
4. Latent spaces prove to be powerful but the correspondence between latent and human interpretable features is often a stumbling block
5. What can humans learn from explanations? (validity, generalizability, reliability of models AND data)
    a. Separability and distinction of concepts to be explained is not always straightforward and interpretable
    b. Lack of representative and complete data for evaluation
6. The data are becoming an undetachable part of the pre-trained (PT) models, but they are often not openly accessible, may not follow different requirements, e.g. GDPR, copyright, etc.
7. Causality is often being substituted by statistical likelihood and plausibility which are not the same and this hinders the explainability
8. Assessment and interpretation of the Risk

The main **industrial applications** were identified:

1. Autonomous driving
2. Health
3. New Chemicals, drugs, toxicity
4. Judicial system, bailing from jail (COMPAS, bias)
5. Earth Observation

Participants further identified the following **key recommendations**:

These projects have received funding from The European Union's research and innovation programme under GA No. 951911 (AI4Media), 951847 (ELISE), 101070617 (ELSA), 101070596 (euROBIN), 761758 (HumaneAI Net), 952215 (TAILOR) and 952070 (VISION).

1. Develop a metric and benchmarks for explainable and trusted/safe AI/ML
2. Regulatory and legal interventions to be balanced - to safeguard the users and citizens, yet to provide a fair ground for development and advance
3. New methodologies and techniques to be developed that address the key challenges
4. Different levels of explanations to address different users.
5. FAIR principles on data, especially for pre-trained models.
6. Transparent error analysis where specialists can audit AI systems when necessary.
7. Equip data-heavy end-to-end approaches with reasoning

The **key results** can be summarised as follows:

1. Inter-/multidisciplinary collaborations are critically important (the workshop was a mini-example)
2. Identified the need for metrics to be developed as well as new methods, techniques, benchmarks
3. Data transparency and diversity (e.g. minority languages, images, etc.) and "data literacy" (prepare the society and users)

**Session 11: Trusting Each Other**

For collaborative decision-making it is essential that each human and agent is aware of each others' points of view and understands that others possess mental states that might differ from one's own – which is known as a Theory of Mind (ToM).

In this breakout workshop, the key challenges on the topic of mental modelling in hybrid human-AI team settings were discussed and identified. As trustworthiness is a concern when humans interact with AI agents, understanding each other becomes paramount. Establishing the level of trust in human-agent interactions becomes a central issue, as it can deeply influence the effectiveness and acceptance of AI systems.

The concept can play a significant role in various fields. For example in Autonomous Driving Situations systems of multiple vehicles have to take into account the behaviour of others sharing the road. Vehicles use communication and coordination to avoid collisions, share information about road conditions, and make lane-changing decisions. On a higher level, multi-agent systems can assist in optimising traffic flow at intersections, reducing congestion, and improving overall traffic efficiency. Vehicles could also negotiate routes to minimise travel time or energy consumption while considering real-time traffic conditions and environmental factors.

In the field of energy-related negotiations, multi-agent systems are employed to optimise energy distribution, consumption, and resource allocation. Key roles include: Smart Grids: where multiple agents (e.g., buildings, households, power plants, renewable energy sources) negotiate energy production, distribution, and pricing. This ensures efficient energy use and minimises wastage. Agents and humans can collaborate to balance energy loads by shifting demand to off-peak hours, reducing strain on the grid during peak times. In warehouse automation, multi-agent systems are used to optimise logistics and coordinate the actions of robots. They could take into account each

These projects have received funding from The European Union's research and innovation programme under GA No. 951911 (AI4Media), 951847 (ELISE), 101070617 (ELSA), 101070596 (euROBIN), 761758 (HumaneAI Net), 952215 (TAILOR) and 952070 (VISION).

other and negotiate and allocate tasks, such as picking, packing, and transporting goods, to optimise overall warehouse efficiency. Robots as well as humans can take into account their navigation through the warehouse while transporting goods, and while avoiding collisions and congestion.Agents and humans work together to manage inventory levels, ensuring that products are restocked when needed and that warehouse space is efficiently utilised.

In Healthcare one can think of collaboration between healthcare professionals and AI agents in advisory roles. For example in clinical decision support: AI agents can provide recommendations and insights to healthcare professionals based on patient data, medical literature, and best practices. Humans and agents can use ToM while diagnosing and treating patients. Or in the setting of allocate resources like beds, staff, and equipment efficiently, especially in emergencies.

Typical scenarios involve the need for negotiations whilst having incomplete information. Scenarios where multiple entities need to interact, negotiate, and collaborate to achieve optimal outcomes. ToM mechanisms improve efficient decision-making, resource allocation, and coordination in complex, dynamic environments across a wide range of applications.

The workshop also discussed the challenges that need to be addressed to make the scenarios mentioned above successful:

The concept of Theory of Mind (ToM) is fundamental in understanding how humans and agents in a hybrid setting perceive and interact with one another. To grasp the essence, an example like “Bob thinks that Alice wants to have a drink” is easy to understand. But when transitioning to a modelling or computational perspective, it can be challenging to understand and deal with the underlying mechanisms that implement the ToM.

Doubts naturally emerge. Is the current approach to ToM modelling the sole and optimal method for facilitating interactions in hybrid human-AI teams? Researchers are discussing and comparing ToM approaches with centralised approaches, and game theoretic results with ToM are not yet mapped in real-life scenarios, e.g. in contemporary AI decision-making support. Another critical consideration is distributed reasoning in larger groups and at higher levels of ToM complexity. Achieving scalability poses a serious challenge. Moving forward, assessing the level of ToM reasoning in others and coping with the computational complexity of that in real-time, brute force ToM calculations might reach the computational limits. Heuristics for reasoning might help, but are yet unknown territory. Deciphering whether collaborative or competitive goals should guide negotiations further complicates the landscape. This activity is part of the act of migrating from game-theoretic settings to real-life scenarios.

The challenges and considerations discussed in the breakout session revolve around the complex task of developing artificial systems that can effectively interact with humans, anticipate their behaviour, and foster trust.

Designing and building networked systems, composed of humans and agents, in which the actors are capable of understanding, predicting, and adapting to each other's behaviour is a new challenge for researchers and engineers.

These projects have received funding from The European Union's research and innovation programme under GA No. 951911 (AI4Media), 951847 (ELISE), 101070617 (ELSA), 101070596 (euROBIN), 761758 (HumaneAI Net), 952215 (TAILOR) and 952070 (VISION).

These challenges touch upon various aspects, including expertise in system design, the emergence of trust, and the translation of theory into practice. A key point is the study of how trust naturally emerges in systems that incorporate the concepts of Theory of Mind (ToM) within their negotiation mechanisms. This exploration is essential for building trust in AI systems that can collaborate effectively with humans.

We have to bridge the gap between theoretical insights, particularly from game theory, and their practical application in real-world scenarios containing human-agent interactions. However, a crucial caveat is recognizing the limitations of ToM, as human reasoning is inherently imperfect. The design of artificial systems that interact with humans must consider the human perspective. This raises questions about the behaviour of ToM agents. Should these agents prioritise qualities such as honesty, impartiality, and transparency in their reasoning and decision-making processes when interacting with humans? Striking the right balance between ethical considerations and the functional aspects of AI systems is a pressing concern.

The development of the next generation of AI systems should be a multidisciplinary effort. Addressing these multifaceted challenges requires expertise from diverse fields, including psychology, computer science, ethics, and more. Collaboration across these domains is vital to ensure that AI systems are not only technically proficient but also ethically sound and capable of fostering trust in human interactions. Successfully navigating these challenges will pave the way for the development of AI systems that can navigate the complexities of human interactions and contribute positively to society.

**Session 12: Human-Aligned Video AI**

This session discussed the dual-use of video AI technology. Video-AI holds the promise to explore what is unreachable, monitor what is imperceivable and to protect what is most valuable. This is no longer wishful thinking. Broad uptake of video-AI for science, for wellbeing, and for business awaits at the horizon, thanks to a decade of phenomenal progress in deep learning. However, the same video-AI is also accountable for self-driving cars crashing into pedestrians, deep fakes making us believe misinformation, and mass-surveillance systems monitoring our behaviour. The research community's over-concentration on recognition accuracy has neglected human-alignment for societal acceptance. Therefore, to make video-AI deliver on its big promise, human-alignment is key. But what exactly defines human-aligned video-AI, how can it be made computable, and what determines its societal acceptance?

Four experts provided their perspectives on human-aligned video AI and its many dimensions and stakeholders. They identified key challenges that need to be addressed, these include

1. **Life-cycle control.** To be able to monitor the entire life-cycle correctness of a video AI system, so it becomes more accepted, more trusted, and more acceptable.
2. **Modular explainability.** To be able to explain not only the overall black-box video-AI systems, but to be able to do so at a modular level so that each specific system function is understood.
3. **Self-awareness.** Make video-AI systems aware of their ambiguity and (out of set) biases.

These projects have received funding from The European Union's research and innovation programme under GA No. 951911 (AI4Media), 951847 (ELISE), 101070617 (ELSA), 101070596 (euROBIN), 761758 (HumaneAI Net), 952215 (TAILOR) and 952070 (VISION).

4. **Unlearning.** How can we force a Video-AI system to forget, or unlearn, undesired properties?
5. **Conflicting requirements.** To be able to identify what makes video-AI acceptable to society given the often conflicting multi-disciplinary stakeholder requirements.

In response to these challenges, the session participants formulated a clear recommendation and action that is needed to make a step towards more human-aligned video AI. First of all, there is a need for a multi-disciplinary research agenda, that supplements technical-AI know how with knowledge from ethics, social and legal scholars, as well as use-case specific domain knowledge. Inspiration from other communities can be very valuable here to learn from their lessons, for example those that have studied explainability for a long time, or the human-computer interaction community that have always advocated for involvement of technology-users from the start. It will be no free lunch, however, Video-AI will require content- and use-specific adaptation.

A natural second recommendation is the call for action to work together across disciplines and expertise areas. Thanks to democratisation of AI-data, -software and -compute, as well as an encouraging European focus on human-centred AI, the moment is now. We need to mobilise the communities, organise more cross-cutting community events like today's workshop, and raise awareness in various research communities. A position paper also came out as a recommended action to get the process started.

Lastly, we discussed the acceptance of discomfort caused by questions that cannot be made computable, or have no objective function to optimise. Research on 'good old AI' has delivered many valuable lessons, still most relevant today. Try to incorporate these lessons from the start, at scale, not as a fix at the end. It was concluded that a joint proposal application for an EU project would be a much-needed next step to leverage the momentum and get the multidisciplinary collaboration with relevant experts and stakeholders started.

**Session 13: Trustworthiness in Robotics: at home, at work, and in the city**

How to ensure trust of humans to a robot? What are the hindrances to build that trust, at home, at work, in the city? How to ensure trust with people suffering from cognitive, physical or sensorial deficiencies? The session will take inspiration from the guidelines of the High-Level Expert Group on AI which recommends that AI systems should meet a set of requirements to be deemed trustworthy including:

- Allowing humans to make informed decisions,
- Implementing technical robustness and safety
- Guaranteeing privacy and data governance
- Ensuring transparency, explanations of decisions made
- Enabling diversity, non-discrimination, fairness
- Respecting societal and environmental well-being
- Putting in place mechanism to ensure accountability for systems and their outcomes

These projects have received funding from The European Union's research and innovation programme under GA No. 951911 (AI4Media), 951847 (ELISE), 101070617 (ELSA), 101070596 (euROBIN), 761758 (HumaneAI Net), 952215 (TAILOR) and 952070 (VISION).

Within the insightful breakout session on Trustworthiness in Robotics, the participants commenced with a nuanced discussion on terminology. Recognizing the potential pitfalls of using the term "trust" in the context of human-robot interactions, they pivoted towards the concept of "confidence" as a more fitting and nuanced descriptor. The concern arose from the bilateral nature of trust, which could lead to issues like betrayal when solely relying on trust.

Delving into the complexities of confidence in robotics, especially within the realm of social robotics, revealed a multifaceted landscape. The intricacies spanned various aspects, including task performance, precision, accuracy, operator safety, social interactions, communication difficulties, embodiment necessities, and concerns regarding privacy and dignity. The extensive array of challenges necessitates careful consideration when instilling confidence in robotics.

A crucial aspect highlighted during the discussions was the cultural variability in human-robot interaction, prompting the need for adaptable solutions. The individuals involved, their profiles, and capacities emerged as pivotal factors, with co-design emerging as a potential strategy to tailor solutions based on user profiles.

It became evident that confidence in robotics is a complex phenomenon, and safety represents only the visible tip of the iceberg. Achieving this confidence involves a delicate interplay between software and hardware, challenging the efficacy of standards as a sole solution for ensuring robot safety. Consensus emerged among participants that an interdisciplinary approach, encompassing social sciences and technology developers, is paramount to navigating the intricacies of building confidence in robots.

Summarizing the key challenges, the participants acknowledged traditional AI challenges related to predictability, privacy, and cybersecurity. In the context of robotics, additional considerations such as operator safety, task performance, and the intricate dynamics of social interactions and embodiments were deemed essential to address. This holistic approach was unanimously recognized as indispensable for achieving the desired confidence in robotic systems.

The discussions extended to the diverse applications of robotics, spanning healthcare, assistive technologies, surgery, rehabilitation, transport, building, energy production, distribution, and agro technologies. The universal concern across sectors was evident whenever human-robot interaction occurred, emphasizing the need for confidence-building measures.

A notable concern raised during the breakout session pertained to the integration of multimodalities in robotic models and the potential consequences for human-robot interactions. The intricacies of this integration, particularly in the context of inspection and maintenance infrastructure, underscored the importance of addressing the challenges posed by evolving robotic technologies.

In essence, the breakout session on Trustworthiness in Robotics brought to light the multifaceted nature of confidence-building in robotic systems, emphasizing the necessity of interdisciplinary

These projects have received funding from The European Union's research and innovation programme under GA No. 951911 (AI4Media), 951847 (ELISE), 101070617 (ELSA), 101070596 (euROBIN), 761758 (HumaneAI Net), 952215 (TAILOR) and 952070 (VISION).

collaboration and tailored approaches to address the intricate challenges inherent in this evolving field.

**Session 14: Ethics in Games AI**

Games is an application domain of AI research that is often overlooked when discussing responsible AI. Yet, given the scale of the industry and the wide use of AI techniques for non-player characters, game/level co-creation, and even matchmaking, it is surprising that concerns related to AI Ethics have not yet been discussed. This panel aims to challenge this by discussing the unique challenges that appear in the games environment (E.g. need for believable characters) while also satisfying ethical values.

In our thought-provoking breakout session on ethics in AI games, the participants delved into the unique challenges posed by this dynamic intersection of technology and entertainment. Recognizing that AI games serve as a cornerstone for AI applications, the complexities arising from the gaming industry's rapid technological advancements and historical lower levels of regulation have exploded.

Key Challenges:

1. **Opacity in the Game Industry**
   The industry's opacity, influenced by rapid technological adoption and limited regulation, poses challenges in addressing emerging critical issues.
2. **Addictiveness and Safety Concerns**
   Successful games often exhibit addictive qualities, raising concerns about users disregarding safety precautions. Hidden costs in gaming may contribute to ethical concerns, especially with the deployment of large AI models impacting climate goals and sustainability.
3. **Responsibility Ambiguity**
   Large-scale deployment of AI by various companies blurs lines of responsibility in the industry, intensifying issues of opaqueness and secrecy.

Key Recommendations:

1. **Leveraging Reputation**
   Emphasizing reputation as a crucial factor, we proposed using it to drive conversations and enhance literacy about AI applications and game design patterns among end users.
2. **Centralized Oversight**
   Acknowledging the fragmented nature of the gaming industry, we recommended centralizing oversight on AI ethics through major publishers and platforms to ensure more effective management, particularly for smaller companies.
3. **Enforcing Carbon Costs**

These projects have received funding from The European Union's research and innovation programme under GA No. 951911 (AI4Media), 951847 (ELISE), 101070617 (ELSA), 101070596 (euROBIN), 761758 (HumaneAI Net), 952215 (TAILOR) and 952070 (VISION).

Recognizing the underexplored aspect of sustainability, we proposed enforcing a carbon cost on large AI models to encourage more eco-friendly practices within the gaming industry.

4. **Preserving Autonomy**
   The importance of preserving autonomy emerged as a recurring theme. Ethical applications and certifications were seen as catalysts for a new industry branch, emphasizing the need for smaller, distributed, and efficient models for sustainability.

Key Results:

1. **Relevance to AI Automation on the Labor Market**
   The discussions outlined key results and proposed a roadmap for the future, suggesting workshops on AI automation in the labor market.
2. **Game Agnostic Toxicity Detection Hackathons**
   Future endeavors could involve hackathons focusing on game-agnostic toxicity detection, aiming to foster innovation and awareness.
3. **Ethical Gaming Facilitation**
   Addressing ethical concerns in virtual sports and gaming, workshops could facilitate discussions and solutions for sustainable AI systems in the gaming industry.

In conclusion, the breakout session not only highlighted the intricate challenges in ethics within AI games but also provided a roadmap for future endeavors, emphasizing the pivotal role of reputation, centralized oversight, sustainability, and the preservation of autonomy in shaping the future of the gaming industry.

These projects have received funding from The European Union's research and innovation programme under GA No. 951911 (AI4Media), 951847 (ELISE), 101070617 (ELSA), 101070596 (euROBIN), 761758 (HumaneAI Net), 952215 (TAILOR) and 952070 (VISION).

# Summary and Conclusion

The high international interest that was expressed in response to the announcement of the 2nd cross-cutting Theme Development Workshop (TDW) on "Trusted AI: The Future of Creating Ethical and Responsible AI Systems" translated into excellent attendance of the event. One hundred twenty participants joined the TDW, ranging from a diverse set of backgrounds. The TDW, therefore, caught the attention of some of the most important actors in the field of Trusted AI and brought together representatives from key companies, supra-national institutions, and academia. The workshop thus successfully provided a platform for discussions between representatives from academia, industry and politics: Discussions that are key in unlocking the full potential of AI in Europe.

The Organising Committee would like to express its deep gratitude to all experts for their valuable input and contributions to this Theme Development Workshop! Their active participation in the workshop and engagement in the breakout session discussions paved the way for the excellent results presented in this report.

These projects have received funding from The European Union's research and innovation programme under GA No. 951911 (AI4Media), 951847 (ELISE), 101070617 (ELSA), 101070596 (euROBIN), 761758 (HumaneAI Net), 952215 (TAILOR) and 952070 (VISION).

# List of participants

(in alphabetical order)

| Name | Affiliation |
|---|---|
| Alin Albu-Schaeffer | DLR-German Aerospace Center, Technical University of Munich, Germany |
| Giuseppe Amato | CNR-ISTI, Italy |
| Argyro Amidi | National and Kapodistrian University of Athens, Greece |
| Antoine-Alexandre André | DG CNECT, Belgium |
| Plamen Angelov | Lancaster University, United Kingdom |
| Isabella Banks | University of Amsterdam, The Netherlands |
| Imre Bard | Radboud University, The Netherlands |
| Dante Barone | Federal University of Rio Grande do Sul, Brazil |
| Max Bartolo | Cohere Inc., Canada |
| Oscar Bastiaens | Breda University of Applied Sciences, The Netherlands |
| Duuk Baten | SURF, The Netherlands |
| Jenny Benois-Pineau | University of  Bordeaux, France |
| Ana Bernardos | Universidad Politécnica de Madrid, Spain |
| Tobias Blanke | University of Amsterdam, The Netherlands |
| Andrea Borghesi | University of Bologna, Italy |
| Barteld Braaksma | CBS, The Netherlands |
| Stefano Braghin | IBM Research, Ireland |
| Edward Brodie | THALES SIX GTS, France |
| Robert Brunner | B'IMPRESS, Germany |
| Alberto Bugarin-Diz | University of Santiago de Compostela, Italy |
| Syed Saqib Bukhari | ZF Friedrichshafen AG, Germany |
| Noemi Luna Carmeno | intellera consulting srl, Italy |
| Rémy Chaput | Laboratoire d'InfoRmatique en Image et Systèmes d'information (LIRIS, France |
| Mohamed Chetouani | Sorbonne Universite, ISIR-UPMC, CNRS, France |
| Manolis Chiou | Queen Mary University of London, United Kingdom |

These projects have received funding from The European Union's research and innovation programme under GA No. 951911 (AI4Media), 951847 (ELISE), 101070617 (ELSA), 101070596 (euROBIN), 761758 (HumaneAI Net), 952215 (TAILOR) and 952070 (VISION).

| | |
|---|---|
| Gabriele Ciravegna | Politecnico di Torino, Italy |
| Ana Corrêa | KU Leuven, Belgium |
| Rita Cucchiara | Università degli Studi di Modena e Reggio Emilia, Italy |
| Harmen de Weerd | University of Groningen, The Netherlands |
| Sabine Demey | Imec, The Netherlands |
| Natalia Diaz Rodriguez | University of Granada, Spain |
| Manuel Dietrich | Honda Research Institute Europe GmbH, germany |
| Lidia Dutkiewicz | KU Leuven, Belgium |
| Elizabeth El Haddad | Inria, France |
| Francesco Ferro | PAL Robotics, Spain |
| Anja Fessler | ZF Group, Germany |
| Bettina Finzel | University of Bamberg, germany |
| Jack Fitzsimons | University of Oxford, United Kingdom |
| Martina Flatscher | ZF Friedrichshafen AG, Germany |
| Guillaume Gadek | Airbus, Germany |
| Swen Gaudl | University of Gothenburg, Sweden |
| Ole Goethe | Correlate, Norway |
| Tamas Gyulai | Széchenyi István University, Hungary |
| Meeri Haataja | Saidot Ltd., Finland |
| Anisa Halimi | IBM, Ireland |
| Andreas Herzig | CNRS-IRIT, France |
| Sam Hill | German Research Center for Artificial Intelligence (DFKI Germany |
| Naoise Holohan | IBM Research, Ireland |
| Janina Hoppstädter | German Research Center for Artificial Intelligence (DFKI Germany |
| George Ioannidis | IN2 Digital Innovations GmbH, Germany |
| Norbert Jastroch | MET Communications, Germany |
| Pippa Jones | Hogeschool van Amsterdam, The Netherlands |
| Katarzyna Kaczmarek-Majer | Systems Research Institute Polish Academy of Sciences, Poland |

These projects have received funding from The European Union's research and innovation programme under GA No. 951911 (AI4Media), 951847 (ELISE), 101070617 (ELSA), 101070596 (euROBIN), 761758 (HumaneAI Net), 952215 (TAILOR) and 952070 (VISION).

| | |
|---|---|
| Dmitry Kangin | Lancaster University, United Kingdom |
| Dimosthenis Karatzas | Universitat Autónoma de Barcelona, Spain |
| Kyra Kiefer | German Research Center for Artificial Intelligence (DFKI Germany |
| Yiannis Kompartsiaris | Centre for Research and Technology Hellas (CERTH) |
| Arzam Kotriwala | ABB Corporate Research Center, Germany |
| Christos Koutlis | Centre for Research and Technology Hellas (CERTH) |
| Noémie Krack | Centre for IP and IT law (KU Leuven), Belgium |
| Danica Kragic Jensfelt | Royal Institute of Technology (KTH), Sweden |
| Agnieszka Lawrynowicz | Poznan University of Technology, Poland |
| Christophe Leroux | CEA, France |
| Paul Lukowicz | German Research Center for Artificial Intelligence (DFKI Germany |
| Maria Makrynioti | Centre for Research and Technology Hellas (CERTH) |
| Giulio Mecacci | Radboud University Nijmegen / TU Delft, The Netherlar |
| David Melhart | University of Malta, Malta |
| Leila Methnani | Umeå University, Sweden |
| Pascal Mettes | University of Amsterdam, The Netherlands |
| André Meyer-Vitali | German Research Center for Artificial Intelligence (DFKI Germany |
| Vasileios Mezaris | Centre for Research and Technology Hellas (CERTH), Greece |
| Krysztof Mieszkowski | Łukasiewicz-PIAP/ WUT CEZAMAT, Poland |
| Mikolaj Morzy | Poznan University of Technology, Poland |
| Wico Mulder | TNO, The Netherlands |
| Wolfgang Nejdl | Leibniz Universität Hannover, Germany |
| Edina Nemeth | National Research, Develpment and Innovation Office; A Coalition of Hungary, Hungary |
| Grigoris Nikolaou | University of West Attica, Greece |
| Andrea Orlandini | National Research Council of Italy, Italy |

These projects have received funding from The European Union's research and innovation programme under GA No. 951911 (AI4Media), 951847 (ELISE), 101070617 (ELSA), 101070596 (euROBIN), 761758 (HumaneAI Net), 952215 (TAILOR) and 952070 (VISION).

| | |
|---|---|
| Zacharoula Papamitsiou | SINTEF AS, Norway |
| Adrien Pavao | Paris-Sud University, France |
| Laurent Perrussel | IRIT Computer Science Research Institute of Toulouse, France |
| Maura Pintor | University of Cagliari, Italy |
| Chaido Porlou | Centre for Research and Technology Hellas (CERTH), Greece |
| Rui Prada | INESC-ID and Instituto Superior Técnico, Universidade d Lisboa, Portugal |
| Monika Reif | ZHAW Zurich University of Applied Sciences, Switzerlan |
| Isabela Rosal | KU Leuven, Belgium |
| Patrick Schramowski<br>Kinga Schumacher | German Research Center for Artificial Intelligence (DFKI Germany<br>German Research Center for Artificial Intelligence (DFKI Germany |
| Mark Schutera | ZF Group, Germany |
| Lionel Scremin | IRT SystemX, France |
| Johanna Seibt | Aarhus University, Denmark |
| Marija Slavkovik | University of Bergen, Norway |
| Philipp Slusallek | German Research Center for Artificial Intelligence (DFKI Germany |
| Cees Snoek | University of Amsterdam, The Netherlands |
| Eduardo Soares | IBM Research, Brazil |
| Christoforos Spartailis | Centre for Research and Technology Hellas (CERTH), Greece |
| Theoni Spathi | Centre for Research and Technology Hellas (CERTH), Greece |
| Marc Steen | TNO, The Netherlands |
| Olga Stepankova | CIIRC, CVUT v Praze, Czech Republic |
| Tamas Suli | ZF Friedrichshafen AG, Germany |
| Andreas Theodorou | Umeå University, Sweden |

These projects have received funding from The European Union's research and innovation programme under GA No. 951911 (AI4Media), 951847 (ELISE), 101070617 (ELSA), 101070596 (euROBIN), 761758 (HumaneAI Net), 952215 (TAILOR) and 952070 (VISION).

| | |
|---|---|
| Marlies Thönnissen | German Research Center for Artificial Intelligence (DFKI Germany |
| Carme Torras | CSIC-UPC, Spain |
| Loukas Triantafyllopolous | Hellenic Open University, Greece |
| Sonja Trifuljesko | University of Helsinki, Finland |
| Filareti Tsalakanidou | Centre for Research and Technology Hellas (CERTH), Greece |
| Despoina Vakkou | Centre for Research and Technology Hellas (CERTH), Greece |
| Jordi Vallerdú | ICREA - UAB, Catalonia |
| Bertine van Deyzen | SURF, The Netherlands |
| Roland Vogt | German Research Center for Artificial Intelligence (DFKI Germany |
| Xia Wang | German Research Center for Artificial Intelligence (DFKI Germany |
| Jianan Wie | Leibniz Institute for High Performance Microelectronics Germany |
| Alan Winfield | University of the West of England, United Kingdom |
| Tobias Wirth | German Research Center for Artificial Intelligence (DFKI Germany |
| Lorraine Wolter | CISPA Helmholtz Center for Information Security, Germa |
| Dimitros Zeginis | University of Macedonia, Greece |
| Willem Zuidema | University of Amsterdam (ILLC), The Netherlands |

In addition to this list, 4 participants of the TDW preferred not to be mentioned publicly by name and affiliation.

**The organisers would like to thank all participants for their valuable input and contributions to the Theme Development Workshop!**