

value-**ALIGN**ed socio-technical systems using  
large-language models (LLMs)

***WP1 - Doctoral Candidate Recruitment and Training***

**D1.3 Doctoral Seminars**

<b>Contractual Delivery Date</b>	30 Jun 2025	<b>Actual Delivery Date</b>	16.09.2025		
<b>Responsible Beneficiary</b>	TUM	<b>Contributing Beneficiary</b>	All		
<b>Security</b>	PU	<b>Nature</b>	OTHER		
<b>Version</b>	1	<b>Date</b>	16.09.2025	<b>Page Nb.</b>	



HORIZON-MSCA-2023-DN-01

Grant Number 101169473



Funded by the  
European Union

## Authors

Name	Organisation	Email
Auxane Boch	TUM	auxane.boch@tum.de

## Document History

Revision	Date	Modification	Contact Person

## Executive Summary

This deliverable reports on the first AlignAI doctoral seminar, held on 15 September 2025. The seminar gathered all doctoral candidates, supervisors, and consortium partners for a structured half-day exchange. The candidates presented their initial research plans, followed by feedback and discussion.

The presentations reflected the diversity of research across the network, covering equity and vulnerable populations in the EU, the global Brussels effect of the EU AI Act, awareness and meta-reasoning in large language models, and the role of Key Enabling Methodologies in aligning AI with human values. Discussions emphasised methodological reflexivity, the careful framing of key concepts such as “vulnerable groups” and “awareness,” and the challenges of bridging legal, ethical, philosophical, and design perspectives.

The event demonstrated the coherence of the doctoral cohort’s work with AlignAI’s overall objectives and confirmed the strength of the project’s interdisciplinary training framework.

## 1. Introduction

Doctoral seminars are a central element of WP1, providing a platform for candidates to present their work, receive feedback, and connect their projects to the consortium’s shared objectives. They serve both as training opportunities and as mechanisms to ensure alignment of individual projects with the overarching goal of developing value-aligned socio-technical systems.

The first seminar was conducted online and structured around four doctoral presentations, each followed by a discussion, creating an environment for constructive critique, knowledge exchange, and the identification of synergies across research domains.

## 2. Seminar Proceedings

### DC1 – Julia Li

*A Multidimensional Conception of Equity for AI and LLMs in the EU*

*Abstract:* Although the question of value-alignment has long existed in the public imagination as well as academic literature (Wiener, 1960), the world is being faced with a number of unprecedented issues which intersect with AI ethics. Moreover, technological innovations like AI have the capacity to highlight existing social inequities and inequalities on various levels (Valenduc, 2018; C. Wang et al., 2025; Weiss & Eikemo, 2017). There is the risk that advancements in AI will either leave the most vulnerable groups behind or make existing injustices worse (Capraro et al., 2024; Ong et al., 2024). At the forefront are large-scale societal shifts related to factors such as climate change, urbanisation, migration and digitalisation, which are part of the backdrop of AI development (Pavaloiu, 2016). Areas of scholarship relevant to AI, such as the digital determinants of health (Chidambaram et al., 2024), digital equity (Gottschalk & Weise, 2023) and human rights discourse (Hogan & Lasek-Markey, 2024), are active in addressing the impacts of AI on vulnerable groups. A unified approach to ethical AI, which is grounded in a holistic view of these societal challenges and how they affect populations, is necessary to proceed with truly ethical and responsible AI development. This doctoral thesis aims to address various types of inequity relevant to AI development in the EU with a focus on LLM technologies. Rather than focusing on narrowly defined ideas of value alignment, the goal is to look at the overall impacts of AI on underrepresented and marginalised populations. The research will be grounded on a combination of perspectives including but not limited to; socio-medical perspectives on wellbeing such as the social determinants of health (Marmot & Wilkinson, 1991; Osmick & Wilson, 2020), conceptualizing harm and vulnerability in ways that are centred in theories of distributive justice which address the experiences of the most disadvantaged members of society (Arneson, 2008; Rawls, 1971), and relational ethics which look at bottom-up approaches that embed AI ethics in social structural influences (Branford & Herzog, 2025). The thesis will involve four chapters/papers, which include a combination of empirical, sociotechnical and normative papers.

The discussion following Julia's talk raised important questions. First, the suggestion was given to extend the scope of her systematic review to cover pre-2017 literature, highlighting that value debates existed before the current wave of AI technologies. Second, participants asked how Julia planned to remain reflexive in a rapidly evolving sector. Julia explained that her methodology was designed to be flexible, combining inductive and deductive

approaches and remaining open to revisions as the field develops. Intersectionality was highlighted as an essential consideration, with questions about how overlapping identities such as migration status or LGBTQ+ identity would be addressed. Other participants raised concerns about the risks of defining “vulnerable groups” too narrowly or problematically.

### **DC5 – Mohaned Bahr**

*Initial Doctoral Plan & Papers Presentation - The Legal Aspects and Regulatory Frameworks of LLMS.*

**Abstract:** This presentation covers the initial road map of comprehending and analysing the current legal frameworks of LLMs and the intersection between law, ethics and human rights in light of LLMs. The aim is to conduct a comparative study between the current global AI/LLMs regulations and address the question "why the EU AI Act could be the right model for future regulation" and "How the EU AI Act could benefit from other regulatory regimes". The outcomes of this work will lead to the provision of a better understanding of how to improve the EU regulation to further innovation, translate regulatory language into technical benchmarks to achieve "trustworthy AI", and enhance the AI user's legal protections.

The discussion that followed engaged with questions of scope and methodology. Participants asked whether the work would include comparisons between EU and non-EU regulatory regimes, and what global consequences might follow from differences. Another participant queried whether the case studies would be limited to Germany or extended across Europe, and Mohaned clarified that his scope would be EU-wide. The conversation confirmed the relevance of his work to both academic debates and policy design.

### **DC10 – Zeynep Kabadere**

*Awareness and Meta-Reasoning in LLMs*

**Abstract:** This presentation introduces the first part of my PhD project, which explores the role of awareness in enabling meta-reasoning in Large Language Models (LLMs). While current LLMs can generate impressive outputs, they often lack the ability to monitor, explain, and adjust their own reasoning. I argue that meta-reasoning - reasoning about one's own reasoning - requires a foundational level of awareness. In this project, awareness

is treated as a functional and measurable concept, distinct from consciousness, and is analysed through three heuristic dimensions: self-awareness, situational awareness, and social awareness. The research aims to clarify the conceptual foundation of meta-reasoning, establish awareness as an epistemic prerequisite, and develop a framework for integrating these ideas into LLMs. The ultimate goal is to contribute to the development of AI systems that are not only powerful but also transparent, trustworthy, and responsible.

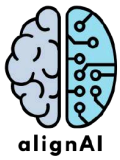
The following discussion was lively, with participants asking how awareness could be measured in machines. Zeynep described her exploration of benchmarking and evaluation frameworks, while acknowledging the limitations of current approaches. Another participant suggested that prompt engineering might already offer techniques to enhance forms of awareness. Next, a broader philosophical issue was raised, questioning whether debates on consciousness should be dismissed too quickly, since they can generate valuable insights for framing technical issues. Zeynep clarified that while she would delimit the scope of her work, she recognised the need to situate it within the broader debate. The discussion was pressed further, reflecting on whether existing LLMs such as ChatGPT already display self- or social awareness, leading to an exchange about proxies such as common-sense reasoning and the challenges of designing reliable evaluation methods.

### **DC11 – Sharvari Bondre**

#### *Key Enabling Methodologies for Aligning AI with Human Values*

*Abstract:* This presentation covers the basics of Key Enabling Methodologies (KEMs), specifically 1) Vision and Imagination, 2) Participation and Co-creation, and 3) Ethics and Responsibility. It also covers the role of historicism and critical theories (Feminist STS and Design Justice) in designing methodologies for AI, while providing an example that the presenter worked on and some other relevant case studies touching upon these concepts.

The discussion raised questions about the role of history in participatory design, with participants asking how past technological trajectories could inform current practices. Sharvari emphasised the importance of historicism in futuring, insisting that lessons from past technologies must inform visions of the future. The conversation also turned to the



importance of critical theories in challenging dominant values such as efficiency and productivity.

### **3. Key Outcomes**

This first seminar achieved its dual objectives: providing structured feedback to the doctoral candidates and reinforcing the interdisciplinary connections across the network. The event underscored the consortium's shared ambition to combine legal, ethical, social, and technical expertise in building value-aligned socio-technical systems. Future seminars will build on this foundation, ensuring continuity of training and knowledge integration across the doctoral projects.

---

DC 1 alignAI & doctoral project

# A multidimensional conception of equity for AI and LLMs in the EU

(working title)

Julia Li, Technical University of Munich, Supervisor: Prof. Dr. Christoph Lütge

**preface**

**Outline of  
doctoral plan**

# Introduction

This thesis aims to address various types of societal inequity relevant to AI development in the EU with a focus on LLM technologies.

Focusses on →

1. The extent of legal safeguards for human societal values in AI, including LLMs, in light of the EU AI Act
2. Perspectives on digital inclusion of vulnerable populations regarding LLM systems in education, mental health and online news consumption
3. The digital determinants of health and LLM technologies in the EU
4. A human rights governance framework for societal risks and impacts of AI and LLMs

End goal →

Build a multi-dimensional body of work on societal equity relevant to LLMs in the EU and beyond, highlighting salient population-level concerns and solutions

---

# Table of Contents

- 1** Chapter 1: Systematic lit review and legal analysis of EU societal values
- 2** Chapter 2: Key stakeholder interviews in the three use-cases
- 3** Chapter 3: Multidimensional assessment of societal values and digital determinants of health for LLMs in EU populations
- 4** Chapter 4: A human-rights based framework extending the concept of systemic risk for societal issues related to LLMs in the EU

**01**

**Chapter 1:  
Systematic lit review and legal  
analysis of EU societal values**

# European societal values on artificial intelligence in light of the EU AI Act: a systematic review and legal analysis

Julia Li, Mohaned Bahr and Prof. Dr. Christoph Lütge

*Main RQ: What are European societal values relating to value alignment and artificial intelligence and to what extent are they adequately protected according to the EU AI Act?*

An SLR of mixed-methods empirical literature using a qualitative approach. Identified EU societal values on AI will then be used as a supporting structure for a legal analysis of the EU AI Act for relevant legal issues and other sociotechnical elements.

### Inclusion criteria for systematic review

<b>Sample</b>	Populations and individuals from European Union member states.
<b>Phenomenon of Interest</b>	European societal values
<b>Design</b>	Published literature of empirical research designs including official surveys
<b>Evaluation</b>	Types of values, ethics, ethical norms, perspectives, attitudes, opinions, and/or standards
<b>Research type</b>	Literature containing qualitative, mixed-methods, quantitative results published from beginning of 2017- end of 2025

# European societal values on artificial intelligence in light of the EU AI Act: a systematic review and legal analysis

Julia Li, Mohaned Bahr and Prof. Dr. Christoph Lütge

## Background

- EU AI Act uses a risk-based system that draws heavily on universal human rights
  - 7 main values in Trustworthy AI; data governance, record-keeping, transparency, human-oversight, accuracy, robustness and cybersecurity. (Billah et al., 2025). More mentioned in Act and other documents as well
- There is a need to unite the normative with the empirical to see if there is congruence in between values inscribed in policy and law, and those expressed by people (Billah et al., 2025)
  - Will also include data from Eurobarometer surveys such as “Special Eurobarometer SP554 : Artificial Intelligence and the future of work” which surveyed ~26,415 respondents
- Open-ended language in mechanism in the Act's provisions raises questions
  - Is setting standards to assess if an AI system poses a risk to the fundamental rights a solution?
  - Or forming the notion of “fundamental values” to technical standards is possible to assess the extent of protection granted to these values (Smuha & Yeung, 2024)?

# European societal values on artificial intelligence in light of the EU AI Act: a systematic review and legal analysis

Julia Li, Mohaned Bahr and Prof. Dr. Christoph Lütge

## Analysis

### Systematic review of empirical literature

JL will extract themes and findings about EU societal values from studies involving EU populations that employ methods such as surveys, interviews, mixed-methods approaches, and reviews.

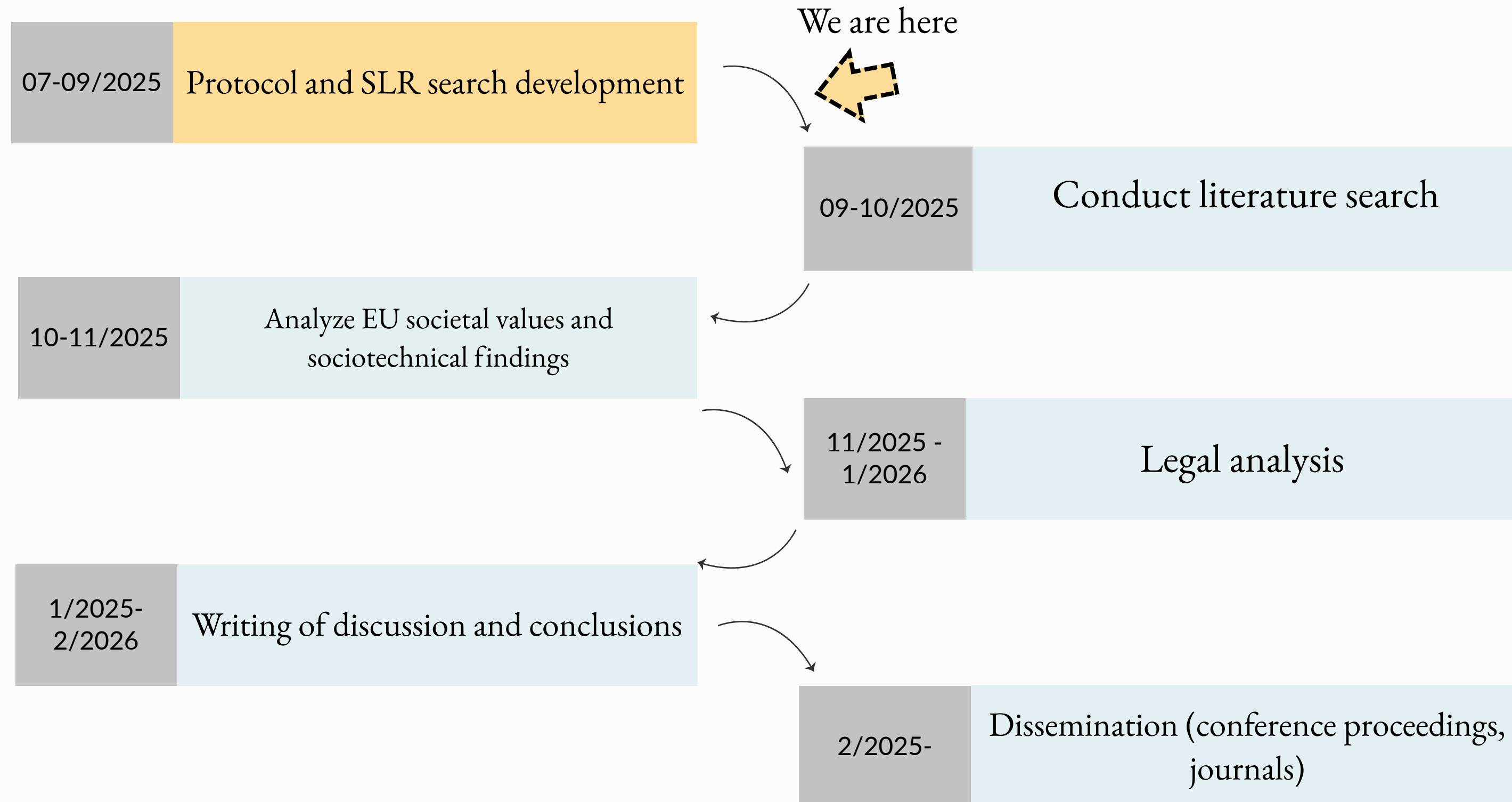
### Legal analysis

MB will identify and analyze the whereabouts of the identified values in the EU AI Act, extent of protection, mechanism provided in the AI Act to assess risks posed by AI systems to these values, and challenge whether these mechanisms are successful enough to create convergence between the legal, ethical, and social comprehension of values and the technical comprehension of the same values.

These will go into a discussion written together where evidence from the literature will be used to contextualized legal findings

# European societal values on artificial intelligence in light of the EU AI Act: a systematic review and legal analysis

Julia Li, Mohaned Bahr and Prof. Dr. Christoph Lütge



02

**Chapter 2:  
Key stakeholder interviews in  
the three use-cases**

## Digital equity and inclusion and LLM value-alignment: perspectives on AI and LLMs from key stakeholders in three EU use-cases

*What are the perspectives on challenges and opportunities of vulnerable groups regarding LLM digital equity in three use cases in education, mental health and journalism?*

The main goals are to:

1. Identify vulnerable groups in the three use cases and gain insights on digital equity which can aid future and current LLM practitioners in making decisions in favour of human value alignment and ethics by design.
1. Provide practical recommendations for the AI practitioners involved in the three use cases which can contribute to digital equity of future users.

# Digital equity and inclusion and LLM value-alignment: perspectives on AI and LLMs from key stakeholders in three EU use-cases

Concept of digital divide refers to gaps in internet usage, access, quality, social support, etc... between people and populations (Stiakakis et al., 2024; OECD, 2020)

Some priorities for digital equity from the EU

- Web accessibility
  - Digital skills
  - Linguistic barriers
  - Connectivity
  - Access
- 
- Socio-economic and demographic profiling variables including age, education, gender, occupation, type of community and geographic location linked to type of internet user/non-user in the EU (Gomes & Dias et al., 2025)

Questions will be focused on the **economic**, **social** and **political** factors that impact digital equity and inclusion in the three use cases.

## Digital equity and inclusion and LLM value-alignment: perspectives on AI and LLMs from key stakeholders in three EU use-cases

*RQ: What are the perspectives on challenges and opportunities of vulnerable groups regarding LLM digital equity in three use cases in education, mental health and journalism?*

### Potential stakeholders/stakeholder groups

**Education** → (~5) Students at TUM and TU/e

**Mental health** → (~5) Peer counsellors for LGBTQ+ youth, such as from LGBTQ+ Denmark

**Online news consumption** → (~5) Key stakeholders from Reporters sans frontières and /or Amnesty International in Switzerland

## Digital equity and inclusion and LLM value-alignment: perspectives on AI and LLMs from key stakeholders in three EU use-cases

Potential stakeholders/stakeholder groups and questions for open-ended discussions

What do you think?

**5 minute activity:** Post at least one sticky note describing your thoughts on the question on Canva

**Join this live session:**

[https://www.canva.com/design/DAGxQpYU6Wg/sk3D2BGG2d-Bgwb4-8gyOQ/edit?utm\\_content=DAGxQpYU6Wg&utm\\_campaign=designshare&utm\\_medium=link2&utm\\_source=sharebutton](https://www.canva.com/design/DAGxQpYU6Wg/sk3D2BGG2d-Bgwb4-8gyOQ/edit?utm_content=DAGxQpYU6Wg&utm_campaign=designshare&utm_medium=link2&utm_source=sharebutton)



03

**Chapter 3:  
Multidimensional assessment of societal  
values and digital determinants of health  
for LLMs in EU populations**

A large-scale multidimensional survey on the impacts of LLMs on the digital determinants  
of health in EU populations

Main research question:

*What are the impacts of LLMs on the digital determinants of health according to perspectives from EU populations?*

The secondary research question is:

*Are there and if so, what are the potential relationships between belonging to one or more vulnerable groups and differences in determinants of health and perspectives of impacts of LLMs?*

A large-scale multidimensional survey on the impacts of LLMs on the digital determinants of health in EU populations

What are the digital determinants of health?

- 1 person-specific determinants;
- 2 community determinants;
- 3 technology-related determinants;
- 4 policy determinants; and
- 5 political, economic, societal and cultural determinants.

From the World Health Organization report “Addressing health determinants in a digital age: project report” (2024) using SLR methods

	Security settings and features	4 (1)	90.62
Policy determinants	AI validation, transparency, explainability, accountability and ethics	5 (1)	91.43
	Regulatory mandate	4 (0)	90.62
	Digital public infrastructure	4 (1)	91.43
	Health/disability status	4 (1)	93.94
	Socioeconomic inequalities	4 (0)	90.62
	Open and transparent decision-making	4 (1)	93.75

Figure 2, cropped

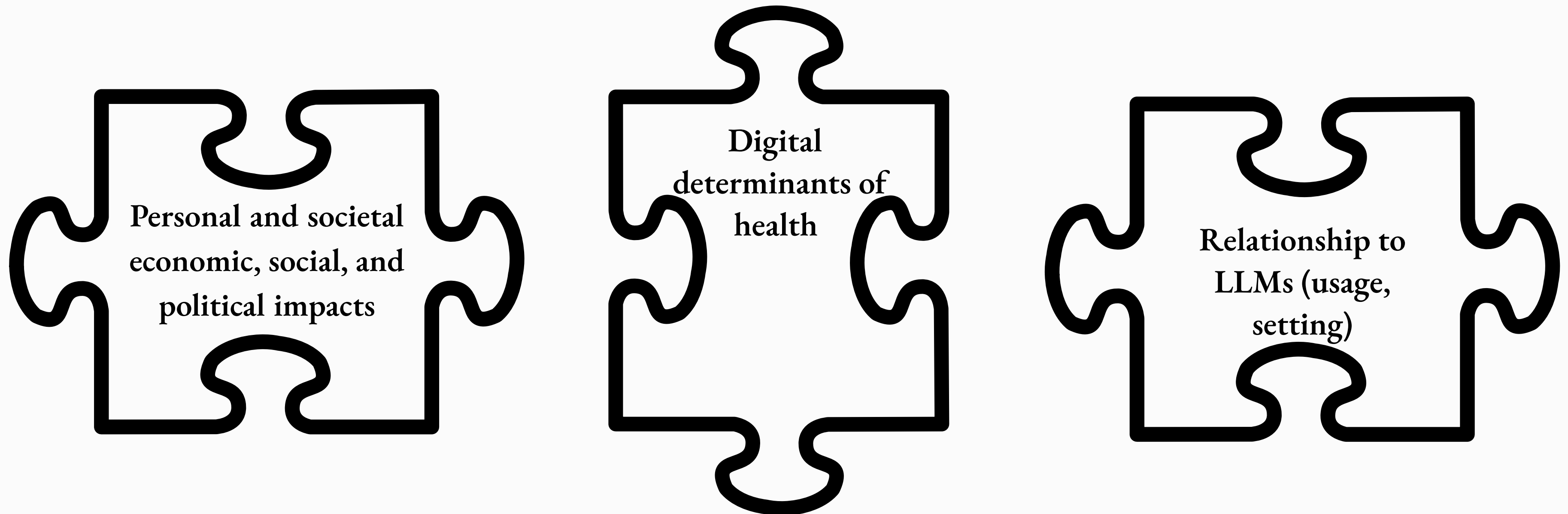
Excerpts from Table 2

**A large-scale multidimensional survey on the impacts of LLMs on the digital determinants of health in EU populations**

Objective of quantitative survey is to assess and link these elements.

Preliminary hypothesis:

Having access to the digital determinants of health lessens negative economic, social impacts that are related to the use of LLMs



**Chapter 4: A human-rights based  
framework extending the concept of  
systemic risk for societal issues  
related to LLMs in the EU**

## A human-rights based framework extending the concept of systemic risk for societal issues related to LLMs in the EU

- Legal regulations on various aspects of AI and technology carry legal and financial consequences
- **EU AI Act (Aug 2024)**
  - Draws on fundamental human rights from the **Charter of Fundamental Rights of the European Union** (entered into force 2009, Treaty of Lisbon) and **International Bill of Human Rights** (containing **Universal Declaration of Human Rights** (UN, 1948))
- Complemented by **General Data Protection Regulation (GDPR)** (April 2016)

# A human-rights based framework extending the concept of systemic risk for societal issues related to LLMs in the EU

Goal is to extend the concept of systemic risk through a risk and impact assessment framework based on human-rights and pertaining to societal-level concerns

There are pressing societal concerns regarding LLMs, which do not solely concern their abilities but also their use and deployment within society.

### Background

- HUDERIA, the human rights, democracy and the rule of law RIA for AIs, a non-legally binding document (Council of Europe Committee on Artificial Intelligence, 2024)
- Smuha's (2021) description of societal harms which threaten the overall aims of human society
- No requirement to conduct a systemic risk and impact assessment for GPAIs, including LLMs, that don't meet the high systemic risk threshold for high-impact abilities in Article 51 of the EU AI Act

# A human-rights based framework extending the concept of systemic risk for societal issues related to LLMs in the EU

### Environmental concerns

- Right to healthy environment (non-legally binding) → Human Rights Council resolution 48/13 (2021), United Nations General Assembly resolution A/RES/76/300 (2022)
- Google's 2024 report:
  - 31 billion liters of water in 2024 across data centers, overall water consumption increased by 28% from 2023-2024
  - Directing energy into cooling innovations to save water e.g. recycling waste water
- “Training GPT-3 in Microsoft's U.S. datacenters may consume a total of 5.4 million liters of water, including 700,000 liters of scope-1 onsite water consumption” (Li et al., 2025)

Should there be a greater emphasis on innovations and governance to improve env. sustainability of LLMs?

How do we balance the tradeoffs between LLM innovations for environmental stewardship and env. impact and should there be stronger thresholds and safeguards against env. damage?

# A human-rights based framework extending the concept of systemic risk for societal issues related to LLMs in the EU

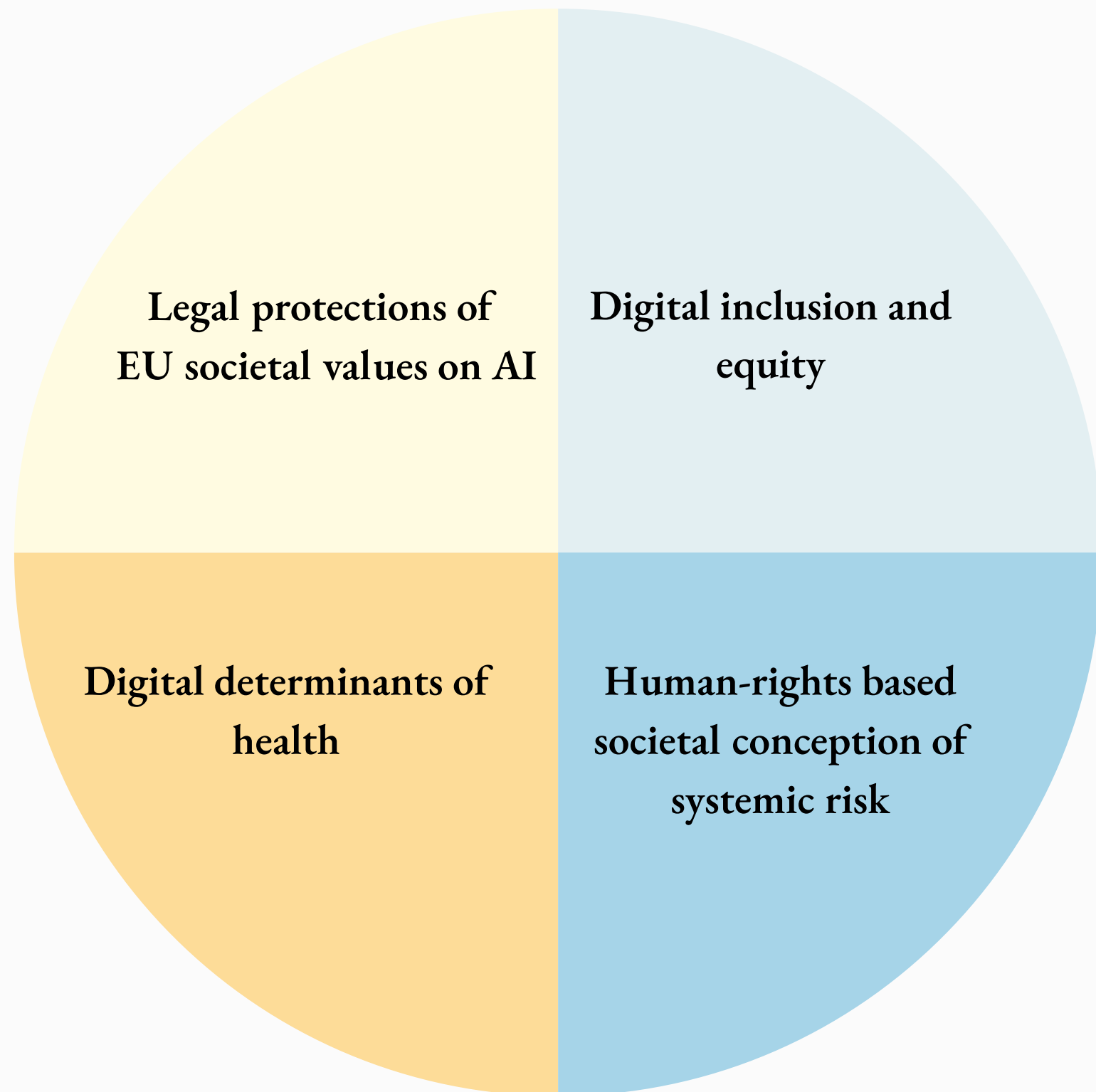
### Labour concerns

- The Universal Declaration of Human Rights (UNDHR), Article 23 → “the rights to work, free choice of employment, to just and favourable conditions of work and to protection against unemployment” (United Nations General Assembly, 1984).
- Concern for regional labour disparity in labor income distribution in EU (Minniti et al., 2025)
- Medium and high skill workers may experience greater negative impacts if they do cognitive and repetitive decision-making, information gathering, writing, etc.. (Tomlinson et al., 2025; Minniti et al., 2025)

How do we anticipate the effect of job loss and the impact it will have on our social systems and society?

Which innovations may affect the labour market the most and where?

# Conclusions



**My informal research questions**

**What population-level concerns are we missing when we talk about value-alignment in LLMs?**

**Who is going to be most affected and the most forgotten?**

**How do we move past the individual in value-alignment and look at the systemic effects of LLMs?**

# References for presentation

## Chapter 1

Billah, M. M., Hamjaya, H. S., Shiralizade, H., Singh, V., & Inam, R. (2025). Large Language Models' Trustworthiness in the Light of the EU AI Act—A Systematic Mapping Study. *Applied Sciences*, 15(14), Article 14. <https://doi.org/10.3390/app15147640>

Hagendorff, T. (2024). Mapping the Ethics of Generative AI: A Comprehensive Scoping Review. *Minds and Machines*, 34(4), 39. <https://doi.org/10.1007/s11023-024-09694-w>

Hogan, L., & Lasek-Markey, M. (2024). Towards a Human Rights-Based Approach to Ethical AI Governance in Europe. *Philosophies*, 9(6), 181. <https://doi.org/10.3390/philosophies9060181>

Lizarondo, L., Stern, C., Carrier, J., Godfrey, C., Rieger, K., Salmond, S., Apostolo, J., Kirkpatrick, P., & Loveday, H. (2020). 8. Mixed methods systematic reviews—JBI Manual for Evidence Synthesis—JBI Global Wiki. In *JBI Manual for Evidence Synthesis*. <https://synthesismanual.jbi.global>.

Luetge, C. (2017). The German Ethics Code for Automated and Connected Driving. *Philosophy & Technology*, 30(4), 547–558. <https://doi.org/10.1007/s13347-017-0284-0>

Scantamburlo, T., Cortés, A., Foffano, F., Barrué, C., Distefano, V., Pham, L., & Fabris, A. (2025). Artificial Intelligence Across Europe: A Study on Awareness, Attitude and Trust. *IEEE Transactions on Artificial Intelligence*, 6(2), 477–490. <https://doi.org/10.1109/TAI.2024.3461633>

# References for presentation

## Chapter 2

Digital divides. (2020). <https://www.oecd.org/en/topics/digital-divides.html>

Gomes, A., & Dias, J. G. (2025). Digital Divide in the European Union: A Typology of EU Citizens. *Social Indicators Research*, 176(1), 149–172. <https://doi.org/10.1007/s11205-024-03452-2>

Patrikakis (Eds.), *Next Generation Society. Technological and Legal Issues* (pp. 43–54). Springer. [https://doi.org/10.1007/978-3-642-11631-5\\_4](https://doi.org/10.1007/978-3-642-11631-5_4)

Stiakakis, E., Kariotellis, P., & Vlachopoulou, M. (2024). From the Digital Divide to Digital Inequality: A Secondary Research in the European Union. In A. B. Sideridis & C. Z.

Wang, C., Boerman, S. C., Kroon, A. C., Möller, J., & H de Vreese, C. (2025). The artificial intelligence divide: Who is the most vulnerable? *New Media & Society*, 27(7), 3867–3889. <https://doi.org/10.1177/14614448241232345>

# References for presentation

## Chapter 3

Dobrovolska, O., & Kolomiets, S. (2024). The Impact of Digitalisation on Social Determinants of Public Health. *Health Economics and Management Review*, 5(3), 128–142.

Mayiwar, L., Asutay, E., Tinghög, G., Västfjäll, D., & Barrafreem, K. (2025). Determinants of digital well-being. *AI & SOCIETY*, 40(4), 3063–3073. <https://doi.org/10.1007/s00146-024-02071-2>

Petretto, D. R., Carrogu, G. P., Gaviano, L., Berti, R., Pinna, M., Petretto, A. D., & Pili, R. (2024). Digital determinants of health as a way to address multilevel complex causal model in the promotion of Digital health equity and the prevention of digital health inequities: A scoping review. *Journal of Public Health Research*, 13(1), 22799036231220352. <https://doi.org/10.1177/22799036231220352>

World Health Organization European Region. (n.d.). Addressing health determinants in a digital age: Project report. WHO/EURO:2024-10917-50689-76724 (PDF), Licence: CC BY-NC-SA 3.0 IGO. Retrieved September 13, 2025, from <https://www.who.int/europe/publications/i/item/WHO-EURO-2024>

# References for presentation

## Chapter 4

Google. (2024). Google Environmental Report. <https://www.gstatic.com/gumdrop/sustainability/google-2024-environmental-report.pdf>

Li, P., Yang, J., Islam, M. A., & Ren, S. (2025). Making AI Less “Thirsty.” *Commun. ACM*, 68(7), 54–61. <https://doi.org/10.1145/3724499>

Minniti, A., Prettner, K., & Venturini, F. (2025). AI innovation and the labor share in European regions. *European Economic Review*, 177, 105043.

Smuha, N. A. (2021). Beyond the Individual: Governing AI’s Societal Harm (SSRN Scholarly Paper No. 3941956). Social Science Research Network. [Smuha, N. A. \(2021\). Beyond the individual: Governing AI’s societal harm.](#)

Tomlinson, K., Jaffe, S., Wang, W., Counts, S., & Suri, S. (2025). Working with AI: Measuring the Occupational Implications of Generative AI (No. arXiv:2507.07935). arXiv.

# Align-AI DN –MSCA

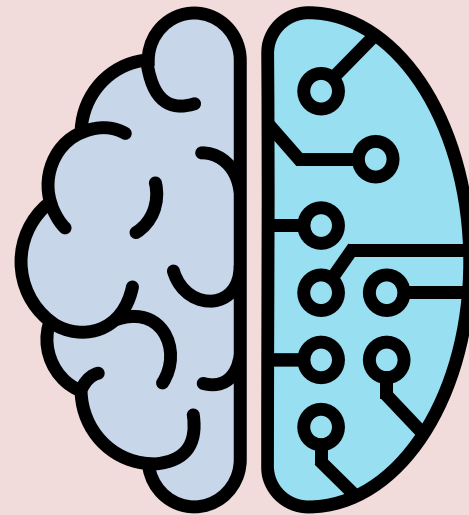
Doctoral Seminar

(Legal and Ethical Aspects of aligned LLMs)

DC.5

Mohaned Bahr | Technical University of  
Munich

Supervisor: Prof. Urs Gasser



**alignAI**

# Doctoral Plan

- DC.5's Role in The AlignAI Project
- Deliverables
- Initial Road Map of the Doctoral Plan
  - Papers
  - Collaborations
  - Others



**alignAI**

Aligning LLM Technologies with Societal Values

# My Role as DC.5



FOCUS: LAW, ETHICS &  
BUSINESS MODELS FOR  
LLMS



WORK PACKAGE 5:  
ENABLING ENVIRONMENT



TASKS: LEGAL FRAMEWORKS,  
BUSINESS MODELS, OPEN-  
ACCESS REPOSITORY

# Objectives & Deliverables



Assess current legal/regulatory frameworks



Propose new mechanisms for user rights protection



Develop ethical business models for LLMs



Create guidelines for SMEs/startups



Draft model terms & conditions


# DC.5's Initial Doctoral Plan

- Legal analysis of EU AI Act, GDPR & other laws = ensure LLMs alignment
    - Paper “ The replica of the EU AI Act”
      - Analyzing AI Act – distilling its regulatory mechanism & potential replicability in other jurisdictions (Brussels Effect)
      - Highlight the AI Act’s role as a rule-setter & experimentality enabler
      - Why EU > US or Chinese model
        - Exterritorial impact
        - Rewarded regulatory benefits
- Contribution: identify areas of improvement – First academic output linked to T5.2 (deliverables) - Regulatory blueprint for other Jurisdictions and companies (Start-ups)

# Deliverables

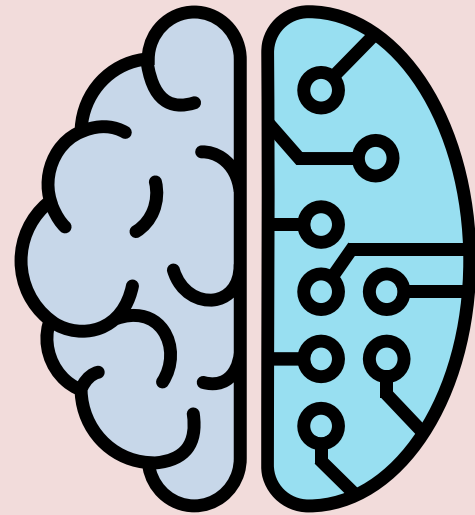
- T5.2 – Legal mechanisms for user rights protection
- T5.3 – Ethical business models for LLMs
- T5.4 – Open-access repository (tools, contracts, guidelines)

# Collaborations & Secondments

- Collaborations with DC1 (Ethics), DC10 (Philosophy), DC11 (Interaction Design)
  - Academic paper on Ethics and Law in the domain of LLMs
    - Identify new values, preferences, and ethical aspects
    - Analyze the text of the EU AI Act and other legal protections awarded to these newly identified human values, preferences, and ethical aspects.
  - Collaborating with TUM ForTe - Start-up consulting Section | TUM
    - life experience of start-ups and practical questions  Legal online repository.

The End

Thank you for listening



**alignAI**

# KEMs for Aligning LLMs with Societal Values

[DC 11] Sharvari Bondre supervised by Jesse Benjamin  
and Stephan Wensveen at TU Eindhoven

Sept 15, 2025

# Agenda

1. Key Enabling Methodologies
  - a. Vision and Imagination
  - b. Participation and Co-creation
  - c. Ethics and Responsibility
2. Participation and Power - State of the Art
3. Critical Theories

# **Key Enabling Methodologies**

# What are KETs



- Additive manufacturing
- Autonomous systems
- Sensor technology
  - Industry 4.0
  - Robotics



- Biomaterials
- 3D printing and design
- Chemicals, polymers, metals, glass
- Rapid prototyping



- Neurotechnology
- Bioengineering
  - AI in biology
- Bioelectronics
- Medical engineering



- Integrated circuit design
- Quantum computing
- IoT sensors and tokens
  - High-performance computing

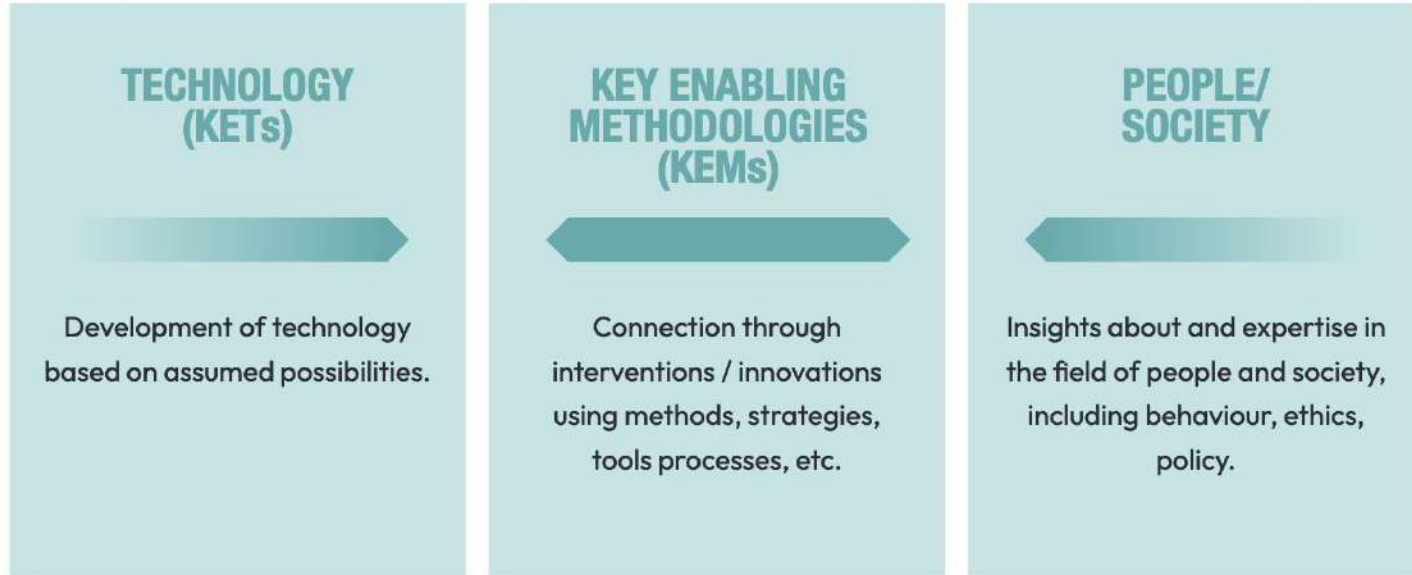


- Deep Learning
- Quantum AI
  - Robotics
- Autonomous systems
  - AI-as-a-service



- Standards (5G, SigFox...)
- Network architectures
  - Cryptography
- IoT networks & protocols
  - Distributed ledgers

# What are KEMs



*Figure 1: KEMs integrate knowledge of people and society with technological opportunities.*

# 11 KEM Categories

- Vision and Imagination
- Participation and Co-creation
- Behaviour and Empowerment
- Experimental Environments
- Value Creation and Scaling
- Institutional Change
- System Change
- Monitoring and Impact Measurement
- Ethics and Responsibility
- Meaning and Awareness
- Data for Exploration and Substantiation



# Mind Map

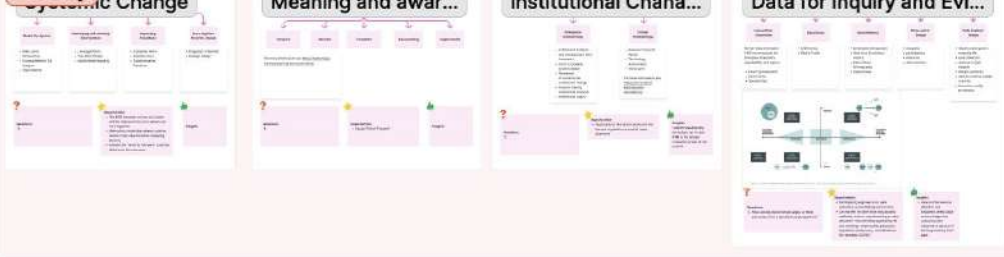
## Primary



## Secondary



## Tertiary



# 11 KEM Categories

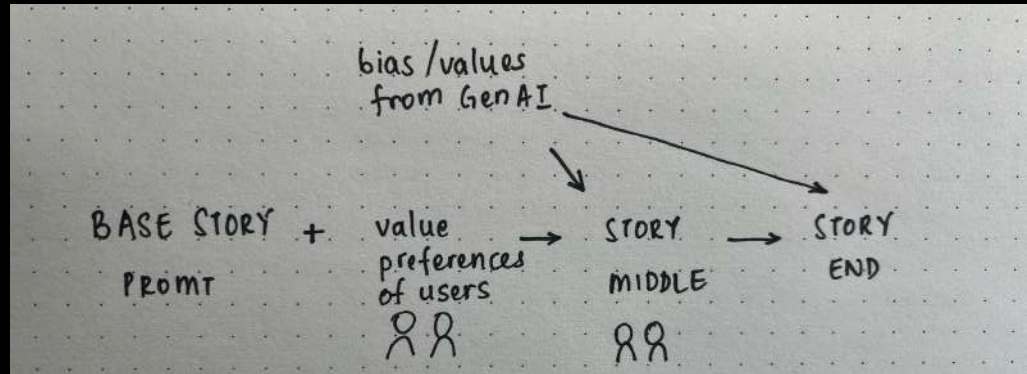
- Vision and Imagination
- Participation and Co-creation
- Behaviour and Empowerment
- Experimental Environments
- Value Creation and Scaling
- Institutional Change
- System Change
- Monitoring and Impact Measurement
- Ethics and Responsibility
- Meaning and Awareness
- Data for Exploration and Substantiation



# Vision and Imagination

- How can Vision and imagination be used for eliciting human values for the future?
- How can we use GenAI for imagining new futures?





# Interactive Story Engine

## Story Beginning

The large apple orchard had been in Samira's family for three generations. With aging parents in a nearby assisted living residence, now it was Samira's turn to try to continue the family business -- necessary to pay tuition for Samira's two children at college, and to save for medical expenses for those parents. However, there were new challenges. Samira was having trouble finding apple-pickers, and the new national pressures on undocumented workers seemed to be driving people away from Samira's land, which was clearly visible from nearby roads. Some other apple-growers had begun to replace human workers with robotic apple-pickers. They said that they could run the machines day and night, and that they could sometimes pick 10,000 apples per hour. But Samira enjoyed interacting with the apple-pickers, and liked having their kids around. Sometimes Samira would get treats for the children, and tried to show up for birthday parties. Samira was also concerned about what would happen to the school district if there were fewer children, as well as the consequences on the nearby outlets. A massive switch to robots could damage the community.

## Value Preference

Community Welfare

Financial Security



Current: 3/5

Balanced Approach

## Middle Story

Generate Middle

Click "Generate Middle" to create the story continuation, or write your own...

0/500 characters

## Story Ending

Generate End

- Complete the middle story section first to generate an ending.

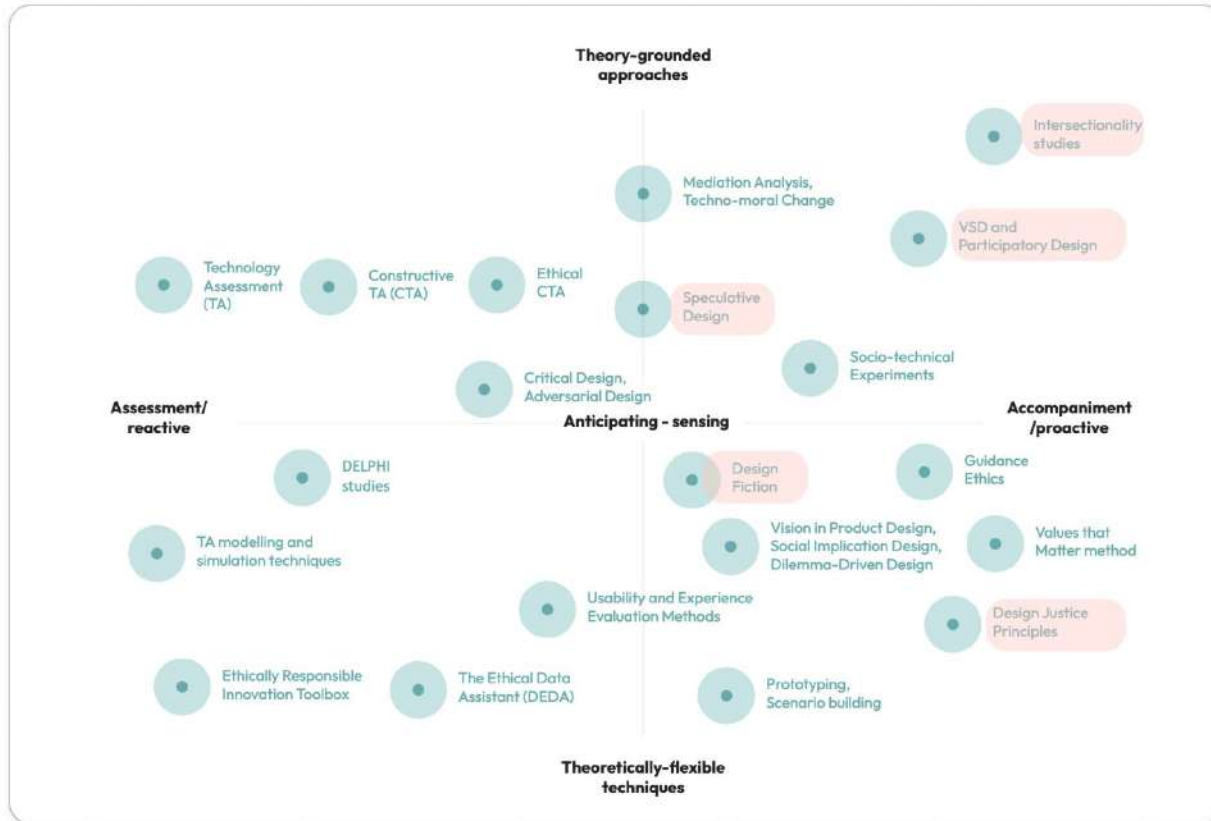
Your story ending will appear here after generation

# Ethics and Responsibility

- How do you make ethics positive and experiential?
- How to align the design, development and implementation of technology with societal and individual needs and values?



# Ethics and Responsibility Approaches



# Participation and Co-Creation

- How do you ensure responsibility in networked innovation processes?
- How do you resolve moral dilemmas and stakeholder value conflicts?
- When is co-creation/participation (not) useful for AI development?
- How do you secure long-term stakeholder value and commitment?





[Case Study]  
Redesigning Psychiatry

# Participation and Power

State of the art



	CONSULT	INCLUDE	COLLABORATE	OWN
<b>PARTICIPATION GOAL</b>	<b>Why is participation needed?</b>			
	To improve the user experience	To better align AI with stakeholders' preferences and values	To deliberate about system features	To shape the system's scope and purpose
	80/80	52/80	30/80	8/80
	<b>PARTICIPATION SCOPE</b>	<b>What is on the table?</b>		
User interface of the system		Underlying datasets (e.g., identification, curation, annotation)	Overall design of system (e.g., task specification, model features)	Whether and why the system should be built
80/80		8/80	8/80	4/80
<b>Who is involved?</b>				
Stakeholders recruited by the project team for discrete feedback	Stakeholders recruited by the project team for domain expertise	Stakeholders designated by the community collaborate in design	Stakeholders designated by community play central role across project lifecycle	
75/80	47/80	6/80	3/80	
<b>FORM OF PARTICIPATION</b>	<b>What form does stakeholder participation take?</b>			
	Giving input on design ideas via questionnaires and interviews	Group discussions with project team	Ongoing collaborative prototyping and decision-making	Reflexively deciding on the participatory approach
68/80	49/80	18/80	0/80	

Fig: Participatory AI Projects Mapped to Conceptual Framework

## Analysis of Level of Participation

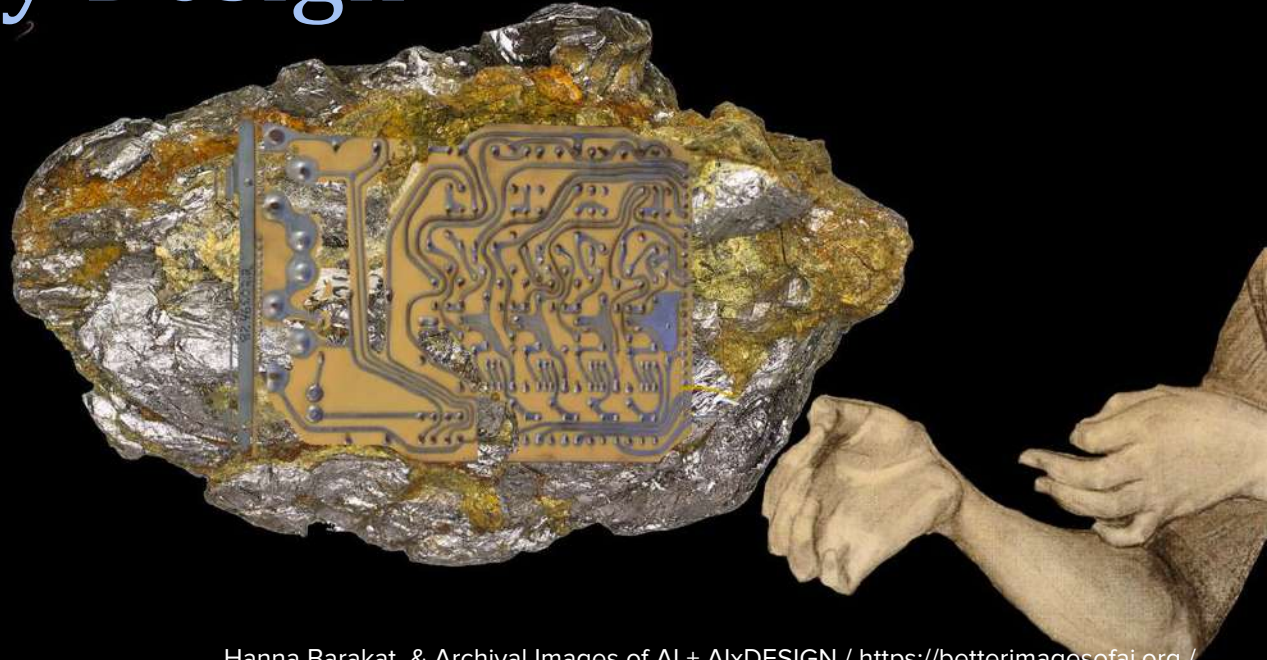


## Participation and Foundation Models

**How do we critically study  
participatory and futuring  
methods for AI?**

How do we **critically** study  
**participatory** and **futuring**  
methods for AI?

# Historicism in Participatory Design



# Drawing from Historicism in CSCW

drawing from STS and post-colonial theories, how historicism can play a role in

- 1) understanding historical trajectories
- 2) investigating histories of the designer, the user, the participant, the non-user, and their historical traces



Source: Wikimedia Commons. From Brockhaus, F. A., ed. *Brockhaus' Conversations-Lexikon*. Vol. 16, 13th ed. Leipzig, Germany, 1887, pp. 142. Public domain.



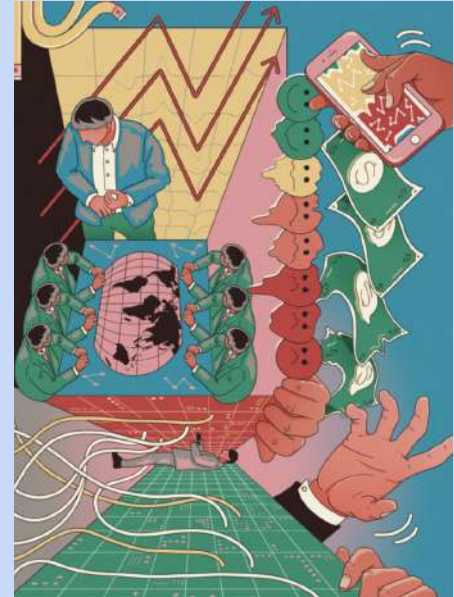
# Critical Theories

# Critical Theories



## Feminist Science and Technology Studies

Feminist Care Ethics  
Feminist Data Science



## Design Justice



## **Feminist Science and Technology Studies**

Feminist Care Ethics  
Feminist Data Science

- Pluralism and situated knowledge
- Studying and dismantling power structures
- Intersectionality
- Interconnectedness
- making labour visible
- overarching impact of Big Tech



## Design Justice

- Post-colonialism
- Design Sites and Participation
- Unequal Environmental Impacts
- Pedagogical Dimensions
- Matrix of domination and Intersectionality



[Case Study]  
Techniques of Use



## Towards Intersectional Feminist and Participatory ML: A Case Study in Supporting Femicide Counterdata Collection

Harini Suresh

hsuresh@mit.edu

Data + Feminism Lab, MIT, USA

Rajiv Movva

rmovva@mit.edu

Data + Feminism Lab, MIT, USA

Amelia Lee Dogan

dogan@mit.edu

Data + Feminism Lab, MIT, USA

Rahul Bhargava

r.bhargava@northeastern.edu

School of Journalism, Northeastern  
University, USA

Isadora Cruxên

i.cruxen@qmul.ac.uk

School of Business and Management,  
Queen Mary University of London  
USA

Ángeles Martínez Cuba

angelesm@mit.edu

Data + Feminism Lab, MIT, USA

Giulia Taurino

g.taurino@northeastern.edu

Khoury College of Computer Science,  
Northeastern University, USA

Wonyoung So

wso@mit.edu

Data + Feminism Lab, MIT, USA

Catherine D'Ignazio

digazio@mit.edu

Data + Feminism Lab, MIT, USA

### ABSTRACT

Data ethics and fairness have emerged as important areas of research in recent years. However, much work in this area focuses

### ACM Reference Format:

Harini Suresh, Rajiv Movva, Amelia Lee Dogan, Rahul Bhargava, Isadora Cruxên, Ángeles Martínez Cuba, Giulia Taurino, Wonyoung So, and Catherine D'Ignazio. 2022. Towards Intersectional Feminist and Participatory ML:

Good example of multi-stage involvement and ownership

**Thank you!**

# Awareness-Based Meta-Reasoning in LLM's: Towards Explainable and Responsible AI

Zeynep Kabadere




Phd Candidate for Philosophy of Artificial Intelligence  
Philosophy & Ethics Group at Eindhoven University of Technology.

[z.kabadere@tue.nl](mailto:z.kabadere@tue.nl)

# OBJECTIVE

LLMs can write, summarize, answer questions. They're already helping in healthcare, education, and more. But in important situations, we need more than correct answers.

We need systems that can:

-  Check their own reasoning
-  Explain their decisions
-  Adjust when something goes wrong

# OBJECTIVE

**Research Question:** How can meta-reasoning strategies in Large Language Models support awareness-based reasoning and help build more explainable, reliable, and responsible AI systems?

**Meta-reasoning** = Thinking about thinking

**Awareness** = Understanding one's own state and the situation



# What Is Meta-Reasoning?

---

Meta-reasoning = Reasoning about one's own reasoning

Helps systems to:

Ask: "Is this the best way to solve the problem?"

Notice: "Did I make a mistake in my thinking?"

Reflect: "Should I try a different approach?"

Improves not just answers, but the thinking process itself



# Awareness: A Key to Meta-Reasoning

---

Meta-reasoning needs awareness to work.

In my research, awareness means:

- Keeping track of what the system knows
- Knowing what it's doing
- Sensing the surrounding context

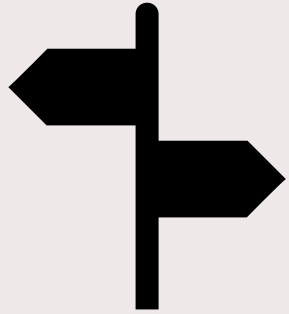
Three key types:

Self-awareness – understanding internal state and limits

Situational awareness – understanding the external context

Social awareness- understanding social context

Without awareness, meta-reasoning is not possible



# ROADMAP

# Step 1: Terminology – From Metacognition to Meta-reasoning

**Goal:** Clarify conceptual usage.

**Connection to thesis:** The project focuses on the higher-level regulation of reasoning processes, so using meta-reasoning instead of metacognition makes the terminology more accurate.

This choice ensures that the project's orientation is clear and consistent.

## Step 2: Awareness as a Prerequisite for Meta-reasoning

**Goal:** Place the claim that meta-reasoning requires a minimal level of awareness in artificial systems.

**Task:** Argue that awareness is an epistemic prerequisite

**Connection to thesis:** For meta-reasoning to function meaningfully and effectively, artificial systems must possess certain levels of awareness.

## Step 3: Methodological Distinction – Consciousness vs. Awareness

**Goal:** Avoid the hard problem → focus on awareness as a workable concept

**Connection to thesis:** Distinguish awareness from consciousness to avoid the hard problem

Consciousness → tied to phenomenal experience & qualia, resists empirical study

Awareness → treated as a functional and measurable concept

Defined as the agent's capacity to access, represent, and respond to internal and external information

## Step 4: Dimensions of Awareness

**Goal:** Make awareness workable for study

**Connection to thesis:** Three proposed dimensions – self, social, situational, and meta-reasoning awareness → heuristic, not final

**Why dimensions?** → To facilitate empirical research. Dimensions as a starting point, open to revision through literature

**Critical roles:** self-awareness & situational awareness in supporting meta-reasoning

# THANK YOU

**Zeynep Kabadere**

**Phd Candidate for Philosophy of Artificial Intelligence**

**Philosophy & Ethics Group at Eindhoven University of Technology.**

 [z.kabadere@tue.nl](mailto:z.kabadere@tue.nl)