DIGHUM – TU Wien

May 5, 2026

AI-RELATED RISK: A PHILOSOPHICAL ANALYSIS

Viola Schiaffonati (with Daniele Chiffi and Giacomo Zanotti)

Department of Electronics, Information and Bioengineering

POLITECNICO MILANO 1863

AGENDA

- A philosophical analysis of AI-related risk
  - The role of **multi-component analysis and its benefits**
  - The role of **multi-risk analysis** within a **socio-technical perspective**
- Why we need philosophy in the age of AI
  - **Theoretical analysis** to fill in **conceptual** and **policy vacuums**

AI AND EXISTENTIAL RISK

"Should we let machines flood our information channels with propaganda and untruth? Should we automate away all the jobs, including the fulfilling ones? Should we develop nonhuman minds that might eventually outnumber, outsmart, obsolete and replace us? Should we risk loss of control of our civilization?"

Future of Life Institute's open letter Pause Giant AI Experiments

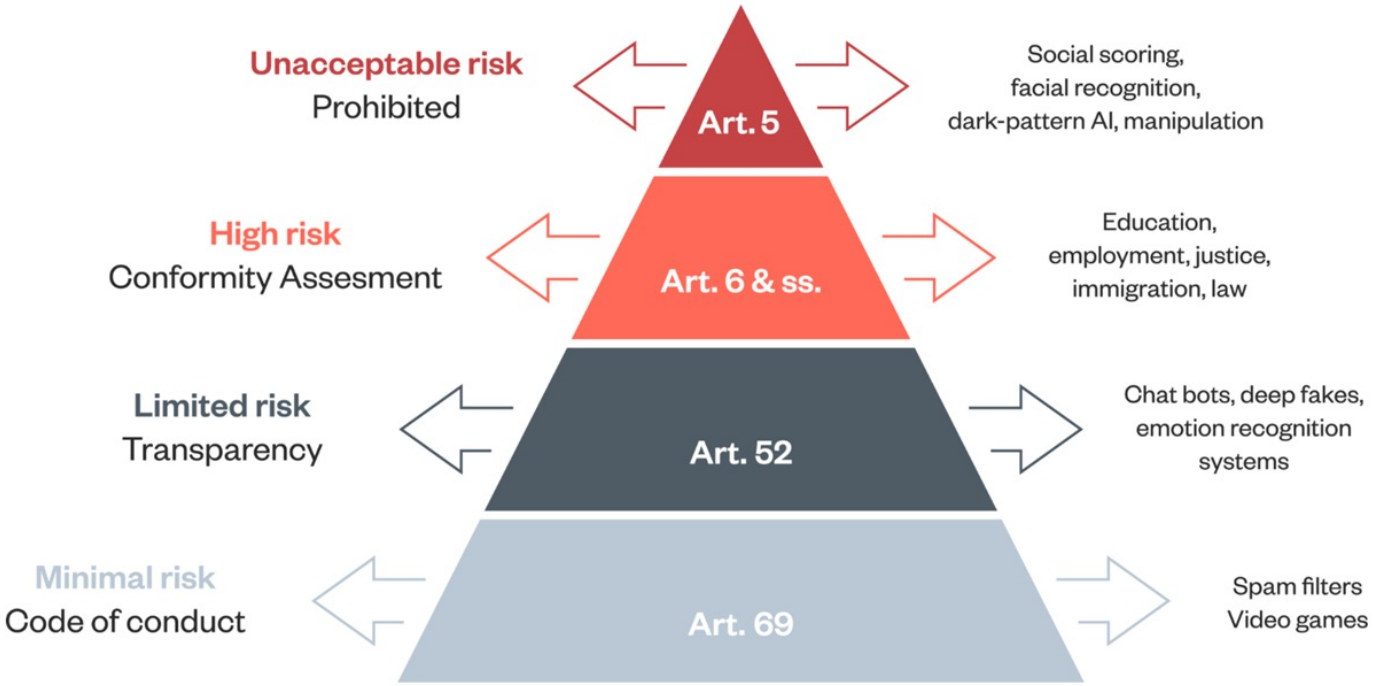
"Mitigating the risk of extinction from AI should be a global priority alongside other societal-scale risks such as pandemics and nuclear war."

Center for AI Safety's Statement on AI Risk

MORE CONCRETE RISKS

- Algorithmic discrimination (Buolamwini & Gebru 2018)
- Privacy violations (Curzon et al., 2021; Mühloff 2023)
- Environmental impact (Tamburrini 2022)
- …

AI ACT



- European AI Act: different levels of regulation for different levels of risk

DEFINING RISK

- No univocal definition of risk
  - Coexistence of technical and non-technical notions
  - Different technical definitions
- The classical approach: risk as the combination of the probability of an unwanted event occurring and the magnitude of its consequences (Royal Society 1983, Rechard 1999)

ZOOM IN: MULTI-COMPONENT ANALYSIS

- We focus on the different components of risk from **natural risk management**
- **Hazard**: the source of potential harm (volcano eruption)
- **Exposure**: people and material assets potentially subject to harm (people and buildings in the vicinity of the volcano)
- **Vulnerability**: the factors increasing or reducing the propensity of people and assets to be damaged by the hazard (vulnerable population such as older adults)

FROM NATURAL RISK TO AI

- The multi-component analysis of risk for AI-related risks can offer some advantages in terms of both conceptualization and mitigation (Zanotti, Chiffi, Schiaffonati 2024)

AI SYSTEMS IN CRITICAL CONTEXTS: HAZARD

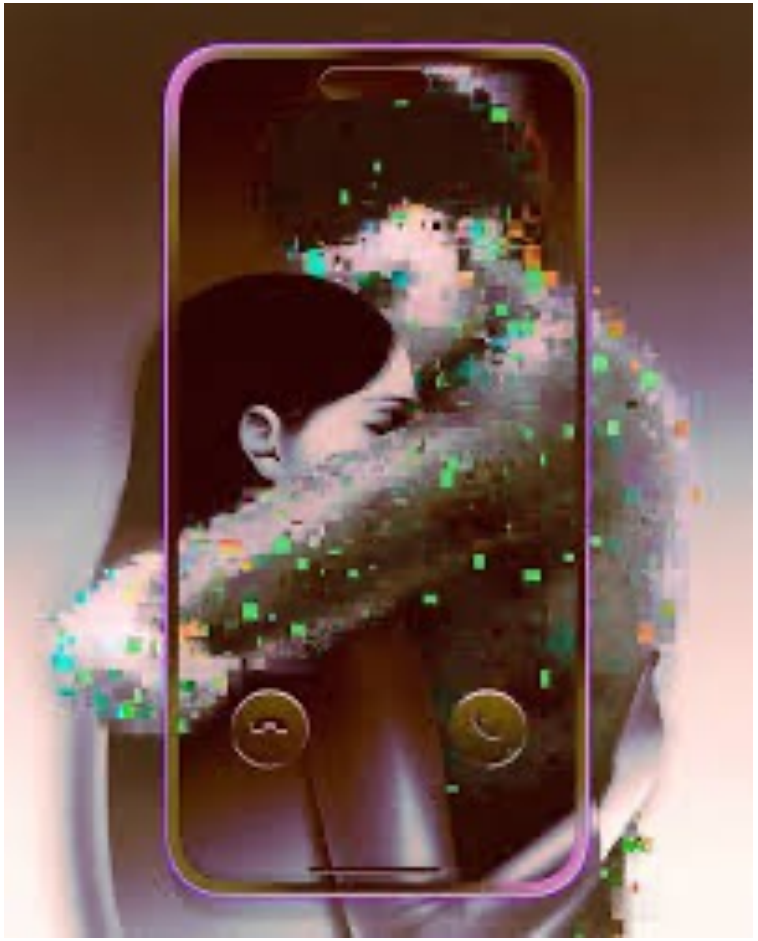
- We tend to focus on the **hazard**: AI systems are used in critical contexts
  - Medical AI (Panayies et al. 2020), courts (Queudot & Meurs 2018), war scenarios (Amoroso & Tamburrini 2020)
- AI systems increasingly delegated with decisions having important **impacts on people's lives**
  - Risks of medical AI: harm to patients, misuses, biases and inequalities, algorithmic opacity, privacy and security, gaps in accountability, obstacles in implementations (Lekadir et al. 2022)

EXPOSURE-CRITICAL SYSTEMS

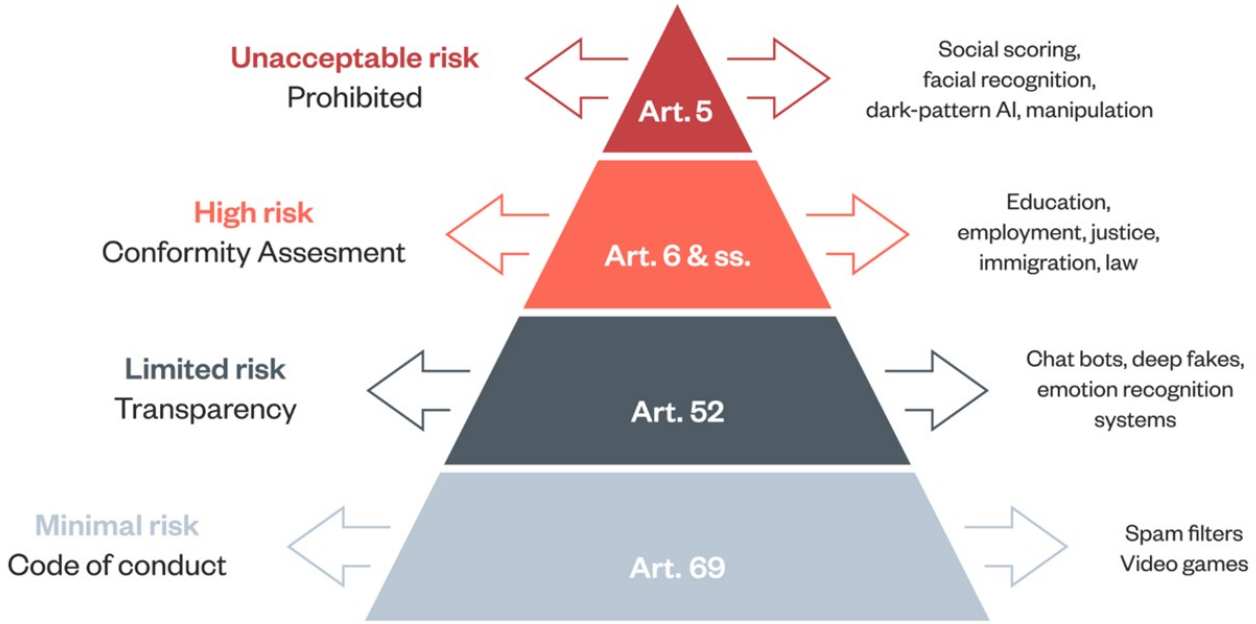
- Some systems are critical due to **exposure** levels
  - Recommender systems embedded in large online platforms (e.g., social networks)
- **Not high risk** if we consider only **hazard**
  - Recommender systems were absent from the AI Act's list of high-risk systems before the amendments of June 2023
  - Their addition can be understood if we consider the **component of exposure** of recommender systems' potential risks
- Not a matter of death or life but **very large-scale influence** (privacy, addiction, manipulation)

VULNERABILITY

- Concerns related to **vulnerability**
  - Affective computing and social robotics: systems interacting with human users in emotionally salient ways
- These systems are increasingly used
  - With **vulnerable populations** (children and elderly people)
  - In **contexts** of vulnerability (e.g., griefbots)
- These systems should be regarded as **highly risky due to the vulnerability** of their **users** and **contexts**



MULTI-COMPONENT RISK ANALYSIS



- Considering risk in AI **only** as a matter of **hazard** might **miss** some **important** aspects
- **Limited risk** (**deep fakes** can be very risky if **exposure** is high; **emotion recognition systems** can be critical if used with **vulnerable** population)

TAKING STOCK

- There is much debate on specific risks related to the use of AI systems, but a general philosophical framework is missing
- A multi-component analysis of risk can fruitfully be applied to AI systems
- This conceptual analysis considering different components of risk is important also for mitigation policies

ZOOM OUT: MULTI-RISK ANALYSIS

- From a multi-component analysis to a **multi-risk perspective**
- Multi-risk perspective can be **beneficial** to properly deal with AI risks connected to other types of risks in complex scenarios
- Again: inspiration from **natural risk analysis** (mainly)

FROM ISOLATION TO COMPLEXITY

- Discussion of AI-related risk considers AI systems – as well as their associated risks – **in isolation**, abstracted from the complexity of their deployment environment
  - E.g., medical AI systems and the risk of misdiagnoses, recommender systems and addiction, griefbots and emotional dependence
- This is a first (necessary) step but **not enough** due to the complexity of scenarios in which AI systems operate

AI SYSTEMS AS SOCIO-TECHNICAL SYSTEMS

- AI systems can be conceived as (part of) sociotechnical systems
- From discussions focusing on the technology (AI) itself to the broader systems in which AI functions (Kudina and van de Poel 2024)

AI SOCIO-TECHNICAL SYSTEMS

- AI systems are (part of) sociotechnical systems
  - Technical artifacts (physical objects fulfilling technical functions)
  - Human agents (developers, deployers, users)
  - Institutions (formal and informal rules)
  - Artificial agents (autonomy, interactivity, adaptativity)
  - Technical norms (related to artificial agents)

IMPLICATIONS FOR AI-RELATED RISK

- Consequences of adopting a socio-technical understanding in AI risk analysis

- **Risks** stem not only from **AI systems** but also from the **integrated socio-technical system** (humans and rules) that it is required to achieve some shared goals

AUTONOMOUS VEHICLES POWERED BY AI

- Different sources of potential harm (hazards)
  - Occasional malfunctions of the system
  - Systematic misclassification errors (affecting the accuracy of autonomous vehicle's perception algorithms)
- Human component
  - Human actions triggering risks (inappropriate behaviours of pedestrians, drivers' choices)
- Technical and human hazards
  - Social engineering exploiting personal information obtained from manipulating individuals to engage into cyberattacks

BEYOND ARTIFACTS AND HUMANS

- Role of **institutions**
- Hazards often occur in **policy vacuums**, in which there is **no regulation** that can effectively mitigate risks
- **Uncertainty** and **complexity** for policy makers (Nordström, 2022)

COMPLEX AI-BASED SOCIO-TECHNICAL SYSTEMS

- **Sociotechnical systems** are characterized by **uncertain phenomena** and **emergent risks**
- These systems do **not always** possess **shared goals** to be achieved (unlike simpler sociotechnical systems) (Simon 1961)
- They are complex socio-technical systems
  - E.g., autonomous vehicles powered by AI and the urban traffic system

COMPLEX AI-BASED SOCIO-TECHNICAL SYSTEMS

- Level of the **single vehicles**: risks emerging also from the interaction between technical artifacts and human agents
- Level of the **interaction between different vehicles** in their socio-technical composition
  - Interaction relying on the collection of significant amounts of data about the passengers and their surroundings, thus risk for individual privacy
- **Lack of shared goals** at the level both of the **single system** and of **the interacting systems**
  - Safety and efficiency vs privacy

THE ISSUES OF SINGLE RISK ANALYSIS

- As AI systems are often embedded within broader socio-technical systems risk assessment must account for the **dynamic interplay** between **technical** and **societal factors**
- Analysing a **single hazard** or a **single risk** is often **not enough**
  - Risks from multiple events can **exceed** the **sum** of **individual risks**
  - Almost always different hazards and risks are **connected** and **mutually interact**

MULTI-HAZARD AND MULTI-RISK

- Like other natural or technological hazards, **AI-related hazards** can **amplify**, **trigger**, or **interfere** with one another in ways that simple aggregation cannot capture

- Adopting more encompassing **multi-hazard** and **multi-risk perspectives** could enhance our capacity to analyse, understand, and mitigate the plural effects of multiple AI-related risks

MULTI-HAZARD APPROACH

- Focus on the component of **hazard** and their different **relations** (Gil & Malamud 2016)

- **Triggering**: a hazard directly causing another hazard

- **Increased probability**: a hazard raising the likelihood of another hazard without direct causation

- **Catalysis/impedance**: a hazard accelerating/intensifying or reducing/neutralizing the impact of another hazard

MULTI-HAZARD AND AI SYSTEMS

- Triggering (hazard directly causing another hazard)
  - In a personalized, AI-based continuous glucose monitoring system embedded in an insulin pump, an incorrect evaluation of glucose could bring about the administration of a wrong dose of insulin

MULTI-HAZARD AND AI SYSTEMS

- Increased probability (hazard raising the likelihood of another hazard)
  - In the context of AI-assisted medical diagnosis, the probability of a final wrong diagnosis following an inaccurate algorithmic output is increased by the doctor's overreliance on the AI system

MULTI-HAZARD AND AI SYSTEMS

- Catalysis and impedence (hazard accelerating or reducing the impact of another hazard)
  - An AI system could suggest a wrong treatment for an individual's medical condition, whose course could be thereby accelerated
  - AI-related hazards (wrong diagnosis and treatment) could be impeded by an unexpected power shortage that prevents the use of many tools in a hospital, including AI-based ones

TOWARD MULTI-RISK

- In a multi-hazard perspective, the interaction between different hazards is evaluated without considering the impact of different hazards on **exposure** and **vulnerability**

- Recently, disaster risk scholars have begun exploring **new multi-risk frameworks** to address the impact of multiple **hazards** concerning overall **vulnerability** and **exposure** (Pescaroli & Alexander 2018)

AN AI-RELATED EXAMPLE

- ML-based glucose monitoring tools trained on specific patients' glucose dynamics and integrated with an insulin pumping module (Medanki *et al.* 2024)

- Risks in terms of personalization (Canali et al. 2025)

AI-RELATED RISK AND MULTI-RISK

- Risks of this personalization are not only in terms of hazards but also of vulnerability
- Case of a pathology developed in a subclinical form (after the training of the device) that changes the glucose dynamics of the patient
  - The development of a new pathology impacting on glucose dynamics is a triggering hazard for a wrong insulin administration
  - This wrong insulin administration, in turn, catalyzes another hazard, namely diabetes-related complications
  - Overreliance on tool and reduced oversight on the system's functioning can make the patient more vulnerable with respect to inaccurate outputs of the system

TAKING STOCK

- The **overall risk** for the patient not emerging if considering only the relevant risks taken in isolation without considering their **relations** also on the **vulnerability** of exposed assets (e.g., patients)

- The **socio-technical lens** allowing to enlarge the view and considering not only risks stemming from **artificial agents** but also from **human ones** (e.g., overreliance mechanisms) and **institutions** (e.g., lack of governance and regulation)

- A **multi-risk** analysis (beyond the multi-hazard one) can **fruitfully** be **applied** to **AI** systems

COMPLEX SOCIO-TECHNICAL SYSTEMS

- The dynamics of hazard and vulnerability interactions emerge only when we go beyond the mere AI-related hazard, situating it in a **wider context of human practices**

- Again: this is important not only for the sake of **conceptual clarity** but also for devising **mitigation strategies**

CONCLUSIONS

- There is much debate on specific risks related to the use of AI systems, but a **general philosophical framework** is **missing**
- A **multi-component analysis** of **risk** can **fruitfully** be **applied** to **AI** systems and offers some benefits
- **Analysing AI-related risks in isolation is insufficient** - their potential interactions and combinations with other risks must be considered
- A **multi-risk perspective** is valuable for understanding AI-related risks within complex sociotechnical systems where risks from multiple events **can exceed the sum** of individual risks
- **Complex sociotechnical systems**, often associated to conflicting goals, require a **multi-risk approach** to address emergent and **uncertain risks**, in particular **systemic ones**

THANK YOU!

REFERENCES

Amoroso, D., & Tamburrini, G. (2020). Autonomous weapons systems and meaningful human control: Ethical and legal issues. *Current Robotics Reports*, *1*, 187–194. https://doi.org/10.1007/s43154-020-00024-3

Buolamwini, J., & Gebru, T. (2018). Gender shades: Intersectional accuracy disparities in commercial gender classification. Conference on Fairness, Accountability and Transparency, New York: PMLR, 77–91

Gill, J. C., & Malamud, B. D. (2016). Hazard interactions and interaction networks (cascades) within multi-hazard methodologies. Earth System Dynamics, 7(3), 659-679

Kudina, O., & van de Poel, I. (2024). A sociotechnical system perspective on AI. Minds and Machines, 34(3), 21

Marzocchi, W., Mastellone, M., Di Ruocco, A., Novelli, P., Romeo, E., & Gasparini, P. (2009). Principles of Multi-risk Assessment - Interaction amongst Natural and Man-induced Risks, https://op.europa.eu/en/publication-detail/-/publication/22eb788f-5d0a-496a-92d4-4759b0b57fde/language-en

Nordström, M. (2022). AI under great uncertainty: implications and decision strategies for public policy. *AI & society*, *37*(4), 1703-1714. https://doi.org/10.1007/s00146-021-01263-4

Pescaroli, G., & Alexander, D. (2018). Understanding compound, interconnected, interacting, and cascading risks: a holistic framework. Risk Analysis, 38(11), 2245-2257

Simon, H.A. (1961). The architecture of complexity. Proceedings of the American Philosophical Society, 106(6): 467-482

Zanotti, G., Chiffi, D. & Schiaffonati, V. AI-Related Risk: An Epistemological Approach. *Philos. Technol.* **37**, 66 (2024). https://doi.org/10.1007/s13347-024-00755-7