Ethical Risk Assessment of AI in Practice Methodology: Process-oriented Lessons Learnt from the Initial Phase of Collaborative Development with Public and Private Organisations in Norway

Authors: Murashova N., Lindblom D., Omerovic A., Dahl H., Bergsjø L.

The Eighteenth International Conference on Advances in Computer-Human Interactions
ACHI 2025

Presenter PhD Candidate Natalia Murashova
Østfold University College,
natalia.murashova@hiof.no

Nice, France, May 18-22, 2025

IARIA

Østfold University College

ENACT

Natalia Murashova (Østfold University College) is a PhD candidate at the Department of Teacher Education. Natalia has background in Pedagogy, Social Sciences and Human Geography. In the recent years she closely worked with the topics of ***Responsible Innovation and Research, Co-creation and Participation, Stakeholder Engagement and AI Ethics.***

In her doctoral thesis she explores methodological approaches of operationalizing guidelines for ethical and responsible AI in organisational practice through participatory and co-creation techniques.

Get in touch!

natalia.murashova@hiof.no

LinkedIn

Ethics guidelines for trustworthy AI

On 8 April 2019, the High-Level Expert Group on AI presented Ethics Guidelines for Trustworthy Artificial Intelligence. This followed the publication of the guidelines' first draft in December 2018 on which more than 500 comments were received through an open consultation.

According to the Guidelines, trustworthy AI should be:

(1) lawful - respecting all applicable laws and regulations

(2) ethical - respecting ethical principles and values

(3) robust - both from a technical perspective while taking into account its social environment

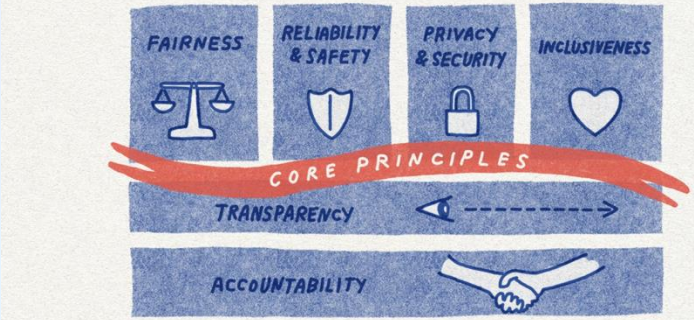
See also

A European approach to artificial intelligence

Related topics

Advanced Digital Technologies

Artificial intelligence



Draft, February 12, 1992

ACM Releases Draft Revision of Ethics Code. ACM's present Code of Professional Conduct was developed in 1972, predating technological developments such as widespread data networks and "computer viruses." With funding from the SIG Discretionary Fund, ACM's Special Interest Group on Computers and Society (SIGCAS) sponsored an Ethics Task Force to revise the ACM Code. The proposed draft ACM Code of Ethics published in this issue was developed by the Task Force and has been reviewed by ACM Council.

All ACM members are invited to comment on this draft. Suggestions to express the code in less USA-centric terms are especially welcome, since the issues are equally important for all ACM members. Based on feedback from the membership, a final draft of the code will be developed and presented to ACM Council for approval.

ACM Code of Ethics and Professional Conduct

Table 2 Comparison of ethical principles in recent publications demonstrating the emerging consensus of 'what' ethical AI should aspire to be

AI4People (published November 2018) (Floridi et al. 2018)	Five principles key to any ethical framework for AI (L Floridi and Clement-Jones 2019)	Ethics Guidelines for Trustworthy AI (Published April 2019) (European Commission 2019)	Recommendation of the Council of Artificial Intelligence (Published May 2019) (OECD 2019b)	Beijing AI Principles for R&D (Published May 2019) ('Beijing AI Principles' 2019)
Beneficence	AI must be beneficial to humanity	Respect for human autonomy	Inclusive growth, sustainable development and well-being	**Do good:** (covers the need for AI to promote human society and the environment)
Non-Maleficence	AI must not infringe on privacy or undermine security	Prevention of harm	Robustness, security and safety	**Be responsible:** (covers the need for researchers to be aware of negative impacts and take steps to mitigate them) **Control risks:** (covers the need for developers to improve the robustness and reliability of systems to ensure data security and AI safety)
Autonomy	AI must protect and enhance our autonomy and ability to take decisions and choose between alternatives		**Human-centred values** and fairness	**For humanity:** (covers the need for AI to serve humanity by conforming to human values including freedom and autonomy)
Justice	AI must promote prosperity and	Fairness	Human-centred values **and fairness**	**Be diverse and inclusive:** (covers the need for AI to benefit as many people as possible) **Be ethical:** (covers the need to make the system as fair as possible, minimising discrimination and bias)
Explicability	AI systems must be understandable and explainable	Explicability	Transparency and explainability Accountability	**Be ethical:** (covers the need for AI to be transparent, explainable and predictable)

For a more detailed comparison see Floridi and Cowls (2019) and Hagendorff (2019)

The classification of the risks posed by AI systems in EU AI Act

Unacceptable Risk

Highest level of risk prohibited in the EU. Includes AI systems using e.g. subliminal manipulation or general social scoring.

Limited Risk

Includes AI systems with a risk of manipulation or deceit, e.g. chatbots or emotion recognition systems. Humans must be informed about their interaction with the AI.

High Risk

Most regulated AI systems, as these have the potential to cause significant harm if they fail or are misused, e.g. if used in law enforcement or recruiting.

Minimal Risk

All other AI systems, e.g. a spam filter, which can be deployed without additional restrictions.

Image source: https://www.trail-ml.com/blog/eu-ai-act-how-risk-is-classified

Challeng

Figure 1- Geographic distribution of issuers of ethical AI guidelines by number of documents released

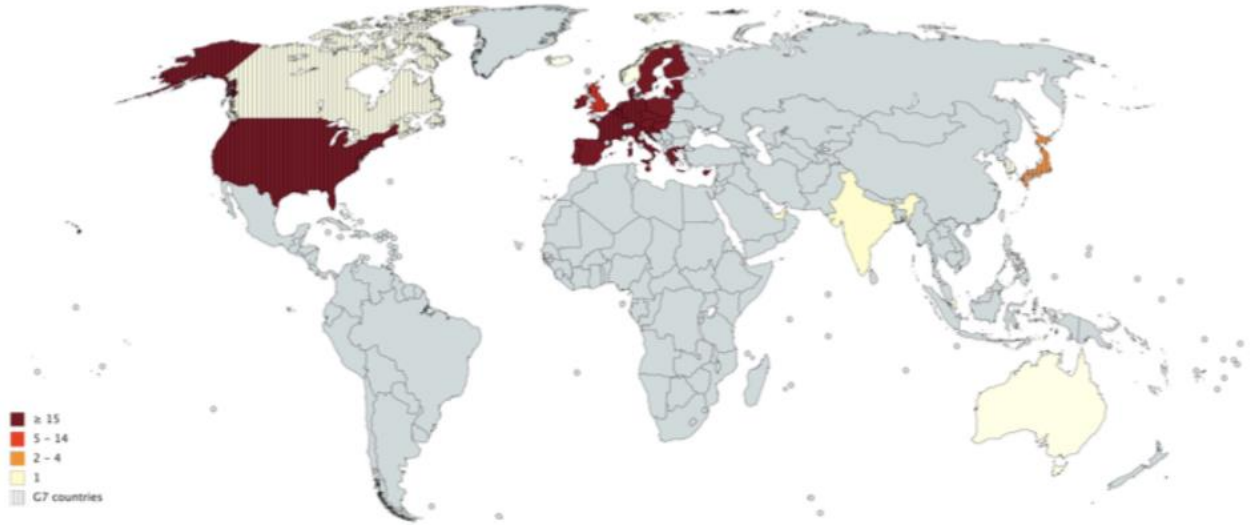
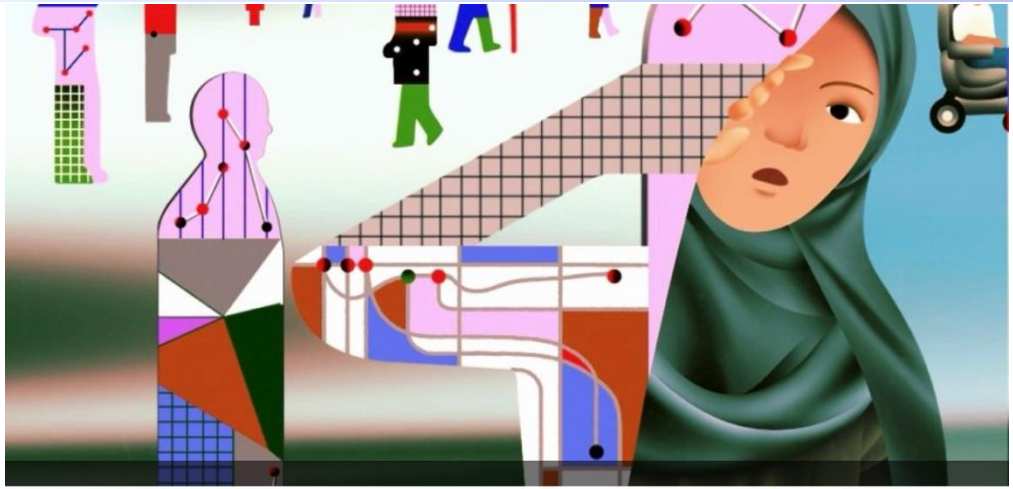


Figure 1: Geographic distribution of issuers of ethical AI guidelines by number of documents released. Most ethics guidelines are released in the United States (n=20) and within the European Union (19), followed by the United Kingdom (14) and Japan (4). Canada, Iceland, Norway, the United Arab Emirates, India, Singapore, South Korea, Australia are represented with 1 document each. Having endorsed a distinct G7 statement, member states of the G7 countries are highlighted separately. Map created using mapchart.net.

From Jobin et. al. (2019)

AI hypes and hopes



< NEWS

12 November 2024

Denmark: AI-powered welfare system fuels mass surveillance and risks discriminating against marginalized groups – report

AI can help shape society for the better – but humans and machines must work together

D Fox Harrell

Collaboration rather than command-and-control is key to creating culturally and ethically positive systems

Kommune tatt for KI-bruk: – Dette er pinlig

Tromsø kommune brukte kunstig intelligens som hjelpemiddel i arbeidet med en viktig rapport. Rapporten inneholdt flere feil, noe KI-ekspert mener kunne vært unngått.

Tromsø kommune brukte kunstig intelligens som hjelpemiddel i arbeidet med ny skolestruktur. Av 18 oppførte kilder i dokumentet, kan bare syv av dem spores opp.
FOTO: SVEINUNG ÅSALI / NRK

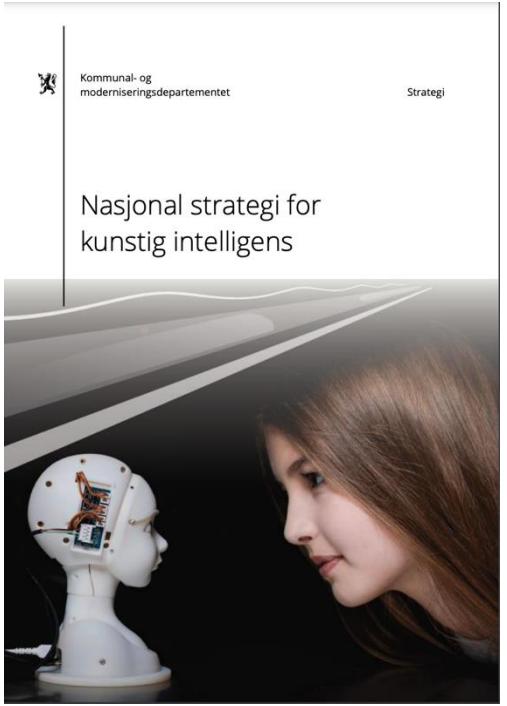
Trygve Grønning
Journalist
+ 3 til

Vi rapporterer fra Tromsø

Publisert 27. mars kl. 20:00
Oppdatert 28. mars kl. 09:47

Norwegian context

- Ca. 5 million inhabitants > aging population
- Focus on digitalization of the public sector.
- National strategy for AI (2020).
- In 2025, 1 billion NOK will be allocated to open 5-6 National AI research centers in Norway.
- National Wealth Fund – 1,7 trillion (March 2025)



ENACT Project

Ethical risks assessmeNt of Artificial intelligenCe in pracTice (ENACT) is a project funded by the Research Council of Norway. The project aims to develop a methodology governing ethical principles and guidelines for the Norwegian public and private sector deploying AI-based systems.

ENACT is a collaboration of academia, social services, finance, healthcare, logistics, education sectors.

Challenges for implementing ethical AI guidelines in organisational practice

Rapid development of AI systems

Professional competence gaps

Different sectoral traditions

Lack of participatory engagement mechanisms

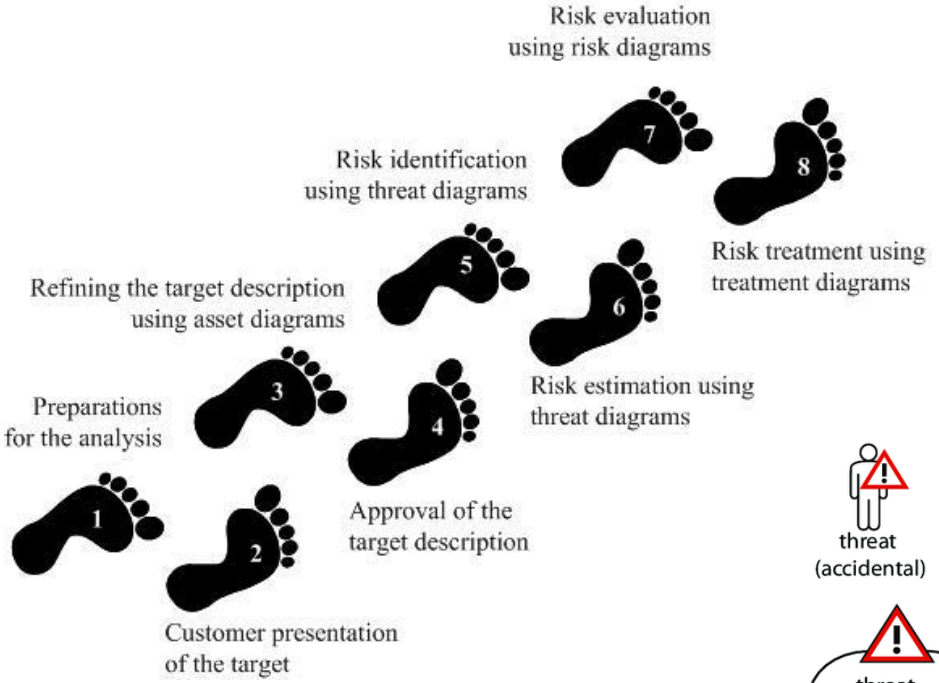
Constant adjustment of products and services

Gap between principles and practice

Ethical Risk Assessment of AI in practice

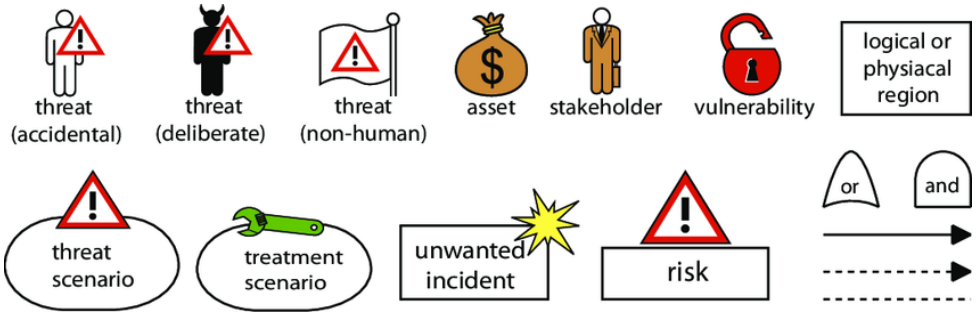
- Open-ended qiuestions, Likert scale, risk visualisation *(Tartaro et. al 2024)*
- Data driven risk assessment with expert knowledge *(Felländer et. al 2022)*
- Relational approach: decision-maker –risk exposed – benificiary (Krijger, 2024)

Theoretical framework

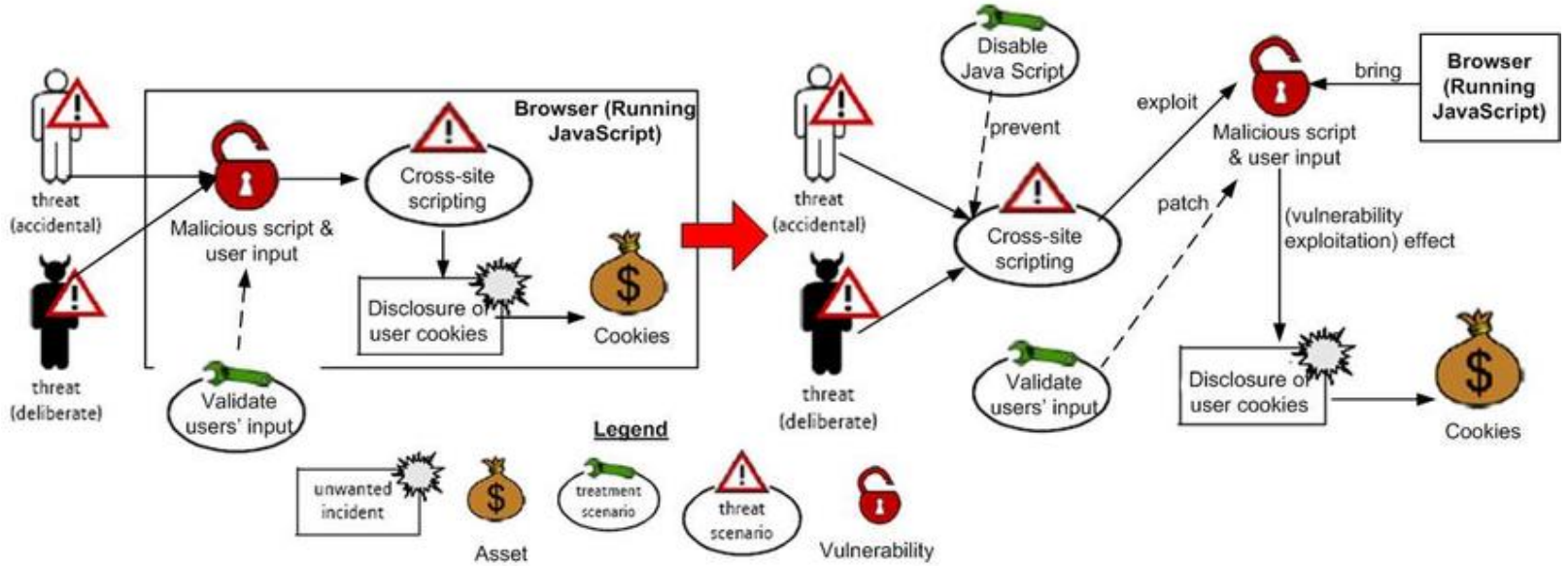


CORAS - Model-based method for security risk analysis (Lund et. al 2010)

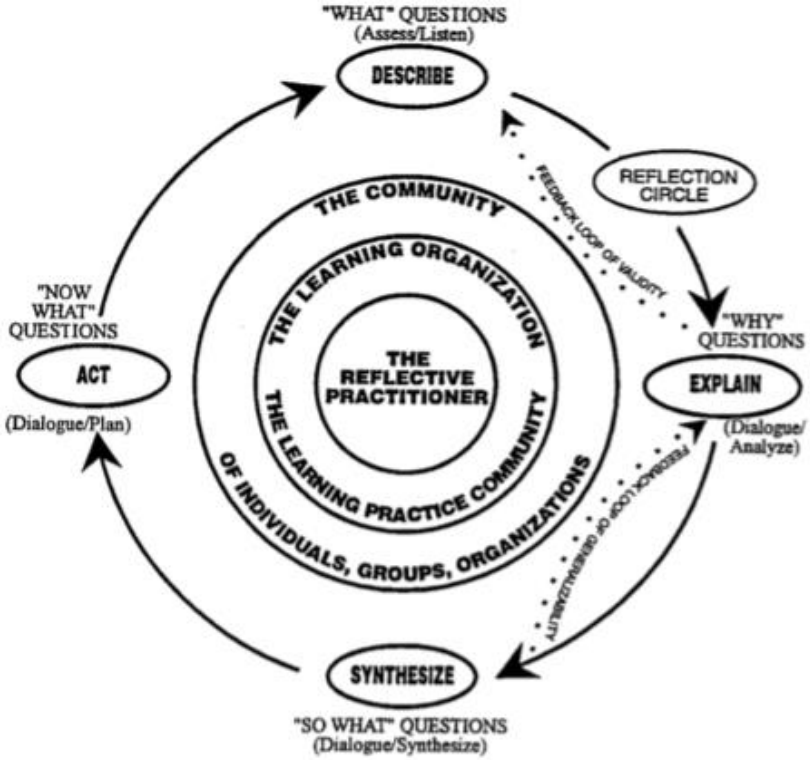
- conducted in three phases: context establishment, risk assessment and risk treatment
- graphical style of the communication, visual modelling, constructive use of language and tighter integration of the assessment outputs in the system development processes



CORAS analysis example



Theoretical framework



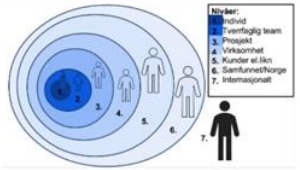
Story Dialog Method (Labonte et. al 1999)
- both a data collection and a data analysis method, based on a structured dialogue and on participants' stories.

- Describe (where WHAT-type of questions are asked)
- Explain (WHY-type of questions are asked)
- Synthesis (where SO WHAT-type of questions are asked)
- Action (NOW WHAT-type of actions are asked)

Workshop structures

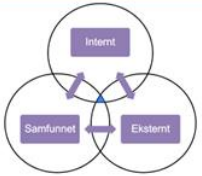
WS1

- Introduction, information and expectations
- Key features of use case, stakeholders, timeframe
- Values and actors in plenum discussion
- High level analysis to identify prioritized values



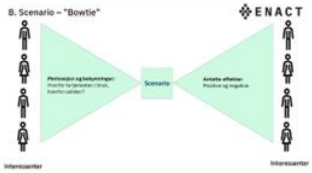
WS2

- Information and expectations
- Values – identifying values and discussion
- Scenarios – identifying values in plenum discussion
- Summary and evaluation



WS3

- Information and expectations
- Scenarios and bowties diagram
- Brainstorming in groups
- Summary of group discussion and initial measures
- Summary and evaluation



Data collection and analysis

- The working group was comprised of the researchers with expertise in *ethics, risk management, technology and pedagogy*.
- Sensemaking through iterative assessment of the notes, collective reflection and synthesis of textual data and researchers' observations.
- Participants selection!

TABLE I
OVERVIEW OF THE PROCESS

Workshops and CORAS steps	Workshop 1: Establishment of the context	Workshop 2: Risk assessment	Workshop 3: Risk assessment
Time and format	60 mins, Digital workshop	60 mins, Digital workshop	60 mins, Digital workshop
Structure of the workshop	1. Introduction, information and expectations 2. Key features of use case, stakeholders, timeframe 3. Values 4. High level analysis	1. Information and expectations 2. Values – identifying values and discussion 3. Scenarios – identifying values and discussion 4. Summary	1. Information and expectations 2. Scenarios 3. Group brainstorming 4. Summary of group discussion and initial measures 5. Summary and evaluation
Participants	10 participants 2 facilitators 3 observers	10 participants 2 facilitators 3 observers	9 participants 2 facilitators 2 observers
Ground for methodology adjustment	Pre-workshop survey, Work group meetings	Post workshop survey, Work group meeting, ENACT business partner meeting, ENACT project meeting	Work group meeting, ENACT business partner meeting
Concepts	Actors	Values	Scenarios
Documentation	Pre-workshop survey, 8.5 pages of structured notes	Post-workshop survey, 8 pages of structured notes	2 pages of notes

Ethical risk assessment of Microsoft Pilot Transcribing in practice

Lesson 1. The scope of ethical risk in cross-sectoral settings

- Common “analysis context”
- Boundaries of cross-sectoral settings
- Different sectoral tradition

Lesson 2. Flexible methodology helps to address organisational needs

- Meeting organisational needs
- Digital format
- Recourse efficient
- Embedding in everyday practice of organisation

Lesson 3. Easing power-relationship for structured dialog and critical reflection

- Reduced number of participant for plenum discussion
- Separation from fellow colleagues
- Business confidentiality

Future work and further development of ethical risk assessment of AI in practice

- sectoral tradition (e.g., similarities and differences between the domain of ethics and security standards with respect to risk assessment)

- group dynamics (e.g., power dynamics in the group, business integrity, approaches to elicit organisational needs)

- confidential information of organisational practices

- format of the workshops, which had to be realistic (e.g., time, digital or physical meetings, resources required) if the businesses were to use the methodology in real world settings

References

1. **A. Jobin**, I. Marcello, and V. Effy, "The global landscape of AI ethics guidelines." Nature machine intelligence 1.9 (2019): 389-399.
2. J. **Morley**, L. Floridi, L. Kinsey, and A. Elhalal, "From what to how: an initial review of publicly available AI ethics tools, methods and research to translate principles into practices," Science and Engineering Ethics, vol. 26, no. 4, pp. 2141-2168, Aug 2020, doi.org/10.1007/s11948-019-00165-5.
3. A. **Tartaro**, E. Panai, and M. Z. Cocchiaro, "AI risk assessment using ethical dimensions," AI and Ethics, pp. 1-8, Jan 2024, doi.org/10.1007/s43681-023-00401-6
4. A. **Felländer**, J. Rebane, S. Larsson, M. Wiggberg, and F. Heintz, "Achieving a data-driven risk assessment methodology for ethical AI," DISO, vol. 1, no. 2, p. 13, Aug 2022, doi.org/10.1007/s44206-022-00016-0.
5. J. **Krijger**, "What About Justice and Power Imbalances? A Relational Approach to Ethical Risk Assessments for AI," DISO, vol. 3, no. 3, p. 56, Oct 2024,doi.org/10.1007/s44206-024-00139-6.
6. R. **Labonte**, J. Feather, and M. Hills, "A story/dialogue method for health promotion knowledge development and evaluation," Health Education Research, vol. 14, no. 1, pp. 39-50, Feb 1999,doi.org/10.1093/her/14.1.39.
7. M. S. **Lund**, B. Solhaug, and K. Stølen, Model-driven risk analysis: the CORAS approach. Springer, pp. 23-43, 2010.

Natalia Murashova
Østfold University College

natalia.murashova@hiof.no

Diana (Saplacan) Lindblom
University of Oslo

dianasa@ifi.uio.no

Aida Omerovic
SINTEF Digital

aida.omerovic@sintef.no

Heidi E. I. Dahl
Posten Bring

heidi.dahl@posten.no

Leonora Onarheim Bergsjø
Østfold University College

leonora.bergsjo@hiof.no

Thank you for your attention!

Get in touch
natalia.murashova@hiof.no

LinkedIn

For more information about ENACT project visit
enactai.no