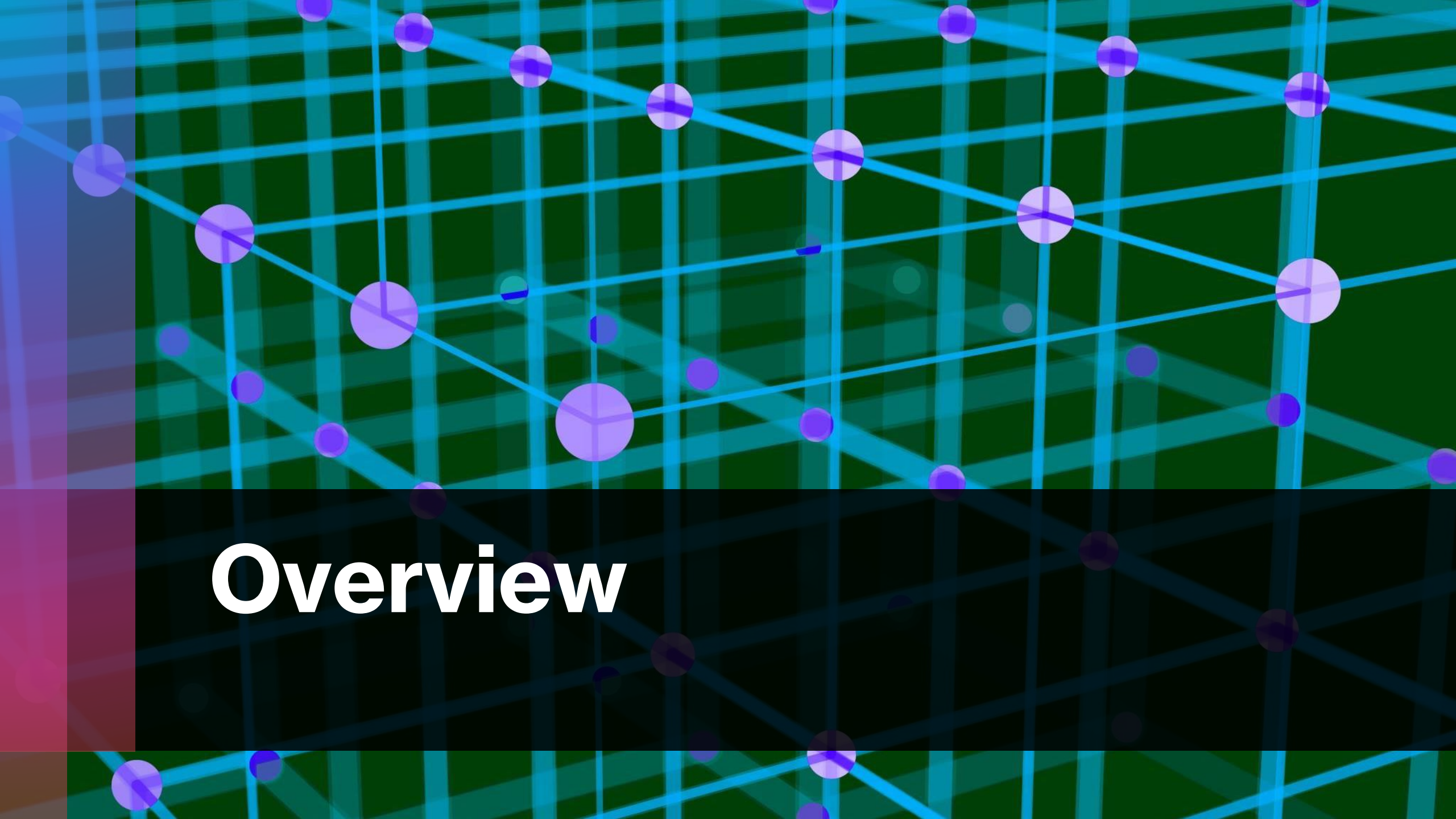
From AI Ethics Principles to Data Science Practice: a Reflection and a Gap Analysis based on Recent Frameworks and Practical Experience

By: Ilina Georgieva, Claudio Lazo, Tjerk Timan & Anne Fleur van Veenstra

Presentation by: Anaqi Amir

Structure

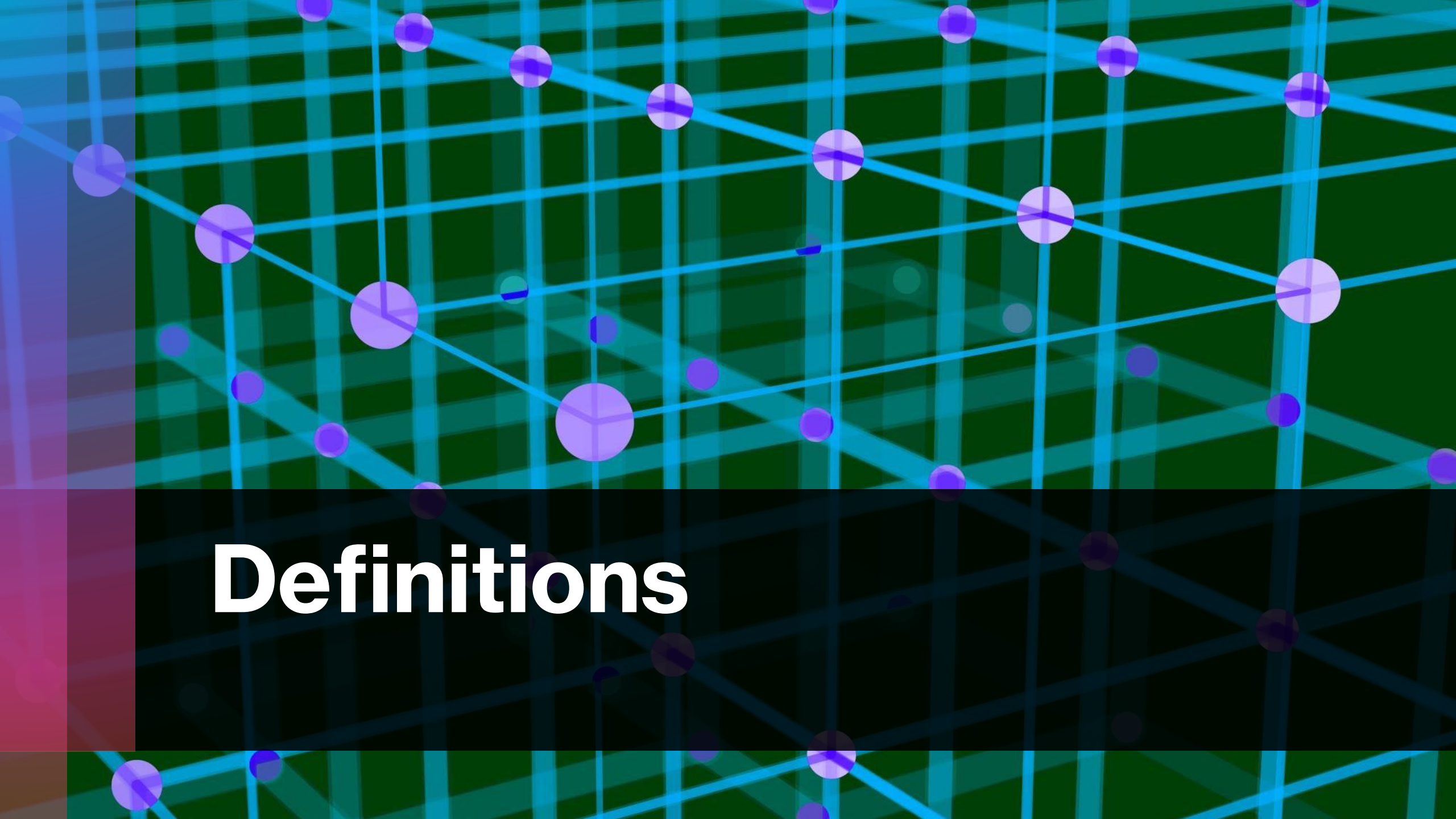
- Overview
- Definitions
- Context & Background
- Methodology
- Results
- Conclusions
- Critiques
- References



Overview

Aim

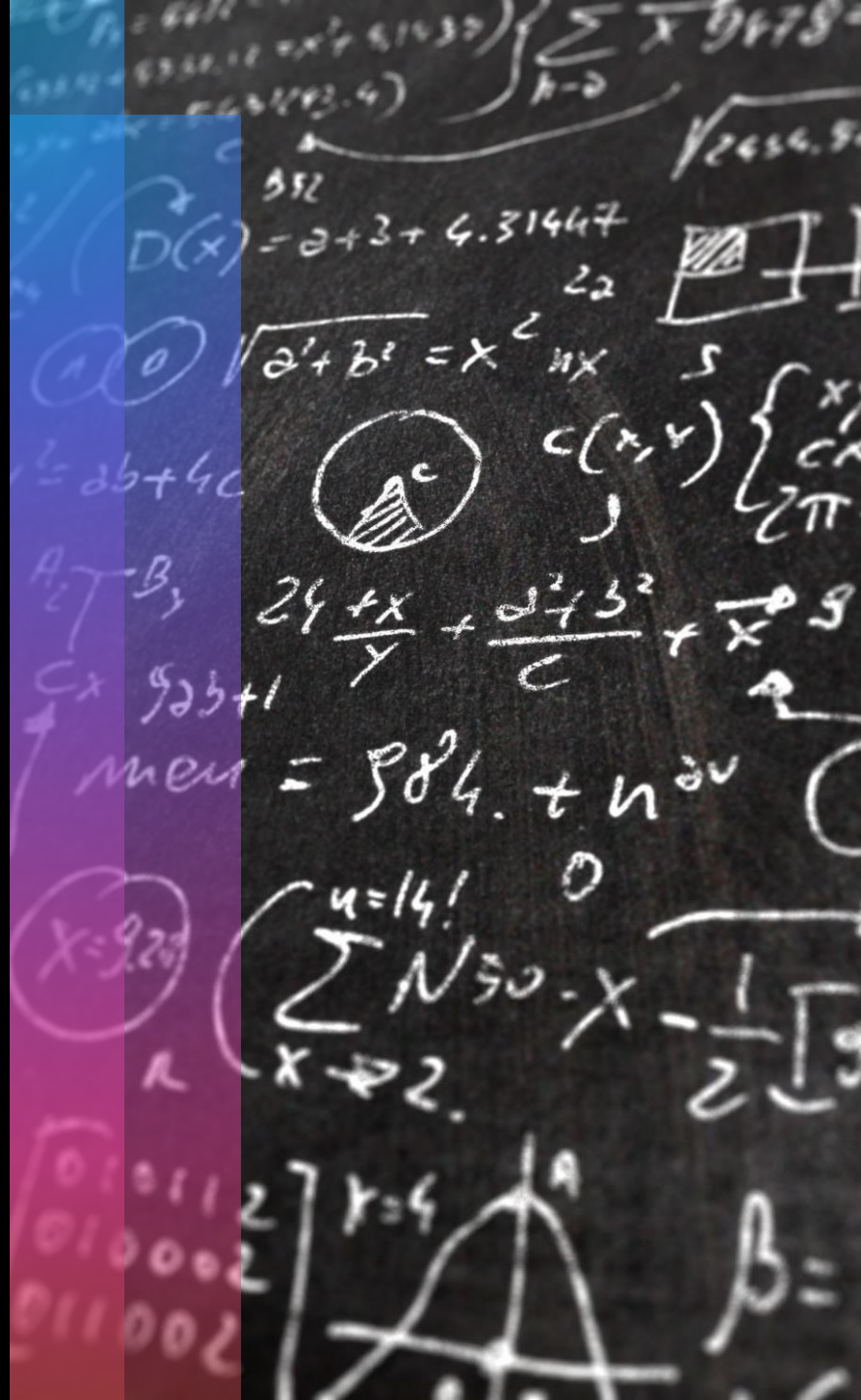
- To contribute to the translation of AI ethics to data science practice as part of the *third wave* of AI Ethic
- To widen the debate on AI governance



Definitions

What is AI Ethics?

- Definition of AI:
  - Machine-based systems that rely on algorithms that perform human-like cognitive functions with varying levels of autonomy
- AI Ethics
  - The design and implementation of human-centric AI

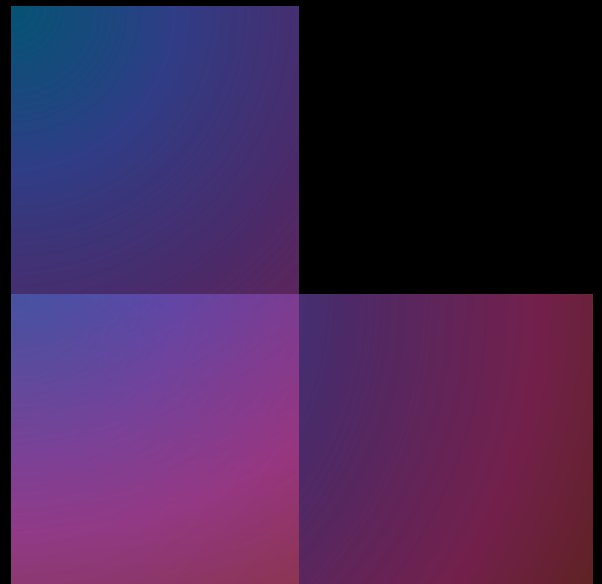


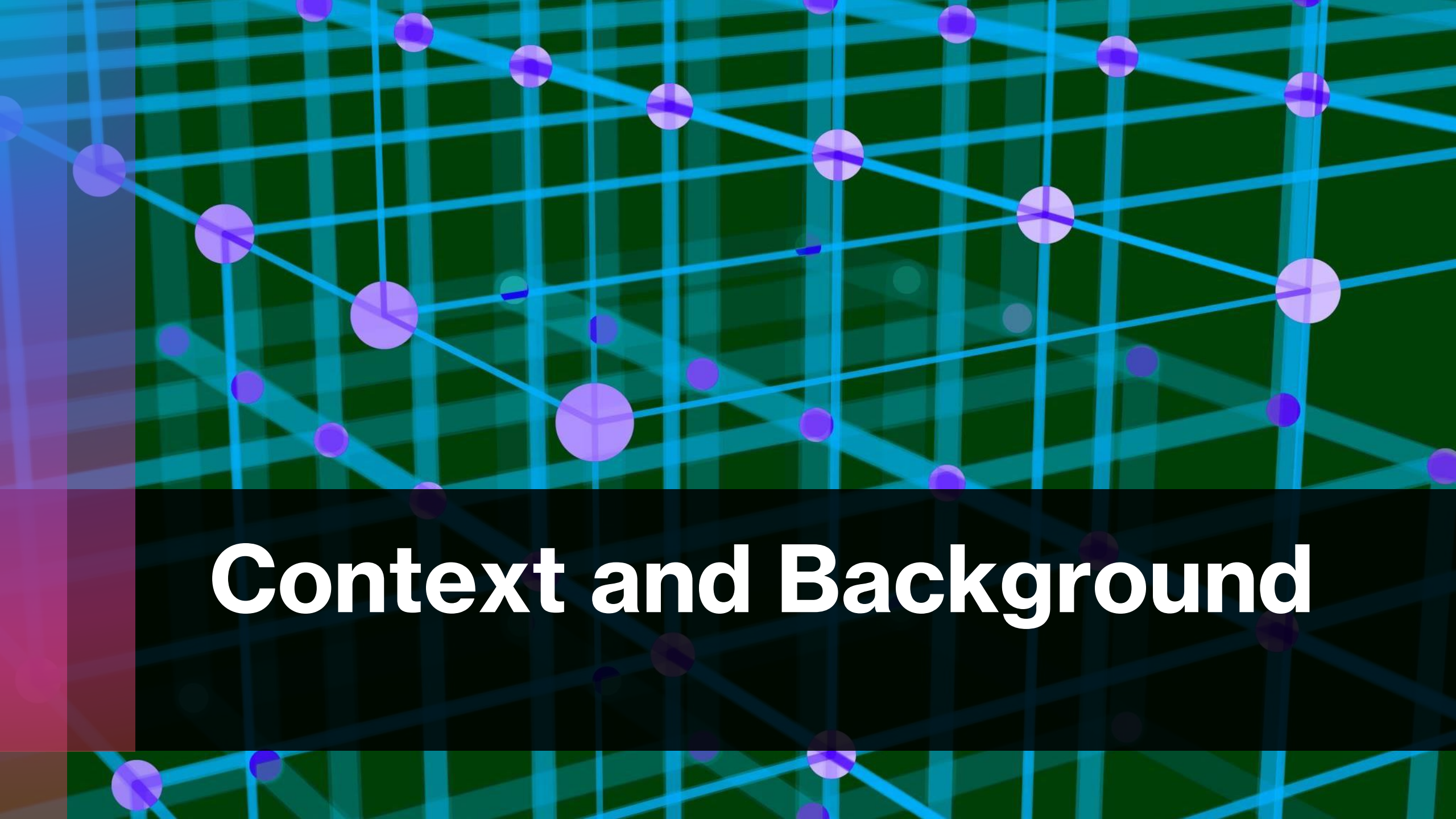
Waves of AI Ethics

First wave:
-Ethical frameworks for the development of socially beneficial or *human-centered* AI and its deployment

Second wave:
-Efforts to assess the potential normative consensus and convergence of these **frameworks**

Third wave:
-Turning AI principles into actionable data science practice and governance
-Applicability of AI ethics in real-world cases
-Highlights issues of standardization in specific sectors
-Questions of enforcement and regulatory control to AI governance





Context and Background

European Commission High-Level Expert Group (HLEG) Requirements for a Trustworthy AI

1) human agency and oversight
2) technical robustness and safety
3) privacy and data governance
4) transparency
5) diversity, non-discrimination, and fairness
6) societal and environmental well-being
7) accountability

Context

Nations have used the AI HLEG guidelines as mere "recommendations"

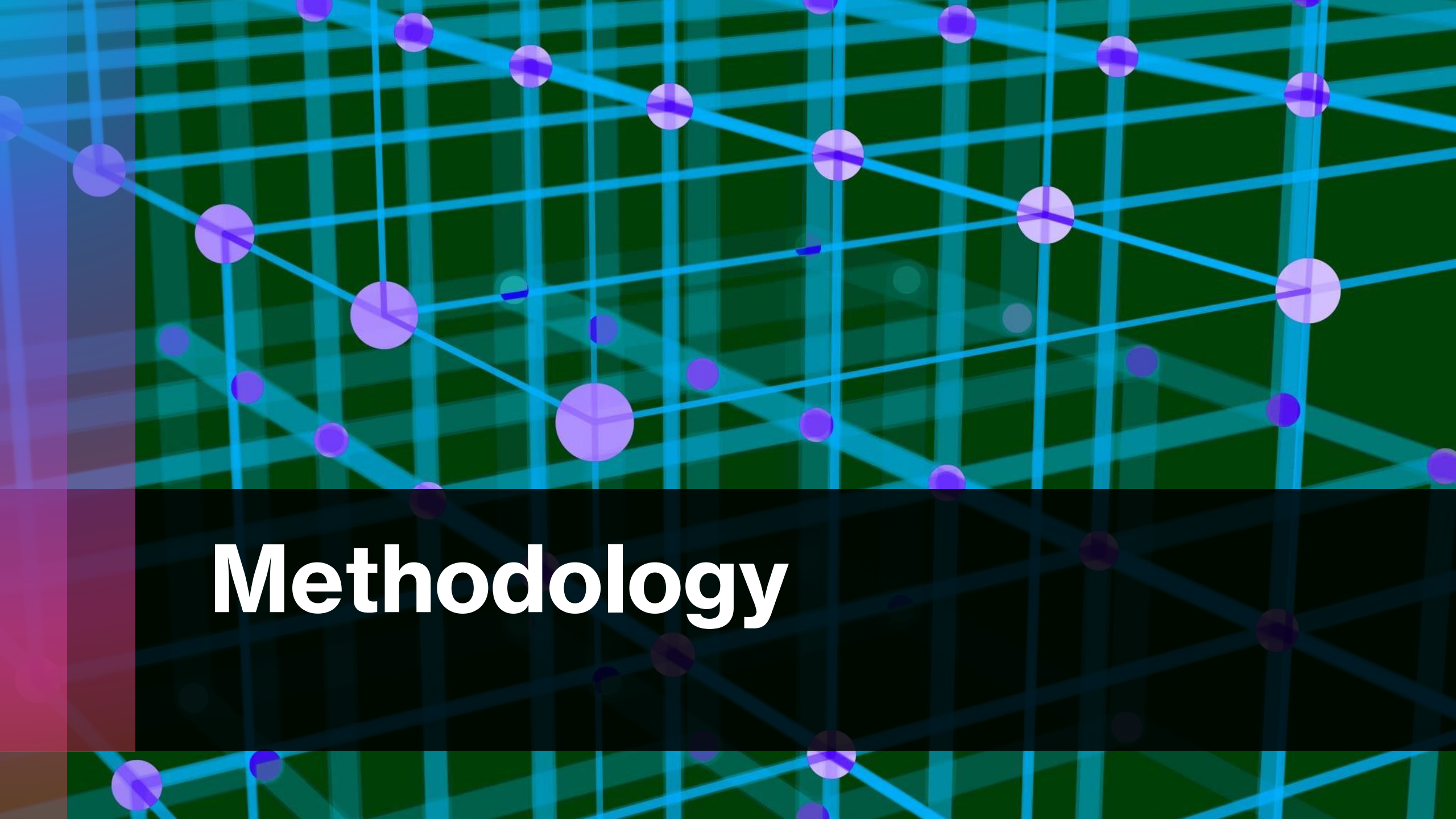
Problem:

Unsure how to implement these AI guidelines into their design and deployment

Unsure whose responsibility it is when it comes to AI ethics

Case Study

- SyRI (Systeem Risico Indicatie)
- Created by the Dutch government
- Ranks its citizens based on how likely they are to commit fraud


Methodology

Methodology

Step 1:

Define AI HLEG and AI lifecycle

Step 2:

Map AI HLEG to each stage of the AI-lifecycle using code

Step 3:

Create scores for each intersection

Results Table

	Goverance Structure	Safety Culture	Reliable Systems					Safety Culture	Trustworthy Certification
	Responsibility	Organization	Team					Organization	Industry
	AI Lifecycle Stage	Business / Use Case Development	Design	Testing & Test Data Procurement	Building	Testing	Deployment	Business / Use Case Development	*(Outside AI Lifecycle)*
AI HLEG	Human agency and oversight								
	Technical robustness and safety								
	Privacy and data governance								
	Transparency								
	Diversity, non-discrimination and fairness								
	Societal and environmental well-being								
	Accountability								

European Commission High-Level Expert Group (HLEG) Requirements for a Trustworthy AI

1) human agency and oversight
2) technical robustness and safety
3) privacy and data governance
4) transparency
5) diversity, non-discrimination, and fairness
6) societal and environmental well-being
7) accountability

AI Lifecycle and Explainability Definitions

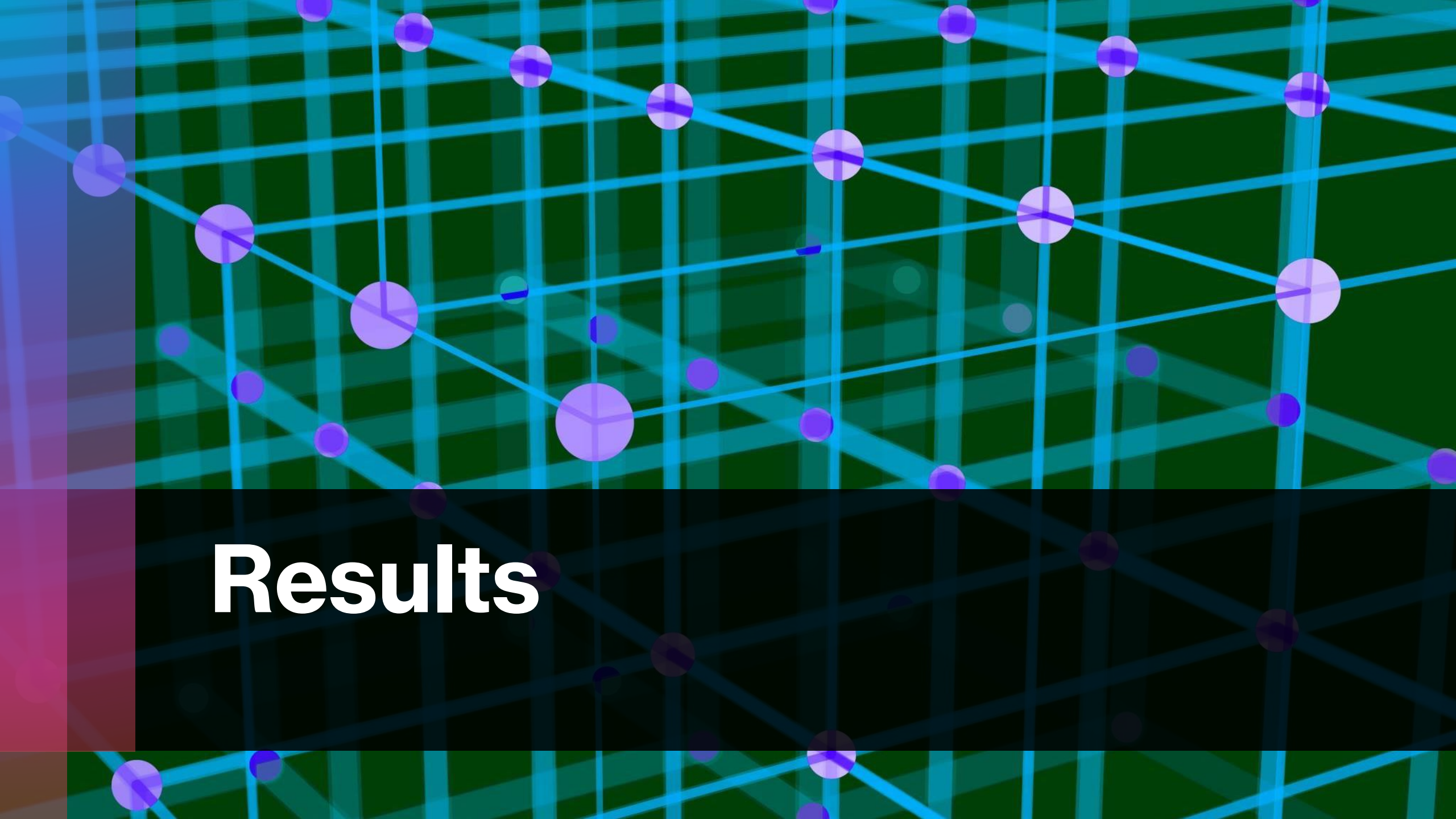
	AI lifecycle [45]	AI governance [58]
1	Business/use-case development	Safety culture (Organization)
2	Design	Reliable systems (Team)
3	Training & test data procurement	
4	Building	
5	Testing	
6	Deployment	
7	Monitoring	Safety culture (Organization)
8	Outside of the AI lifecycle	Trustworthy certification (Industry)

Section	Coding	Classification
"Explainability concerns the ability to explain both the technical processes of an AI system and the related human decisions (e.g. application areas of a system)"	(Not coded)	
"Technical explainability requires that the decisions made by an AI system can be understood and traced by human beings. Moreover, trade-offs might have to be made between enhancing a system's explainability (which may reduce its accuracy) or increasing its accuracy (at the cost of explainability)"	Technical explainability	Deployment; monitoring
"Whenever an AI system has a significant impact on people's lives, it should be possible to demand a suitable explanation of the AI system's decision-making process. Such explanation should be timely and adapted to the expertise of the stakeholder concerned (e.g. layperson, regulator or researcher)"	Process explainability	Deployment; monitoring
"In addition, explanations of the degree to which an AI system influences and shapes the organizational decision-making process, design choices of the system, and the rationale for deploying it, should be available (hence ensuring business model transparency)"	Business explainability	Business/use-case development; monitoring

HLEG x AI Lifecycle

	Governance structure	Safety culture	Reliable systems					Safety culture	Trustworthy certification
	Responsibility	Organization	Team					Organi-zation	Industry
	AI lifecycle stage	Business / use case development	Design	Training & test data procurement	Build-ing	Testing	Deploy-ment	Monitor-ing	*(Outside AI lifecycle)*
Requirements for trustworthy AI (HLEG)	Human agency and oversight	0	2	0	0	0	5	5	2
	Technical robustness and safety	0	5	0	1	6	7	1	0
	Privacy and data governance	0	1	4	0	2	2	0	1
	Transparency	1	0	0	0	0	6	4	0
	Diversity, non-discrimination and fairness	2	4	2	2	2	1	2	1
	Societal and environmental well-being	1	0	2	2	0	2	3	0
	Accountability	0	2	0	2	0	2	4	1

*Refer to Appendix 1 for specific sub-requirements



Results

3 sections

- Practical Findings
- Conceptual Findings
- Normative & Political Findings

Practical Findings and Implications

Practical Findings and Implications

- Lack of ethical requirements in design phase
- Lack of resources for meeting ethical requirements
- What the guidelines focused on:
  - accessibility
  - stakeholder participation
  - security
  - others (through impact and risk assessments)

Practical Findings (Cont.)

Managers

- No tangible method of applying on a practical level
- AI development requires non-tangible results such as
  - inter-departmental coordination
  - organizational change capacity (AGILE)
  - risk tendency

Developers

- Implementation and validating adherence to the guidelines is difficult
- Methodological issues:
  - (1) ethical requirements are not in alignment with AI development practices
  - (2) no standards for validation of ethics in software

Practical Findings (Cont.)

Problem 1

- Desirable structural translation method needs to be introduced
- Software Engineering (SE) + Requirements Engineering (RE)

Problem 2

- Problems:
  - Translation
  - Formulation
  - Validation
- SE community needs to step up in:
  - operationalization* of values in software
  - tractability of ethics in software
  - fairness in software

*operationalization = turning abstract conceptual ideas into measurable observations

Design for Values (DfV) community

- DfV is a methodological design approach that aims at making human values a part of the technological design, research, and development
- ICT products such as AI systems should embed human values through a ‘value hierarchy’, a hierarchy structure of values, norms and design requirements

Practical Findings

Problem

Translation and Operationalization

Solution

Design-for-Value (DfV)

Conceptual Findings and Implications

Conceptual Findings and Implications

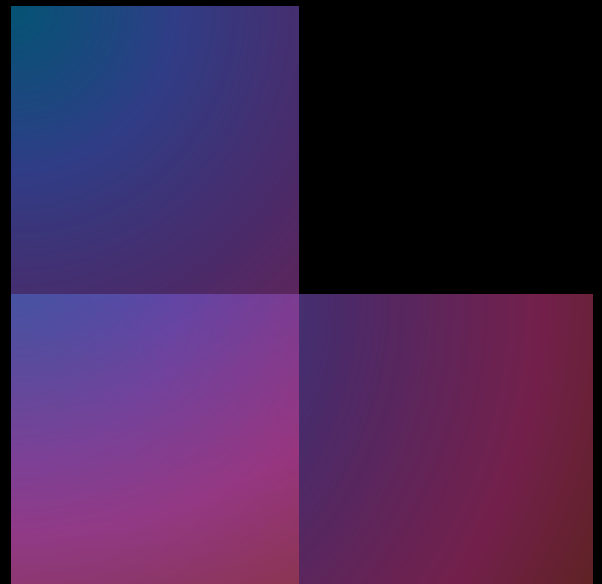
Lacks engagement in business and design stage

Focuses on deployment and monitoring phase

Focus points:

- Human Agency
- Transparency
- Accountability

HLEG framework has an interventionist approach, not a human-centric approach



Conceptual Findings (Solution)

- Governance level by Shneiderman's structures model
  - Human-centred AI
    - engage with diverse stakeholders
    - emphasize user experiences
  - Focus on:
    - measuring human performance and satisfaction
    - valuing customer and consumer needs
    - ensuring meaningful human control

Conceptual Findings

Problem

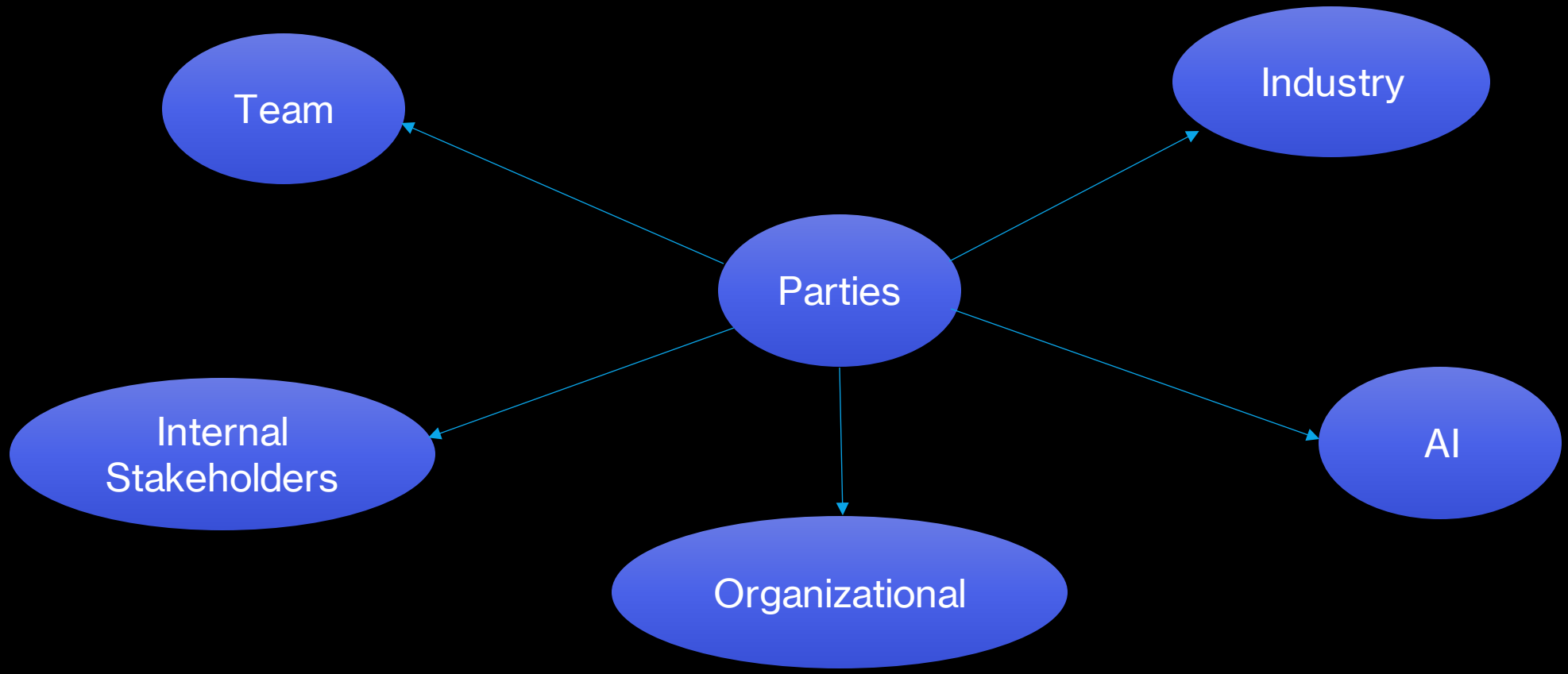
AI HLEG Framework isn’t human-centric

Solution

Scheidermann's Model (human-centric AI)

Normative & Political Findings and Implications

Normative & Political Findings



Normative & Political Findings and Implications

- Problem:
  - Accountability is only focused on the later part of the AI lifecycle
  - Ethical AI debate is placed exclusively on data scientists and not on business developers
  - Failure to acknowledge that standards, regulations, and agents are embedded within the AI lifecycle, the debate supports an **idealized** process
  - Hard to determine who's accountable

Normative & Political Findings (Cont.)

Option 1

- self-regulation needs widespread voluntary adoption

Option 2

- enforcing such regulation needs competent auditing institutions (which we don’t have)

Normative & Political Findings (Solution)

- One of the key elements for operationalization is having an applicable and effective governance framework
- Taking into account the:
  - design process
  - organizational reality
  - industry development
  - evaluation of business goals
  - risks for the company
  - end users
  - implicated actors of developing AI-based services

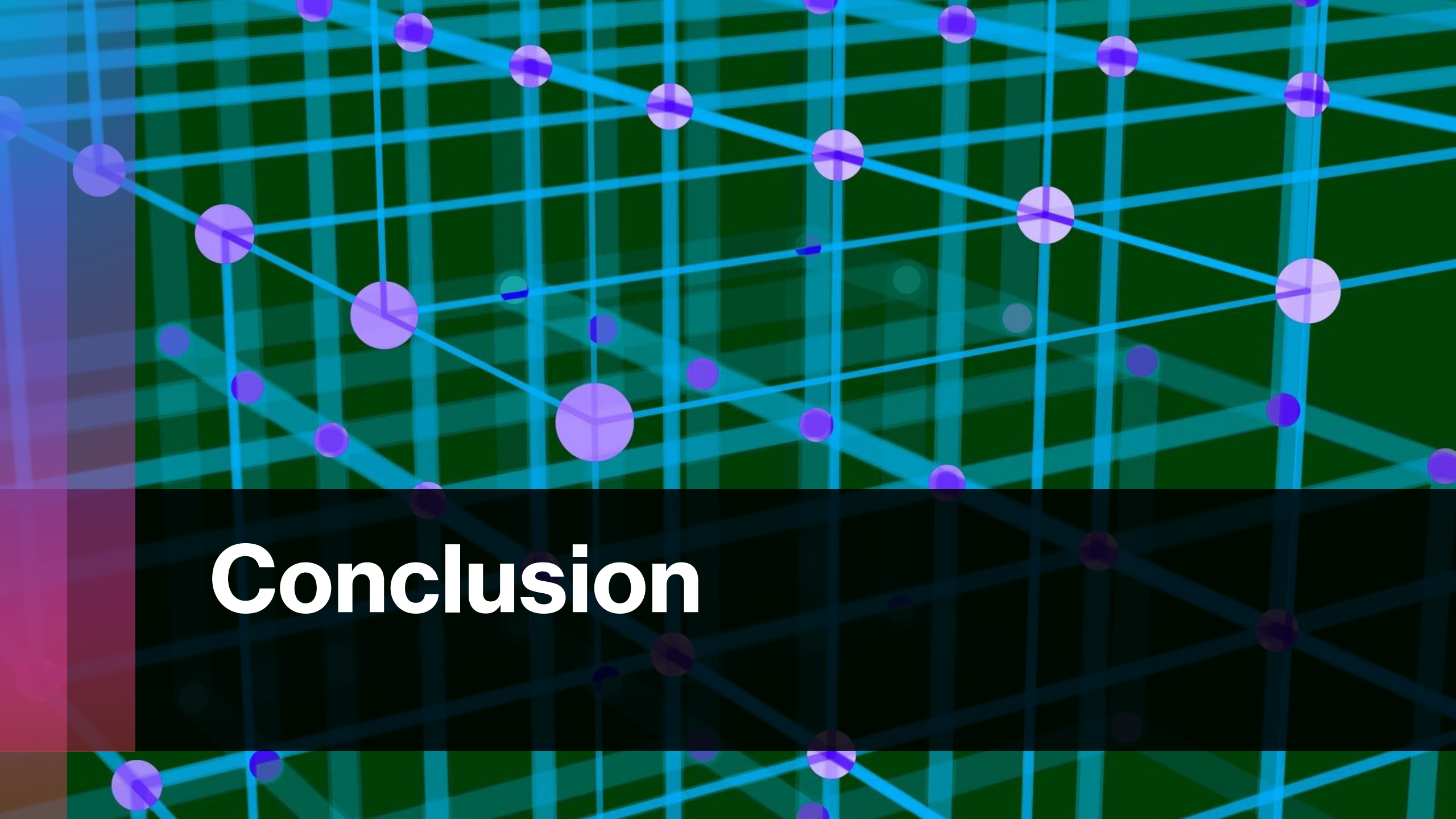
Normative & Political Findings

Problem

Hard to determine accountability

Solution

Create a model that focuses on other parties as well


Conclusion

Context

Nations have used the AI HLEG guidelines as mere "recommendations"

Problem:

Unsure how to implement these AI guidelines into their design and deployment

Unsure whose responsibility it is when it comes to AI ethics

Methodology

Step 1:

Define AI HLEG and AI lifecycle

Step 2:

Map AI HLEG to each stage of the AI-lifecycle using code

Step 3:

Create scores for each intersection

HLEG x AI Lifecycle

	Governance structure	Safety culture	Reliable systems					Safety culture	Trustworthy certification
	Responsibility	Organization	Team					Organi-zation	Industry
	AI lifecycle stage	Business / use case development	Design	Training & test data procurement	Build-ing	Testing	Deploy-ment	Monitor-ing	*(Outside AI lifecycle)*
Requirements for trustworthy AI (HLEG)	Human agency and oversight	0	2	0	0	0	5	5	2
	Technical robustness and safety	0	5	0	1	6	7	1	0
	Privacy and data governance	0	1	4	0	2	2	0	1
	Transparency	1	0	0	0	0	6	4	0
	Diversity, non-discrimination and fairness	2	4	2	2	2	1	2	1
	Societal and environmental well-being	1	0	2	2	0	2	3	0
	Accountability	0	2	0	2	0	2	4	1

*Refer to Appendix 1 for specific sub-requirements

Practical Findings

Problem

Translation and Operationalization

Solution

Design-for-Value (DfV)

Conceptual Findings

Problem

AI HLEG Framework isn’t human-centric

Solution

Scheidermann's Model (human-centric AI)

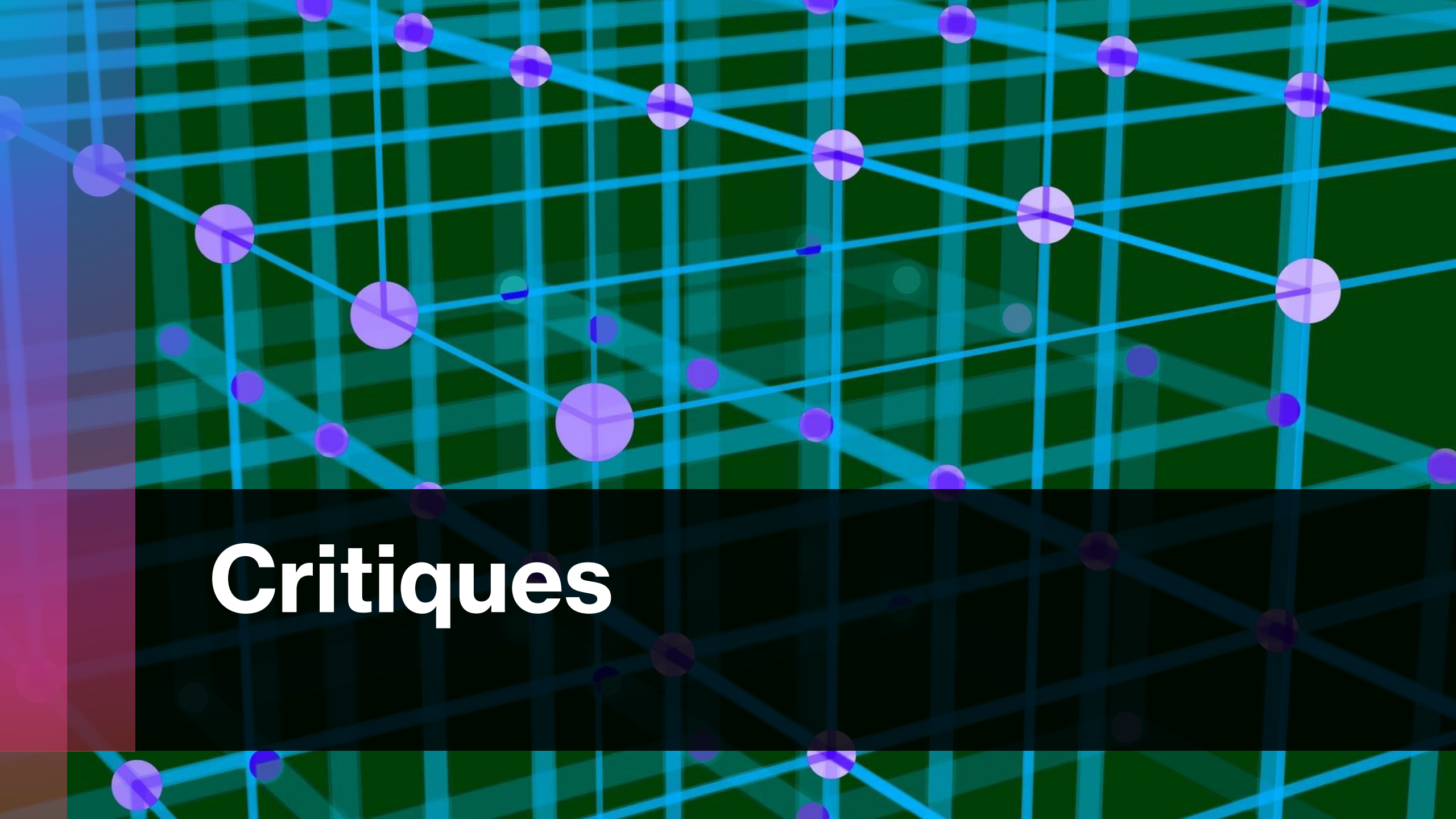
Normative & Political Findings

Problem

Hard to determine accountability

Solution

Create a model that focuses on other parties as well



Critiques

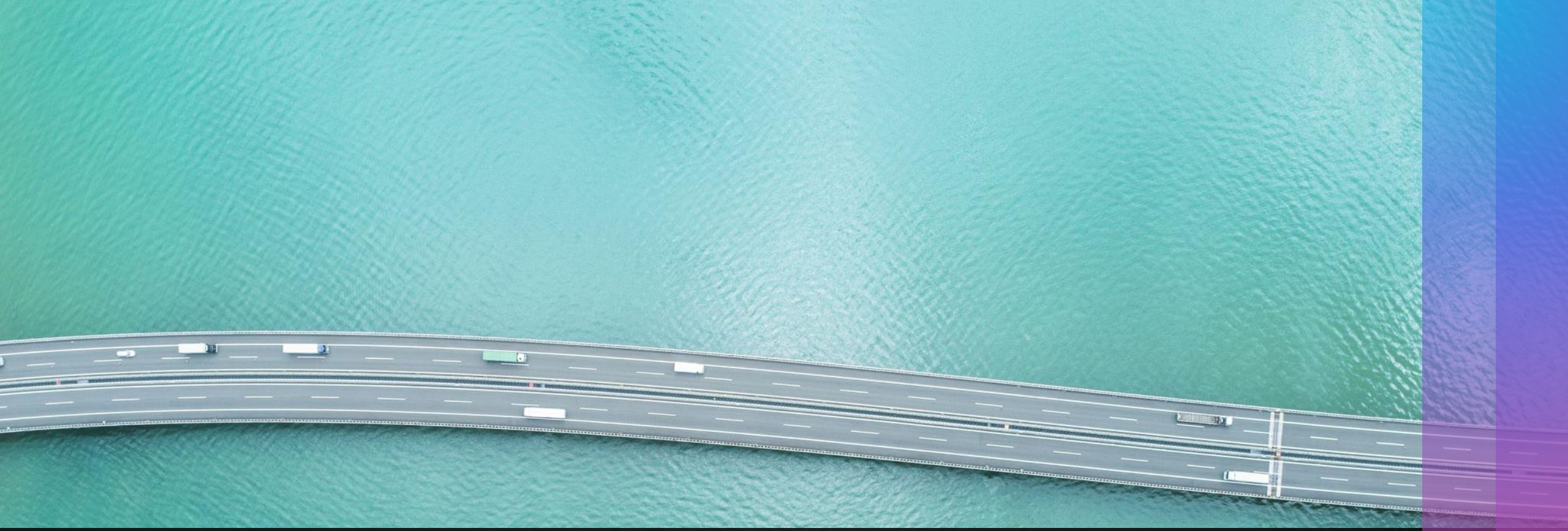
Critiques

1. Only using one case study
2. Only using one guideline through most of the study (AI HLEG)
3. Table lacks intuitive reading
4. Never expanded on the solutions section
5. Convoluted language

References

References

- Georgiva, Ilina, et al. *From AI Ethics Principles to Data Science Practice: A Refection and a Gap Analysis Based on Recent Frameworks and Practical Experience*. Springer Nature Switzerland, 7 Dec. 2021.



THANK YOU :)