Addressing **Ethical Dilemmas** in **AI**: Listening to Engineers

IEEE SA STANDARDS ASSOCIATION
ane ASSOCIATION OF NORDIC ENGINEERS
UNIVERSITY OF COPENHAGEN
DATAETHICS

Table of Contents

Introduction

The rapid development of artificial intelligence (AI) systems brings immense opportunities as well as considerable ethical concerns. There are numerous discussions among policy makers, industry professionals and academics about how to capitalise on the significant potential these technologies offer and seek ways to minimise the negative impacts of AI systems, which could include unemployment, inequality and threats to democracy. To confront the growing challenges, engineers and technology developers could potentially make use of the many guidelines and ethical principles available to them. However, it is unclear whether and how the newly emerging guidelines and standards for ethical and responsible AI, such as the IEEE Ethically Aligned Design effort, might be integrated into existing technology development processes.

Engineering through its very essence seeks to have an impact on society, and this impact bears responsibilities, obligations, and rights. Concerns around the development of AI systems demands a shift in the mindset for engineers, their organisations, as well as for policy decision-makers about how technologies are designed. Therefore, it is important to engage engineers and technology developers in the discussion about AI and ethics in order to understand the challenges and opportunities for responsible technology development.

This report[1] is based on the proceedings of the online hackathon "Ethical dilemmas in AI - engineering the way out", conducted in September 2020 by the Association of Nordic Engineers (ANE), the Data Ethics ThinkDoTank (DataEthics.eu), the Institute of Electrical and Electronics Engineers (IEEE) and the researchers from the Department of Computer Science at the University of Copenhagen. The goal of the hackathon was to identify the main challenges for integrating existing ethical principles and guidelines into the engineering processes that power AI development.

The report presents the findings from the discussions with professional engineers about ethics in AI development conducted as part of the hackathon. We detail engineer needs and concerns in seeking routes to the development of ethical and responsible AI. The standards and guidelines currently being developed are essential but require processes that ensure their implementation. It is of paramount importance to define specialised responsibilities and put in place local and clearly determined structures for accountability. Principles alone cannot guide our technology development. Engineers stressed the need to better govern decisions about AI technologies *in practice*.

Further research and policy decision-making is needed to build on, challenge and adapt the recommendations presented here. Ethics in AI is a process, which requires direction that must be

1 This report has been jointly developed by Professor Irina Shklovski's team from the Department of Computer Science at the University of Copenhagen (DIKU) with the Association of Nordic Engineers (ANE), and the Data Ethics ThinkDoTank (DataEthics.eu), with support from the Institute of Electrical and Electronics Engineers (IEEE)

refined through consistent collaboration with engineers at all levels and from diverse sectors, and in cooperation with other multi-disciplinary experts. Initiatives to spark collective efforts towards discussing AI ethics in practice should be promoted within all workplaces, and throughout all levels of employment.

RECOMMENDATIONS FOR GOVERNMENT BODIES AND ORGANISATIONS

1. **PUTTING IN PLACE A GOVERNANCE FRAMEWORK** – without proper governance measures and some level of external engagement it is impossible to implement the necessary standards for ethical and responsible AI. There is a need for institutional unity to ensure oversight over obligations towards ethical practices. Governance should emerge through negotiation between engineers, their organisations, and various government institutions to define responsibilities and processes to regulate these responsibilities. Inspiration can be found in the challenges identified by the engineers and in the nodes of certainty they proposed as steps to take towards a supportive governance framework. These governance structures will be ultimately responsible for implementing standards, best practices and audits for AI systems, as well as training programs and certifications for people who develop and use AI.

2. **DEFINING RESPONSIBILITIES AND ACCOUNTABILITY** – the engineering profession cannot shoulder all of the responsibility for the wrongdoings in AI. There is a need to define and introduce clear frameworks for how responsibility is distributed and accountability is performed. Distributing responsibility successfully requires the use of standards and regulations, and necessitates a no-blame culture in which taking on responsibility does not lead to punitive consequences if unexpected ethical issues arise. A 'no-blame' approach allows opportunities for learning from prior mistakes, thus developing best practices for addressing ethical challenges. Accountability frameworks should include appeal instances and clearly defined whistleblower protection mechanisms to support individual and collective processes for challenging unethical practices or outcomes. This is especially important as new concerns come to light due to broadening implementation of advanced AI systems.

3. **CREATING SPACES FOR**: 1) education and training through which additional interdisciplinary competences for addressing ethics in AI could be acquired; 2) discussions about ethical challenges and concerns with AI within the workplace, at external organisations supporting interdisciplinary debates across domains and disciplines, and on a political level between politicians and engineers; 3) engineers to be able to voice concerns about ethical issues and challenges in AI development, and engage with external support and whistle-blower protection. These spaces are intended to enable ethics as an iterative process by providing engineers with spaces to identify ethical issues and challenges as well as with spaces for debating and challenging current standards and processes, ensuring support for adaptive regulatory mechanisms.

4. **SUPPORTING DIFFERENT FORMS OF DOCUMENTATION TO ENSURE TRANSPARENCY** of the development processes. This requires new recommendations for what to document, by whom, how, and when. In practice, the process of AI system development and design decisions already complies with a range of traditional technical documentation practices. Technical documentation should also include policy and decisions being made in system development, as well as descriptions

of system architecture, data sources and flows, and information for the end-user, going beyond basic requirements.

5. PUSHING FOR EXPLAINABILITY – explainability is an obligation, a challenge, and a design choice from an engineering point of view. Where full explainability is not possible, human-in-the-loop approaches and testing solutions can help bridge the gap between complex AI models and the need to navigate the problems of potential bias and unfairness in automated decision-making outcomes. When it comes to testing, there is a need for frameworks and requirements for extensive real-life testing scenarios to be conducted prior to full system launch as well as infrastructures to support testing throughout system life cycle.

RECOMMENDATIONS FOR ENGINEERS AND THEIR ORGANISATIONS

1. TRANSLATION: Translation focuses on how ethical guidelines and principles can be translated into specific engineering practices, and on how to learn to recognise whether something may be or may become an ethical issue.

– Examples from findings:

i. Engineers need to engage with stakeholder groups during system design to deliver practical solutions that foster the adoption of AI sytems while mitigating potential risks. To achieve this they need to be aware of appropropriate standards and best practices. Where standards and best practices are not available or need to be adopted, engineers should engage in multidisciplinary, open and consensus based processes, on how ethical principles can be implemented into practice.

ii. Engineers must ensure that in working on AI systems they have the ability to translate (i.e., ensuring appropriate expertise in taking broader issues into account) and to engage in expanded documentation practices to support transparency and explainability of AI systems they develop.

iii. Organisations must support expanded documentation practices (i.e., through open audit trails and real-time oversight, going beyond traditional technical documentation and creating documentation oriented toward different audiences).

iv. Organisations must support engineer engagement in spaces for education (i.e., to encourage growth in competences through lifelong learning) and spaces for discussion (i.e., supported attendance and activity in collective discussions).

2. IMPLEMENTATION: Implementation refers to the practices developed to mitigate ethical concerns.

– Examples from findings: Since ethical issues relating to AI technologies often arise from unforeseen consequences of AI system design, technical mitigation, likewise, involves engineering practices.

i. Engineers and their organisations must support high levels of engagement inside and outside of organisations with public authorities or stakeholders involved with the development and implementation of AI systems. Examples of technical mitigation of ethical concerns include the development of new forms of documentation, instituting administrative and automated oversight for identified ethical concerns, and new kinds of certification of individual expertise and organisational processes for AI system development.

ii. It is important to ensure that technical mitigation approaches are not imposed top-down but negotiated democratically through spaces for discussion.

3. **ACCOUNTABILITY: Accountability refers to clarity about who ought to be accountable and to whom for the outcomes of technology use, who is responsible when things go wrong and how they should be held to account.**

– Examples from findings:

i. Engineers bear the responsibility to voice their opinions and to propose solutions for appropriate structures of accountability. Through spaces for discussion, engineers can negotiate how to distribute responsibility and accountability, while spaces for concern offer opportunities for challenging existing practices, norms and standards and should provide structures for the performance of accountability.

ii. Organisations shall strive to develop forms of no-blame culture that can facilitate admissions of mistakes, enabling the organisation to learn and move forward. It is important to consider how responsibility and accountability can be distributed to ensure good performance but not dampen the ability of engineers to question and express concerns.

iii. Organisations shall seek appropriate standards, certifications, documentation and testing. They constitute fundamental infrastructures of support that involve active implementation on the part of engineers, providing routes to accountability through normative agreements on *what* to translate and *how* to implement it. Norms and standardised tools can be used for internal audit and forms of self-certification.

Section 1. Main challenges

The vigorous societal debates about the importance of ethical considerations in the design and development of technologies in general, and AI systems in particular, have resulted in a multitude of guidelines and principles intended to provide a foundation for better technologies of the future. In some cases, such documents have an impact through influencing political discourse, addressing social problems, and identifying policy solutions[2]. Most of the documents, however, lack clear directions on how these principles are to be achieved and what might happen if some were achieved and not others. **Most importantly, there is a lack of practical insight and guidance for how ethical principles may need to be implemented to achieve desired outcomes.** After all, principles alone are not enough to guide decision-making given the complexities of technology design and development in practice[3]. Where documents such as the IEEE's Ethically Aligned Design[4] or the EU HLEG Ethical Guidelines for Trustworthy AI[5] bring up important points of concern, they also place a fair amount of responsibility for how AI systems are designed and developed on engineers. Recent efforts to define frameworks for ethical AI systems have focused on new standards and regulations, as well as on seeking approaches that move beyond the reliance on legal compliance. Yet **critical gaps remain between principles, guidelines, and recommendations for addressing ethical issues in the development of AI systems and the realities of engineering practice.**

Engineers today hold many responsibilities in their role as developers of new technologies. As anxieties about how technological decisions might play out in society increase, engineers are expected to take into account broader societal contexts, thus going beyond the traditions of requirement specifications and technical development. The difficulty in moving beyond engineering traditions is rooted in the ***translation gap*** between societal conceptions of fairness, dignity, ethics, and the clearly defined terms of engineering practice. The efforts to develop frameworks for responsible, ethical, and trustworthy AI 'by design' have resulted in recently released standards and guidelines attempting to bridge the translation gap. Yet there remain few proven methods of translating principles into practice where AI development is concerned. There is an ***implementation gap*** between the ideal of achieving ethics or privacy 'by design' and the realities of how these should be implemented in practice in the development of AI technologies. Finally, as engineers design technologies, they must navigate the ***accountability gap***[6] – the lack of clarity about who ought to be accountable for the outcomes of technology use, to whom, and how.

2 Schiff et al, "What's next for AI Ethics, Policy, and Governance? A Global Overview."

3 Mittelstadt, 2019 "Principles alone cannot guarantee ethical AI."

4 https://ethicsinaction.ieee.org

5 https://ec.europa.eu/futurium/en/ai-alliance-consultation/guidelines/1

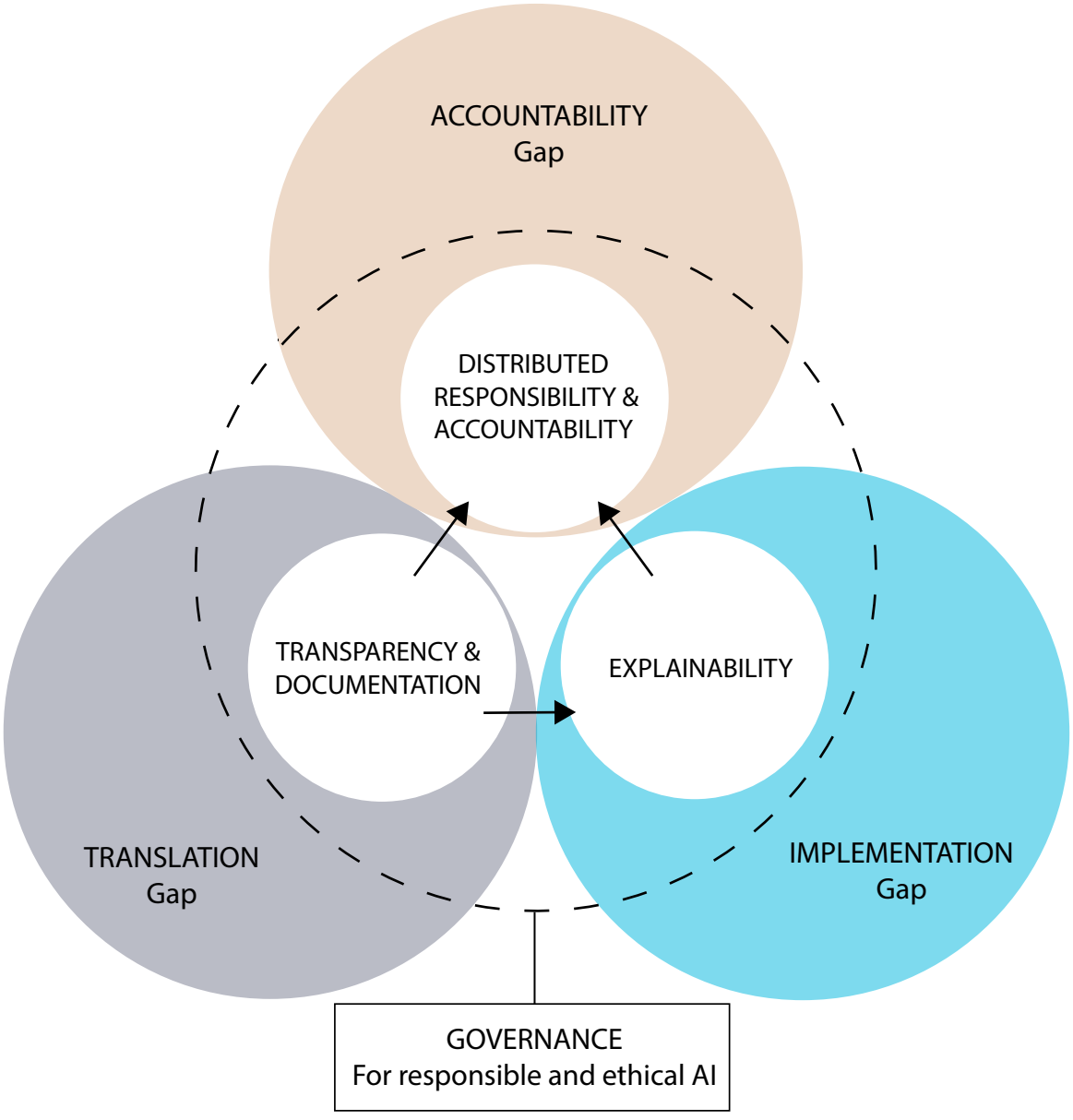
6 Mittelstadt et al., 2016 "The Ethics of Algorithms: Mapping the Debate."

The findings from the Hackathon point clearly and emphatically to an overall lack of clarity of how to address the gaps between the ethical principles and the current engineering practices in AI system development and implementation.

When faced with the uncertainties of how to recognise and navigate ethical issues and challenges, engineers in the hackathon looked to identify the responsibilities that need to be in place to sustain trust and to hold the relevant parties to account for their misdeeds. The hackathon itself offered a space for exploring existing challenges through discussion. **Engineers identified four critical areas of concern they felt need to be addressed to bridge the gaps between ethical principles and the practice of AI system design and development (see Figure 1).**

FIGURE 1.

Bridging the gaps between ethical principles and practice through critical areas of concern.



Drawing on the four critical areas of concern, the participants pointed to what we have termed *nodes of certainty*, areas of intervention that can provide direction for further development of AI ethics as part of seeking solutions. These are described in detail in Section 2. The overall approach and methodology of the hackathon can be found in Section 3 of this report.

1.2 FOUR CRITICAL AREAS OF CONCERN FOR ENGINEERS

Although they are relatively common terms nowadays, neither AI nor ethics have easy and straightforward definitions. During the hackathon participants in general agreed with the sentiment that AI systems *"just do pattern recognition and pattern following … because there is no real-world understanding of anything"* (Ansgar Koene, EY, University of Nottingham, and EMLS RI)[7]. They placed responsibility for societal outcomes not in the algorithms themselves but in the human decisions about which algorithms to use and for what purpose.

Some confusion existed around the term ethics because it was not clear which frameworks and principles ought to be used to contextualise the choices engineers must make as they develop AI systems and how to use them. Ethical reasoning requires reflection, discussion, and deliberation. The hackathon offered space and a framework for such active discussion and deliberation, enabling the participating engineers to engage in the practice of ethical reasoning. As such, they debated the necessary obligations that engineers must take on to address ethical concerns in AI, focusing on four critical areas: (1) transparency and documentation, (2) the challenge of explainability, (3) responsibility and accountability in AI development, and (4) governance for responsible and ethical AI. It is important to note that these four areas of concern are intimately interconnected: documentation is a key enabling determinant of transparency, which is necessary to achieve explainability. Transparency and explainability together form the primary mechanisms for the performance of responsibility and accountability. Governance represents an overarching framework, which determines who must be responsible for what, accountable to whom and in which ways.

1.2.1. TRANSPARENCY AND DOCUMENTATION

Transparency is perhaps the most frequent ethical principle featured in the current sources on AI ethics, although there are differences in how it is interpreted, justified, applied, and expected to be achieved. Transparency can mean different forms of disclosure of data, decision making processes or other forms of communication. While some see transparency as a means of limiting harm, others see it as a route to engendering societal trust. Debates about issues of fairness and ethics in AI systems often seek to assuage concerns by seeking to make various aspects of AI systems and the processes of their development visible and transparent. In this way, the principle of transparency is often proposed as a means of addressing the **translation gap**. However, recommendations of what and to what extent should be made transparent, legible or visible vary. Calls for greater transparency are typically linked with efforts to improve communication and disclosure of how AI systems work, often with the intention to foster trust, minimise harm, and underline benefit for legal reasons[8].

7 Throughout the document all quotes have been approved by the participants. All attributions follow participant requests.

8 Jobin et al., "The Global Landscape of AI Ethics Guidelines."

During the hackathon**, a consensus emerged among the engineers that transparency is an important goal. In practice, the process of making transparent AI technologies and the design decisions taken in their development was seen as dependent on a range of engineering documentation practices:**

> ***"The answer must be documentation, for transparency. It's well-documented systems that reveal the thoughts of the engineers and the developers who have created the AI system"***
>
> **– Rune Krøvel-Velle, Microsoft Certified Systems Engineer/Microsoft Certified Solutions Expert**

Technical documentation can include policy and technical decisions being made in system development, descriptions of system functionality and architecture, of data sources and flows, as well as of which information will necessarily be delivered to the end-user. Such technical documentation is a traditional part of the engineering work process but often varies in quality and the amount of detail provided. Careful consideration of the type and extent of information that is documented is key to a well-functioning engineering design:

> ***"... documentation is an engineering responsibility for how information needs to flow, as frameworks are driving information to the engineer from the corporation and from stakeholders. Documentation is a way for information to flow from the engineer back into those systems, and that requires the engineer to be involved in testing and validation to assure that the engineering designs are actually functioning effectively"***
>
> **– Randy Soper**

Traditional technical documentation is usually highly formalised and abbreviated, intended for a specified and technical audience. Such documentation is less useful for intentions of fostering trust through transparency and providing a means for answering questions about fairness and ethics. Yet, as one participant commented to the general agreement: ***"when it comes to documentation, I think we have to be clear that there are multiple needs there"*** (Randy Soper). Documentation is key for different kinds of audit and for supporting operation at scale. Thus, there need to be facilities for creating both narrative documentation of process and decision-making for design and engineering choices as well as appropriate meta-data to support development operation processes or automation of controls. There was ample agreement that engineer responsibilities for documentation must go beyond traditions of technical documentation:

> ***"It's well-documented systems that reveal the thoughts of the engineer and the developers who have created the AI system"***
>
> **– Rune Krøvel-Velle, Microsoft Certified Systems Engineer/Microsoft Certified Solutions Expert**

Such documented systems are necessary to provide context for engineering choices that might impact the effectiveness of the system and its capabilities. Engineers debated the extent of such documentation – how much information should be provided to the user and the other stakeholders, what kinds of information, and in what format. This expansion of the traditional purpose and practice of documentation is something that needs to be institutionalised to be effective. In the plenum discussions of the hackathon, the engineers readily acknowledged that they must make choices in the development of AI technologies and that these choices can have ethical implications. Documentation of these choices was the acknowledged responsibility of the individual engineer, but organisational frameworks needed to contextualise these choices were seen to be missing.

Thus, **for documentation to become a key and effective mode of enabling forms of transparency in the development of AI, engineers must shoulder greater documentation responsibilities. These new responsibilities require new organisational frameworks that can support new practices.** It is important to acknowledge that traditional technical documentation is typically time-consuming, and its expansion would potentially require additional training as well as adjustments for already tight system development timelines.

Technical documentation serves many purposes, offering a means for external review as well as a mode of internal communication between different engineering teams working on the same systems. One such aspect is documenting the purposes of particular system functionality. Given the ever-present concerns of the dual use of technology, where for example an AI system may be used for both support of its users and their surveillance, engineers in the hackathon wanted to go beyond traditional documentation practices. The questions they debated were how to ensure a common goal towards the purposes of a particular AI development, and how to ensure that it will not be used for other purposes. To achieve this, documentation of decisions needs to include the expected contexts in which the technology is to be used. For instance, a robot used specifically for elderly care may be deemed unsuitable for childcare. The problem thus lies in the misapplication of technologies along with the purposes they had been programmed to serve.

The variable context of transparency decisions is often grounded in the technologies' practical applications and participants expressed a need to document the choices for engineer intentions. Documentation could be used not only to document individual development choices but also as a collective, written agreement among co-workers to propagate awareness of the project goals and purposes throughout all levels of development. In this way, although individual engineers would continue to be held accountable for their decisions, their responsibilities can be premised on a set of collective agreements that necessitate cooperation.

Our participants saw documentation as a key mechanism to achieve transparency where the development of AI systems is concerned. Expansion of purposes and potential audiences for documentation would necessarily demand greater time commitments. Time constraints accompany the burden of many responsibilities that engineers bear and could limit the development of necessary documentation. In addition, commercial secrecy and competitive concerns limit certain types of documentation release. Finally, certain types of algorithmic systems function through opaque mechanisms that can only be documented to a limited extent[9].

9 Ananny and Crawford, "Seeing without Knowing."

Achieving transparency through documentation requires a clear process of decision-making about what to document, how, and when. This means going beyond the accepted traditions of technical documentation and including documentation of decision-making practices at different levels in the organisation. Such a process must also determine who is responsible both for the provision of documentation and for ensuring some basis for holding relevant parties to account should things go wrong. However, there are no clear routes for how to define such processes, the demands and limits of responsibility, or how accountability might be achieved and by whom.

1.2.2. THE CHALLENGE OF EXPLAINABILITY

While documentation is one way of achieving transparency by making information available, the mere availability of documentation may not be enough to understand the logic of AI systems. The European Commission High-Level Expert Group on Artificial Intelligence states that "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."[10] The challenge lies in determining what is a suitable level of explanation of algorithmic processes involved in various forms of automated decision-making. How much should be explained, to whom, and in what way?

Hackathon participants agreed that explanation is ***"an obligation, a challenge, and a design choice"*** (Randy Soper). It is an obligation because explaining is the most important job of the engineer, making sure that all the relevant stakeholders can understand what the capacities and limits of any system are. It is a challenge because it is impossible to explain everything and it is about choosing what to explain and how. It is a design choice because explainability is more than a question of an understandable narrative in some form of documentation, it is a question of what must be implemented and how to ensure the possibility of explanation. There is substantial agreement that explainability is an important goal. Yet despite the volume of effort to develop different methods for making systems legible to different stakeholders an **implementation gap** remains between the principles and guidelines and the development of AI systems in practice. Given the complexity of AI systems, complete explainability is not possible, but how much is enough remains a question.

> ***"This is a standard engineering problem. A trade-off: How AI functions is often inherently unknowable and usually a trade secret. At one extreme, 100% explainable AI means no machine learning, you can't have its benefits. The other extreme is zero explainability. The system does something, accomplishes the goal that it was given, with many possible side effects, and we have no clue how it does it. That is obviously extremely dangerous. So we are looking for a middle ground. For example you might want it to output interim results and perform constant public tests. But it's an annoying situation because you can't have a universally correct solution to all AI."***
>
> **– Erik Mowinckel**

10 High-level Expert Group on Artificial Intelligence, "Ethics Guidelines for Trustworthy AI."

Achieving the right amount of explainability is an attempt to find a balance between two extremes of 'no explainability' and 'too much explainability'. Too much comes at the expense of efficiency and precision of AI systems, commercial considerations and competitive advantage, or even leaving the system too open for malicious attacks, while too little may leave the stakeholders involved unable to trust the technologies. Similar to the questions of documentation, what to explain, how and why are the important questions.

> ***"You don't actually need to know exactly the inner workings of [an algorithm]. Explainability is not about revealing the source code. It's about trying to give an explanation of the mechanism ... because it is more interesting to understand, why did I get this specific response."***
>
> **– Mikael Anneroth, Ericsson Research**

The engineers agreed that not everything in an AI system can be explainable and discussed alternative approaches to what might be explained and how. Two primary proposals for achieving forms of explainability, despite the inherent opaqueness of many AI systems, centred on human-in-the-loop and testing solutions. The idea of human-in-the-loop approaches to AI system development received broad support during the hackathon. Human-in-the-loop was seen as a necessary and obvious way to navigate the problems of potential bias and unfairness in automated decision-making outcomes.

> ***"It's a very good solution to have a human expert between an AI and its output. A human expert tends to have high specificity (will not make a false positive), but not great sensitivity (may miss a true positive). Machine learning algorithms tend to have very high sensitivity, but somewhat lack specificity. Thus, an expert will catch exactly the kinds of errors a machine learning algorithm tends to make, while still getting the benefits of its high sensitivity. This also avoids a bunch of ethics problems, because you now have a human expert making the final decision, and who is responsible for the outcomes. This can be hard to do with a machine learning algorithm, because if seven teams worked on it at different stages and it makes mistakes, whose fault is that?"***
>
> **– Erik Mowinckel**

A lot of discussions distinguished between decisions that may seriously impact human lives and those where errors are unlikely to cause significant harms. In the struggle to determine how to navigate potential ethical quandaries of algorithmic support for difficult decisions such as medical care, dispersal of public benefits, job hires or admission to educational institutions, engineers were comfortable defaulting to human involvement. Once potentially sensitive decision-making instances are identified, trying to implement conditions for a human-in-the-loop was regarded as a way to address the whole swath of ethical concerns, shortcutting the challenges of explainability. Algorithmic systems have made the costs of predictions based on data very small, but predictions are not very

useful without judgement for when they are relevant[11]. Such judgement can often require the kind of knowledge and awareness that is not possible to include into the models underlying AI systems and thus require human involvement, making judgement the costly part of this enterprise.

Alongside the potential of human biases and rising costs of human-in-the-loop solutions, there is another problem. One of the advantages of algorithmic systems is speed and scale – the ability to process large amounts of data and to produce decisions quickly. This makes it difficult to put a human-in-the-loop for every instance of decision-making no matter how sensitive or important these might be. One proposal to address this problem was through developing new system testing approaches. The discussion of testing focused on the idea that current system development testing is insufficient and on the need for more extensive in-the-wild testing scenarios to be conducted prior to full system launch. This was seen as especially important for systems that utilise algorithmic approaches that are more difficult to explain, for example, neural networks.

> ***"With many systems, no one, including the designers of the system can guarantee it will do a specific thing. So, you have to test before system release. Your test needs to be public before it is released. There is a lot of economic potential from fiddling with the system."***
>
> **– Paul Knutson, NITO Studentene**

Of course, what and how to test needs to be specified and codified further. Software developed for different purposes must typically comply with a range of different standards, some of which require performance reviews and benchmarking prior to system release. New forms of system testing for explainability purposes could be integrated into existing software development and release processes. However, it is also clear that no set of formally devised tests will ever be able to envision all the potential contexts and problems that might emerge over time as AI systems are implemented.

Despite the reliance on a human-in-the-loop as a better bet for difficult decisions, engineers acknowledged the simple fact that all humans are biased in different ways. Human decision-making can also be opaque not only to outsiders but also to the decision-makers themselves, as we struggle to come up with post hoc explanations for our actions. We expect clearer logic for the systems we build, however, reliant as these are on our specifications. Yet we must be aware that how we choose to present explanations, whether using the human-in-the-loop arrangement for rendering final judgment or making decisions about how a system might explain itself to its stakeholders, may be biased or wrong.

While the discussions of explainability generally focused on explaining the logic and reasoning behind a particular decision produced by the AI system, some engineers also brought up a more human-centred view of explainability as an approach to communicating decisions taken in the development of an AI system and of its capacity to explain itself.

11 Agrawal et al., "How AI Will Change the Way We Make Decisions".

"Explainability should not only focus on explaining the algorithm but also on explaining the thought processes of the people that decided what the algorithm should be doing."

– Ansgar Koene, EY, University of Nottingham, and EMLS RI

Explainability is a necessary part of transparency in that it implies an action – the transformation of opaque processes into something intelligible – rendering certain forms of transparency possible and, in turn, contributing to traceability, auditability, and ultimately accountability.[12] Decisions about transparency and explainability are context-dependent. For instance, an algorithm used for a specific purpose in one domain (i.e., elderly healthcare in a hospital) may be unsuitable in another domain or for another purpose (i.e. grade assessment in education), not only because of the shift in context, but also because the available explanations for its function are insufficient or irrelevant to the new context of implementation. Taking into account different contexts necessitates ongoing maintenance of relationships between stakeholders and organisations or industries, requiring engagement and dialogue to accompany regulations13. To enable this engagement, participants suggested that appropriate guidance for responsibilities and accountabilities needs to be in place.

1.2.3. RESPONSIBILITY AND ACCOUNTABILITY IN AI DEVELOPMENT

Questions of responsibility and accountability are central to the development and implementation of any autonomous systems, whether or not they interface directly with people or deal with personal data. Yet it is rarely clear who ought to be responsible for what and in which contexts. Similar questions arise in considerations of accountability: who or what can be held accountable when the technologies are put into the world and either fail or function mostly as intended but with unexpected side effects? What kinds of accountability need to be instituted and how might such accountabilities need to be governed? Our hackathon participants identified the issues of accountability starkly. As participants presented outcomes of group discussion, one participant representing a group summarised their questions as follows:

How do we actually establish that accountability? Is it distributing responsibility across the company, establishing management control frameworks? Who has individual responsibilities? Where does the liability lie? What kinds of certifications are necessary? How do you provide incentives that align to accountability, and then roles, and how can engineers play multiple roles and how are those complicated in AI because of the multidisciplinary nature of teams?

– Randy Soper summarising group discussion

12 Beaudouin et al., "Flexible and Context-Specific AI Explainability."

13 Aitken et al., "Establishing a Social Licence for Financial Technology."

This kind of questioning points to the problem of the **accountability gap**, noting that despite the willingness of engineers to perform their responsibilities and to be held to account – neither the responsibilities nor the frameworks for accountability with respect to AI systems are well defined.

Our participants saw responsibility as an integral part of the engineering profession but acknowledged that it is not an easy burden. As philosopher Judith Simon notes, the terms responsibility and accountability should be thought of in terms of what matters and what does not matter, what is known, what can be known and what is not.[14] AI systems themselves reconfigure accountability by challenging and changing this dynamic. Given the range of demands, AI development is typically sustained through a collective and often multidisciplinary effort. As such, many engineers brought up the issue of distributed responsibility, pointing to a need for a cooperative governance framework. The discussions spanned from governance on a political level to governance on a more individual level. Engineers situated practices within their immediate contexts but also challenged the responsibilities of actors in governmental, municipal and corporate contexts and the effect they may have on the more local-level decision-making within engineers' workplaces. Within more localised spheres of discussion, **engineers expressed the burden of many responsibilities on their shoulders.** To address this issue, we asked engineers, "who should be held accountable for (various ethics issues)?", and there seemed to be a consensus for distributed forms of accountability and responsibility:

> ***"... the challenge, of course, is that AI is kind of a blunt tool and gets you into problems where algorithms are not effectively being nuanced in the way it's delivering refined control over content management. But human controls are too slow ... And so that's the kind of balance that we need to think about in terms of how we distribute responsibility within an organisation"***
>
> **– Randy Soper summarising group discussion**

By answering questions about responsibility and accountability, participants were prompted to address the accountability gap by debating who or what can be held accountable when technologies are put into the world and either fail or function mostly as intended but with unexpected side effects.

Demands for forms of distributed accountability and responsibility recognise the importance of ethical concerns but arise from a degree of resistance to individualised responsibility. The increasing individualisation of responsibility can come to be seen untenable especially because many engineers felt it was difficult to identify what constituted an ethical issue and to know how to respond. They agreed in principle that:

> ***"engineers must be willing and able to take a big picture view of the task they are being asked to solve and know how to do problem identification discussions"***
>
> **– Ansgar Koene, EY, University of Nottingham, and EMLS RI**

14 Simon, "Distributed Epistemic Responsibility in a Hyperconnected Era."

However, many did not feel equipped or empowered to do this, acknowledging the pressures of corporate power dynamics[15]. Distributing responsibility successfully was said to require not only specific actors to ensure best practices (i.e., the use of standards and regulations) but also to necessitate a no-blame culture in which taking on responsibility does not lead to punitive consequences if unexpected ethical issues arise. In a 'no-blame' approach, there is a need to aim for collective responsibility, something that is prevalent in the codes of conduct in the Nordic countries. How do you know you are right or wrong? How do you know what will result in the wrong outcomes? How can you be certain that following a standard will result in positive outcomes? These questions ought to be answered collectively but it was clear that there is a glaring lack of spaces and opportunities for having collective discussions and debates about ethical issues that can lead to practical outcomes.

> ***"What frameworks do you need as an engineer, because engineering decisions are not ethical decisions, they're engineering choices in design. So, what you need are some sort of informative frameworks that can help to make those choices in a way that can align those design decisions with the greater organisational position of ethics."***
>
> **– Randy Soper**

In Silicon Valley, many technology companies have instituted some options for ethical discussions and even introduced a new role for an ethics-specialised employee or an "ethics owner." These employees are charged with providing a kind of check on potential ethical issues and a route for engineering teams to question check-in on potentially problematic design options. Ethics owners make efforts to implement ethics within companies, but often do not benefit from existing practices to guide their actions. This is because **ethics often continues to be seen as something that needs to be implemented more or less ad hoc rather than something to design organisations around.**[16]

1.2.4. GOVERNANCE FOR RESPONSIBLE AND ETHICAL AI

Throughout the hackathon, discussions frequently converged on the idea that addressing ethical issues requires the development of new modes of governance. This is necessary to support ethical reasoning through clarifying responsibilities, creating appropriate accountability frameworks, demanding new and more extensive forms of documentation, and implementing processes to ensure explainability. This became especially important as conversations focused on more collective mechanisms of control over decision-making throughout AI development. Moving from engineering traditions towards transdisciplinary AI development efforts, developing novel implementation practices and producing accountability frameworks to support decision making, all require addressing the challenge of governance for responsible and ethical AI. This challenge incorporates all three gaps – translation, implementation, and accountability – by enabling actionable insights in each. In

15 Hagendorff, "The Ethics of AI Ethics."

16 Metcalf et al., "Owning Ethics."

one discussion participants debated what was necessary to fulfil the responsibilities and obligations of responsible innovation in AI:

> ***What is it that we have to do to ensure that we're fulfilling our responsibilities and obligations? What do we need to have the correct metrics in place? We talked a lot about professionalism, what kinds of things that implies about licensing, about public trust and public safety, the kinds of certifications that are available, and also how that relates to insurance, the need for an audit trail, whether or not our frameworks are sufficiently informative and provide the correct explainability, whether there are standards that are sufficient for the kinds of design that are going on, and whether or not the code of ethics that's available for AI engineers is actually providing the correct framework for them to be held accountable to the designs that they're doing.***
>
> **– Randy Soper summarising group discussion**

Despite extensive discussions in groups and in plenum, none of the questions summarised above had satisfactory answers. Ethical concerns carry the burden of a range of responsibilities that individual engineers may or may not have the capacity to take on. Although participants were aware of the available codes of conduct from professional organisations such as the Association of Computing Machinery (ACM) and the IEEE, these were deemed too vague to really apply to the specific technologies they might find themselves creating. It is important to keep in mind that **engineers operate within hierarchies of power, which define their capacity to carry out tasks and put limits on the technical and organisational choices they can make. Many remain uncertain about what infrastructures of support are available for performing responsibilities and being accountable for ethical concerns in the development of AI systems.**

The hierarchies of power can be further complicated when traditional methods of software development span cultural and national boundaries. If responsibilities and accountabilities around ethical reasoning are not made clear at the outset, these can become even more complicated where cultural and structural differences intervene:

> ***"... We assume that people are working in Europe and that we have a certain culture where each individual has quite a lot of influence over their work situations. Whereas in other cultures, maybe it's more of a hierarchical work culture, and in a more hierarchical work culture, I think there is a greater risk that the individual just assumes that someone else is taking the responsibilities."***
>
> **– Inger Annie Moe**

In this case, being located in Europe was equated to terms of individualised responsibility and at least some measure of control over one's own "work situations". However, engineers do not expect themselves to have full responsibility and accountability when faced with ethical issues. **An appropriate governance framework would allow necessary hierarchical structures with specified responsibilities of actors that engineers could turn to in case of doubt.** The main uncertainty expressed throughout hackathon discussions had to do with a gap in whom to address when ethical issues might arise. Some participants looked to their organisations to provide a means for debate and advice, while others considered whether this may be a role for civil society organisations such as the trade unions.

Hackathon participants did not believe that the algorithms themselves could be judged ethical or held responsible for questionable outcomes. The responsibilities for outcomes were laid firmly at the feet of different stakeholders, with the engineers themselves recognising their position and obligations:

> ***The algorithm itself is never responsible for its outcome. It's always people. It's the management. It's the ones that decided that this algorithm was going to be used. But there is also a responsibility on the side of the engineers, for instance, there is a responsibility to have the right kind of skills when you're actually taking on this kind of job.***
>
> **– Javier Poveda Figueroa, Catalan association for artificial intelligence**

With machine learning toolboxes becoming simpler to use and available to a broader audience, there is less need to understand what these techniques do and what are the assumptions that must be considered. The responsibility remains on the side of engineers for understanding the tools they use and the solutions they create and put forth into the world. This was a reason for looking at professional certifications as a way to ensure that those developing AI solutions understand which are the right tools for a particular purpose and why. Such a position questions the increasing efforts to "democratise" the use of AI through simple tools available to users with expertise in domains.

Yet AI development is no longer the purview of engineering and computer science alone. Participants debated how to know whether someone has the right skills for the job of developing AI and what those skills might be. Many sided with the argument that pure technical expertise is no longer enough and that engineering education itself must change and incorporate broader societal views. Some went so far as to propose special certifications for engineers that might evaluate capacity for ethical reasoning and reflection as well as computer science and engineering training. Yet many recognised that mistakes will continue to be made and there needs to be capacity for discussion of those mistakes, a "no-blame culture" for discussing failures and learning from them.

Participants generally felt that **there is insufficient support for engineers and designers who seek safe spaces for expressing concern, perhaps facilitated and administered by an external party whom they could consult.** We must recognise that despite the best intentions, there are situations where employees will feel that their opinions or ethical concerns cannot be voiced within the power structures of the workplace. This is a paradox where fostering trust in AI systems

may create new problems. There is a **possibility that increasing trust in AI may diminish scrutiny and erode the social obligations of developers and engineers. This challenges the idea that building trust is fundamental to successful, ethical AI governance**.[17] There is an inherent tension between high levels of distrust in AI that can make it impossible to allow AI system development and high levels of trust in AI that can deem data-driven decision-making unassailably trustworthy. Although most people are aware of both the possibilities and the challenges of AI systems, they nevertheless seek some sources of certainty when faced with dilemmas around ethics and AI. The hackathon event consolidated what participants expressed as necessary measures to fill the accountability and governance gaps, or what we refer to as 'nodes of certainty'. The lack of accountability was often seen as connected to the lack of interpersonal discussion, collective support, and clear infrastructures through which individuals' sense of certainty and trust could be cultivated and encouraged.

17 Jobin et al., "The Global Landscape of AI Ethics Guidelines."

Section 2. Spaces for Ethics and Nodes of Certainty

Discussions about ethics in practice are full of uncertainty because it is difficult to predict how technical decisions might play out once AI systems are implemented and used in the world. The translation, implementation, and accountability gaps are difficult to bridge because few obvious and certain practical solutions are available. Ethics in AI is a wicked and multifaceted problem. Drawing on the four critical areas of concern outlined above, hackathon participants elaborated on their worries and presented ideas for routes toward a more supportive governance framework for ethics in AI.

Engineers recognised that finding good ways to engage with and address ethical challenges in AI system development requires going beyond existing and familiar practices. **The repeated and occasionally quite public failures of AI systems demonstrate that business as usual in engineering practice is no longer enough.** Developing AI systems responsibly demands additional skills and expertise for how to make technology development decisions while taking broader societal concerns into account. Finding ways to integrate these additional skills with existing practices requires allocating time for reflection and *spaces for collective discussion and debate*. Such spaces need to offer opportunities to challenge what is familiar and accepted, defining what should be the responsibilities and obligations of engineers, and how these might be fulfilled. Opportunities for generative discussion and debate need to be made available internally in organisations but also externally through trade unions, civil society organisations and government initiatives.

Engineering relies on several traditional practices that create structure and certainty for an enterprise that is always looking to create a new future through technical innovation. These practices, such as reliance on community-defined standards and traditions of technical documentation, offer familiar points of departure to tackle ethical dilemmas in AI and form what we have termed *nodes of certainty*. These are familiar practices that can be re-envisioned and augmented to accommodate the needs of responsible and ethical AI, debated through spaces for discussion and supported through policy initiatives.

The European context and, more specifically the Nordic setting, offers some advantages, starting with recognising that Nordic countries are already digitised (i.e., nationally maintained and centralised health data), have supportive, democratic welfare governments that most citizens trust (i.e., with different consent practices that underlie Nordic countries' data accumulation[18]), and generally a very low level of corruption. These are all qualities of the Nordic countries that provide a stable

18 Tupasela et al., "The Nordic Data Imaginary."

context for reflection and addressing uncertainties of ethical dilemmas in AI. There is a need to further enhance these qualities by stressing the need for a no-blame culture within workplaces so that engineers and designers feel open to discuss their concerns. The role of trade unions and the power of an organised workforce should not be underestimated here. The purpose of the trade unions has traditionally been to protect workers' rights and their working conditions. As AI development, maintenance, and operation become an ever-greater part of engineering practice, addressing questions of ethical and accountable AI development can be seen as an integral part of maintaining appropriate working conditions. Therefore, trade unions have a role to play in creating safe spaces for discussions of challenges in AI development as well as offering potential mechanisms for "pushing back against reprehensible"[19] development of new technology.

2.1 SPACES FOR ETHICS

Providing an infrastructure of support for ethical and responsible AI development demands time and space. Yet how might spaces be provided for ethical deliberations when engineers feel there is not enough time for ethics to begin with? While spaces for ethics were a demand among all participants, on an individual level they did not feel that they had the time to consider that additional responsibility. This calls for external intervention that would help establish necessary spaces for ethics within workplaces, as well as internal attention to these growing needs.

2.1.1 EDUCATION & TRAINING SPACES

Some engineers expressed that there is a need for additional education concerning ethics as part of engineering education in general and as additional training as part of life-long learning initiatives. Engineering AI systems is an inherently interdisciplinary endeavour requiring collaboration with people from many different backgrounds. Yet there was a sense that engineers themselves need to take additional responsibility for identifying potential ethical issues and discussing these.

> ***"I think this is a very important point. Engineers do not receive any kind of education in humanities and social sciences at all. They should be aware of basic concepts. Later on, they will join an interdisciplinary group with experts in other research areas. But at least, engineers should be able to question themselves and their colleagues about basic challenges, which nowadays does not happen. In general, at least from my experience at the university, this is lacking in our education."***
>
> **– Jordi Domingo-Pascual, Universitat Politècnica de Catalunya – Barcelona TECH (UPC)**

Engineering is a profession that comes with many responsibilities and obligations. In some countries, such as the iron ring ceremony in Canada[20], these are taken seriously through commitments

19 https://www.wired.com/story/timnit-gebru-exit-google-exposes-crisis-in-ai/

20 "Iron Ring", 2020

reminiscent of the Hippocratic oath tradition in medicine. As professional organisations such as the ACM and the IEEE redevelop their codes of conduct, these too bring new commitments attempting to provide some basis for dealing with emergent ethical challenges. Nascent efforts to create additions for reflection and ethics in technical education are also being developed although these currently tend to be ad hoc and uneven.[21]

> ***"Engineers must be willing and able to take a big picture view of the task they are being asked to solve, know how to identify problems and discuss these."***
>
> **– Ansgar Koene, EY, University of Nottingham, and EMLS RI**

Engineers also discussed a need for educational and training spaces that would allow working engineers to develop the necessary qualifications to engage with broader concerns in practice, to support the development of new forms of quality technical documentation, and to provide proof of competence in AI development.

2.1.2 SPACES FOR DISCUSSION

Where education is a structured process of developing new competencies, concerns about the lack of opportunities to discuss and debate what constitutes ethical issues in engineer practice were common. These concerns echoed similar discussions presented in the Nordic Engineer Stand on Ethics report, produced by the ANE in 2018[22].

Such spaces for discussion were imagined on three levels – within workplaces, as interdisciplinary events organised by external organisations, and political debates that could include an engineering point of view. At all levels, such spaces for discussion are ways of addressing the translation gap – an opportunity to debate and perhaps in time agree upon how to identify ethical issues and how to address these. Dealing with ethical dilemmas is about making decisions about trade-offs in an environment of high uncertainty. Such decisions need to be made collectively, effectively distributing responsibility for potential mistakes, while creating opportunities to learn from these mistakes going forward. Many engineers expressed concern about discussion spaces in their workplaces because they did not feel that there are sufficient incentives for all sectors to become aware of the goals behind the technology production (insufficient transparency). Neither are there sufficient means of communication to enable these incentives.

While some of these discussions must happen internally within companies, there is also a place for external organisations, such as trade unions, professional or civil society organisations to provide such opportunities as well. Engineers were looking to external sources of expertise that might help manage uncertainties around what constitutes an ethical issue in AI.

---

21 Fiesler et al. "What Do We Teach When We Teach Tech Ethics?"

22 "Nordic Engineers' Stand on Artificial Intelligence and Ethics Report". ANE, 2018.

> ***"The engineers are encountering dilemmas or issues every day, there is a lot of responsibility on the shoulders. So, we have to provide support through discussion in organisations and between the engineers. Because these are not questions that one person can give the right answer to. No. So discussions are very important."***
>
> **– Inger Annie Moe**

Interdisciplinary discussions organised internally or externally could, for example, support the previously addressed education training spaces and enable communication on matters of AI ethics across disciplines. Working across disciplines, engineers and others can discuss domain-specific contexts that are necessary to address.

On the political level, participants proposed that politicians and engineers should have iterative discussions on AI ethics. They stressed that the European voice on matters concerning ethics standards in AI development is paramount and should be positioned to provide a structure of support to enhance public trust[23]. **Implementing the three levels of spaces for discussion – internally within organisations, externally through independent organisations and at the political level, would enable necessary action to tackle ethical concerns.**

2.1.3 SPACES FOR EXPRESSIONS OF CONCERN

Since engineers work within the structures of hierarchy in their workplaces, they are constrained in what they can say and to whom due to commercial secrecy or non-disclosure agreements for example. In case they were to foresee, witness or suspect undesired ethical consequences, they often do not have the necessary support to voice these concerns. Participants discussed a need for neutral spaces in which independent, external bodies (either individuals or a group) could be approached for advice on ethical matters.

> ***"We thought that it would be good to have some sort of independent body or a person or organisation to whom you could report or with whom you could discuss these ethical concerns that you're having"***
>
> **– Mikhail Takmakov**

These bodies would need to be outside of and unrelated to the workplace to establish a 'safe space' for concern. Such spaces for concern can thus set expectations from outside the organisation and, in this way, perform accountability through communication governance.[24] In other words, expectations need to be strengthened to effectively address the accountability gap. The need to voice concern also called for 'whistleblower protection' or a supportive body to ensure that engineers

23 Aitken et al., "Establishing a Social Licence for Financial Technology."

24 Kerr et al., "Expectations of artificial intelligence and the performativity of ethics."

who voice their concerns do not get punished. Throughout the discussion, several participants suggested having an ethics protection officer in place.

> ***"... if you do speak up, it would be nice that the legal system could ensure that you don't lose your job, for example, or indeed, in an extreme case, you don't need to change your country... we had this idea about an 'ethics officer' in a company for example"***
>
> **– Teemu Viljamäki**

All groups mentioned a need for whistleblower protection, facilitated by mechanisms to allow engineers to flag ethical issues. If such issues are made public, the so-called whistleblower must be protected. To ensure such a mechanism, the role must be enabled from outside the organisations, through governmental measures, to safeguard control of the process in favour of the whistleblower in question.

Since 2014, ANE has been advocating for whistleblower protection on a European-wide level seeking to put in place horizontal and cross-sectoral legislation to protect professionals who expose company wrongdoings. The advocacy efforts carried out in cooperation with European trade union organisations representing professionals and managers, culminated in October 2019, when the EU Directive[25] on whistleblower protection was finally adopted. The EU Member States have until December 2021 to transpose it into their national laws.

Trade unions have a crucial role to play in establishing the culture and measures that promote whistleblowing. Currently, the members of the Danish Society of Engineers (ANE member-organisation) can benefit from legal advice and economic compensation provided by the trade union in case they are laid off following a decision to whistleblow. At European level, the whistleblowing toolkit[26] has been developed by Eurocadres, the European trade union organisation representing professionals and managers.

2.2 NODES OF CERTAINTY

Where spaces for discussion offer a means of making sense of emergent challenges and a route to learning how to identify what constitutes an ethical issue in AI, nodes of certainty are familiar points of departure for how to address ethical issues once identified. The infrastructural nodes of certainty presented below are based on what engineers discussed as necessary to fulfil the responsibilities that are yet to be put in place. Suggestions broadly included documentation, certification, oversight, and punitive measures, and pointed to the specific demands for governance and particular responsibilities.

25 https://ec.europa.eu/info/aid-development-cooperation-fundamental-rights/your-rights-eu/whistleblowers-protection_en

26 https://www.eurocadres.eu/publications/whistleblowing-toolkit-eurocadres-best-practice-guide/

2.2.1 DOCUMENTATION

A large part of individual engineers' own responsibility, documentation allows for the flow of information about the development process others to enable others to replicate development, understand the development process, or understand the purpose and step-by-step use of the technology. One participant suggested that documentation should be reviewable and auditable "a year or two down the road". There are different types of documentation and "narrative" documentation was brought up, signalling that it may also be appropriate for metadata, for example, to support development operation processes or automation of controls. Documentation can have many stakeholders from other developers in teams with different tasks in the AI system production to the end-users. However, the purpose of documentation is similar throughout even if the form may be different depending on the audience:

> ***So they understand contextually why you made a choice and how that impacts the effectiveness of the use of the capability, whether that's control around us, or whether that's providing information to the user at an appropriate time or what have you. All that is an aspect of documentation.***
>
> **– Randy Soper**

Understanding why the developer made a particular choice and how it impacts the effectiveness of use and the capability, and thus providing explainability, could help prevent inappropriate use. In this way, documentation could become an interactive process and lead to greater transparency. Traditional technical documentation approaches need to be re-thought and perhaps re-imagined to include a broader set of audiences and to enable documentation to become a necessary tool for achieving the needed levels of transparency and explainability for an AI system.

Documentation requires initiatives to translate internal processes into feasible accounts for both developers and users, but also to implement ethics throughout the development process. As documentation is a time-consuming process, there have been attempts to automate some of it, which may be a problematic direction if ethical reviews and audit trails come to hinge on documentation of decisions as well as technical content.

2.2.2 TESTING

Although human-in-the-loop approaches might seem like an ideal solution for an embattled AI decision-support system, the sheer scale of automated decision-making processes renders it impossible. At the same time, engineers and developers themselves cannot foresee the effects of the systems they develop. As a result, the development process often relies on extensive testing for effects before system deployment. However, testing does not involve all the variables and changing contexts of the real world – so how do we ensure that testing is done in the best way possible? The challenge lies in explaining the effects of AI systems in use and forging a route to accountability. Thus, despite the excitement around testing systems pre-deployment to guard against potential ethical concerns,

testing is an approach that has its own problems. Ali Hessami, a guest presenter at our Hackathon event, suggested that testing should be augmented by a form of real-time monitoring and oversight:

> ***"What we really need ... is real-time validation and monitoring rather than testing because testing itself requires a whole host of scenarios and cases that may not be representative or relevant to an adaptive system. What you really need is some form of smart contract style monitoring of a complex AI product in real-time, rather than at periodic phases, or in the old-fashioned style of going for some form of audit every three months or six months or one year. We are dealing with adaptive, non-deterministic systems. Testing is not going to be an adequate measure of assurance though this is obviously highly desirable if you didn't have anything else. But in preference to periodic testing, we really do need some form of predictive oversight, and real-time monitoring."***
>
> **– Ali Hessami, Vice Chair & Process Architect, IEEE Ethics Certification Programme for Autonomous & Intelligent Systems**

Testing is typically done prior to deployment but given the fact that AI systems by design learn and change over time given a particular context of implementation, testing processes need to be rethought. At the very least testing ought to be done iteratively and systematically throughout the system lifetime. Whether via iterative testing or real-time monitoring, solutions must be able to deal with changing social and technological contexts as well as account for system adaptability.

2.2.3 STANDARDS

Traditionally, standardisation deals with technical issues, such as quality, interoperability, safety or security. Discussions about standards mostly pointed to the fact that engineers expect standards to be available especially in areas where outcomes can have significant negative side effects. To help organisations apply abstract AI ethics principles to concrete practices, the IEEE Standards Association has been developing both technical and socio-technical standards[27]. The standards focus on things like process frameworks for incorporating values into innovation and engineering projects, defining different levels of transparency for incremental needs, data governance, age-appropriate design[28] and impact assessment of AI systems[29] on human well-being and the environment.

> ***"I think engineers need to talk about ethics. You also need to orient on some ethics standards when we solve problems."***
>
> **– Yaroslav Gosudarkin**

27 https://ethicsinaction.ieee.org/p7000/

28 https://standards.ieee.org/project/2089.html

29 https://www.techstreet.com/ieee/standards/ieee-7010-2020

Throughout the hackathon, engineers noted that standards on trustworthy AI are being developed both nationally and internationally, although it was not yet clear how useful these standards might be. Technical and socio-technical standards and certifications, developed through an open and transparent process, can establish a means for all stakeholders involved to conform with agreed-upon norms and principles. Such standards and certifications could serve as reliable "elements of certainty" and important governance instruments for regulators, industry, and the ordinary citizens, but most importantly for engineers themselves as they seek some certainty in technical decision-making.

2.2.4 CERTIFICATION

Alongside standards, certification was a common concern. Certification broadly involved discussions on licensing, and professionalism, as well as testing and potentially even labelling companies as 'ethically aligned' although the latter was hotly contested because such a label cannot account for changes in company practices over time. Certification is a common practice in engineering, with various programs and certifications available for individual skills and development processes alike. Thus, it is no surprise that professional certification became a big point of discussion.

> ***"... such kind of certification might come from professional organisations, or from the employing organisations, especially if we're talking about larger employment bodies, such as the public sector. There is a need to establish what is the proper kind of certification and this requires public consultation, because the affected stakeholders, in this case, are the public. Such a public consultation would address what kind of certification on ethics would be necessary."***
>
> **– Ansgar Koene, EY, University of Nottingham, and EMLS RI**

Certification of skills in particular domains, such as information security, maintenance or reliability – are common forms of extending areas of expertise through life-long learning opportunities of different kinds. Even though there was considerable agreement that ethical considerations, challenges and concerns with respect to AI systems are important and must be addressed, many engineers felt they lacked the necessary expertise to do so well. The Hackathon participants were a self-selected group of people already deeply interested in ethics in AI willing to spend two days on these discussions. Their hesitation was indicative of a broader sense of anxiety and inadequacy among engineers in translating the abstract principles of ethical AI into practice. Obtaining certification for ethical AI would signal professional capacity and training to address ethical challenges in AI system development, providing a sense of certainty both to those certified and those that could benefit from this extra capacity of their colleagues.

In addition to a certification of skills, it is also important to address the certification of products and services to assure expected and acceptable system behaviour throughout the AI system's entire lifecycle. With actionable assurance criteria, conformity can be certified through audits performed by assessors that can also include methodologies to provide real-time monitoring capabilities in the

future. For example, the IEEE's Ethics Certification Program for Autonomous and Intelligent Systems (ECPAIS) program[30] has developed a set of criteria for this purpose. When launched, ECPAIS certification marks will communicate to consumers, business partners, and public bodies how AI systems have been audited, providing specifics on safety and increasing trust.

Certification and licensing provide one mechanism for bridging the translation gap and the implementation gap by formally prescribing processes to achieve a desired outcome through standards. In this way standards and certifications can provide the capacity to implement the abstract principles of ethical AI in practice. They can also help in addressing the governance challenges by establishing the necessary responsibilities and accountabilities.

2.2.5 OVERSIGHT

Oversight is needed to make sure that the ethical standards and regulations already in place are put into use. Participants discussed the need for oversight on a governmental level and internally within workplaces. Some suggested the use of AI systems to provide real-time oversight and validation of development processes with monitoring that could potentially be used for real-time testing. However, administrative and automated oversight require different responsibilities and forms of governance.

Administrative Oversight

Much of the discussion on governmental oversight hinged on punitive measures and ensuring that there were consequences for clearly unethical decisions. Many engineers commented that commercial concerns would typically trump ethical considerations in industry situations and governmental pressure and oversight were some of the ways to force companies to spend time and resources on ethical deliberation. At the same time, external oversight is a form of shifting responsibility for paying attention to emergent ethical challenges and it demands clear accountability structures for when things go wrong.[31] These mitigation measures are procedures that are taken to reduce risk, preceded by necessary impact assessments that identify what may go wrong.

There are several existing mechanisms for external oversight processes that currently address some aspects of ethical AI, for example, compliance with GDPR. The tradition of different impact assessments points to a possibility for a broadly conceived ethical impact assessment process, which would demand clear commitment from the companies and organisations developing AI systems. However, some forms of administrative oversight also require a depth of technical expertise, which may be difficult to achieve given the rapid development of the state of the art in the area of AI systems.

Automated Oversight

Risk assessments, testing, auditing and creating documentation are all part of the responsibilities and obligations of engineering. These tasks, however, are time-consuming and do not directly con-

30 https://standards.ieee.org/industry-connections/ecpais.html

31 Beaudouin et al., "Flexible and Context-Specific AI Explainability."

tribute to system development. As is common with tasks that are perceived as uninteresting busy work, oversight is ripe for automation, especially given the fact that AI systems continue to change and evolve post system deployment and thus require longer-term involvement and maintenance solutions that go beyond merely asking whether the system "works". However, fully automated oversight solutions would of course result in the same challenges they are intended to solve by having to address potential ethical challenges.

For any AI system, emergent ethical issues are often unexpected, which is something an automated system would have a hard time identifying, as revealed by recent studies on automated auditing[32]. As such, rather than replacing human oversight, automated oversight seeks to automate *part* of that responsibility. Automating to *some* degree unburdens engineers' responsibility by distributing part of the task to automation. Automated, real-time oversight could provide a way to arrange such measures, however, this should not come at the expense of administrative oversight that necessitates human involvement.[33]

Human-in-the-loop approaches

Beyond administrative or automated approaches to oversight, human-in-the-loop approaches were often seen as a catch-all solution to many existing concerns about automated decision-making systems. If a particular decision may be ethically sensitive, then domain specialists ought to be the final judges of the outcome, not algorithmic systems. This approach is somewhere in between administrative oversight and automation and brings up questions of how human expertise might be valued. Where domain experts may be best positioned to render certain types of judgments despite inherent human biases, the speed and scale of AI systems make use of particular and high-level expertise untenable. There remains, however, the precarious and contingent workforce around the world, often used for these kinds of solutions through gig-economy platforms. This brings up a different ethical concern of fair labour costs and questions of whether such arrangements qualify as the employment of appropriate domain expertise.

All forms of oversight, whether administrative, automated or human-in-the-loop, have their limitations and require broader implementation. No matter the type of oversight, responsibilities of engineers and accountability frameworks deployed for oversight must be clearly developed. Although automated oversight is a potentially efficient and economically favourable tool to ensure AI ethics, this should not preclude human involvement and should be carefully balanced to consider the level of explainability required to sustain trust. Where the human-in-the-loop systems are common, the ethical issues around the use of such human workforce are now being seriously considered.

2.2.6 PUNITIVE MEASURES

How to decide which decisions are unethical and under what circumstances such decisions should be punishable was a central issue debated by all participants. Despite the importance of punitive

32 Munoko et al., "The Ethical Implications of Using Artificial Intelligence in Auditing."

33 Kerr et al., "Expectations of artificial intelligence and the performativity of ethics."

measures, engineers argued that facilitating and promoting a no-blame culture is paramount. A no-blame culture would enable engineers to feel certain that as long as they do not deliberately intend to cause harm and participate actively in spaces for discussion on ethics, they will not be punished for their decisions. No-blame culture is a way to encourage employees to learn from mistakes.

> ***"...there's the often-cited difference between the health sector and the aviation sector when something goes wrong. In the aviation sector, there is a 'no blame' approach in principle. So, there is an open investigation, as long as you participate properly in the investigatory process, and there isn't evidence of criminal neglect, then there won't be a liability for the problem. It's so that the industry together can come to solutions and learn from mistakes. Whereas in the health sector, it's more of a, "you're going to get sued for any kind of malpractice". So problems don't get discussed openly and might not actually lead to solutions."***
>
> **– Ansgar Koene, EY, University of Nottingham, and EMLS RI**

Within the necessary no-blame context, we do, however, need some mechanisms for consequences in all areas of accountability. Engineers also agreed that commercial companies may need some extra motivation through the threat of punitive measures to devote resources that would allow for example, for spaces for discussion, reflection and concern to be developed. Participants, who suggested certification of companies as 'ethically aligned', argued that if a company did not live up to the standard, punishment (i.e., losing certification) could be a mechanism for consequence. Many also pointed out that consequences for ethical problems produced by AI systems are important to signal commitment and to motivate engineers to seek certification, to engage in education, and to simply pay attention to emergent concerns. If there is a clearly unethical practice, then punitive mechanisms need to be in place, even for individuals.

Section 3. The AI Hackathon: process and methodology

The 2020 AI Hackathon focused on locating the gaps and uncertainties in existing discussions of AI and ethics, as well as building on prior work of the organisers in this area. In particular, we took our departure from the following work: the ANE's 2018 Nordic Engineers' Stand on Artificial Intelligence and Ethics Report[34], the IEEE's extensive work on Ethically Aligned Design[35] as well as the documents produced by the related working groups[36], the extensive work on data ethics by the DataEthics ThinkDoTank[37], and the tools produced by the VIRT-EU project[38] recently completed by Irina Shklovski from the University of Copenhagen. **The goal of the workshop was for engineers to define problems from an engineering point of view, identify the responsibilities and obligations available for engineers, discuss what is necessary to fulfil these responsibilities and obligations, and to pinpoint forms of accountability.**

The workshop brought together a diverse group of engineers from across Europe. Participants included members from the Nordic Engineering unions in Denmark, Sweden, Norway, Iceland and Finland as well as members of the IEEE organisation coming from the UK, Spain, Portugal, Russia and the US. Most participants were practitioners, some were directly engaged in policy conversations and standards development work, while a few were engaged in academic research on the topic. Several participants have previously been active in developing strategies and approaches to addressing ethics in AI as part of the IEEE Ethically Aligned Design[39] effort. To ensure a fruitful outcome of the Hackathon, we asked all of our participants to prepare for the event by reading a short framing paper developed jointly by the event organisers and then responding to an online questionnaire in advance so that we could integrate their ideas into the proceedings. Answers to the preliminary questionnaire made clear that **many of the current standards, guidelines and principles on Ethics and AI are not being actively used among engineers, even though most of the respondents were aware of several such documents.** We took this finding as a basis for developing the framework for the hackathon, including questions about challenges, responsibilities and concerns about ethics and AI from an engineering point of view.

Given the online format, we were able to record all of the proceedings. However, this created concerns regarding participant privacy since discussions about ethics can potentially be quite

34 "Nordic Engineers' Stand on Artificial Intelligence and Ethics Report". ANE, 2018.

35 "Ethically Aligned Design". IEEE, 2019.

36 "IEEE-SA Working Group Areas". IEEE Standards Association, 2020.

37 "Data Ethics Principles". DataEthics ThinkDoTank, 2017.

38 "VIRT-EU service package". VIRT-EU, 2019.

39 "Ethically Aligned Design". IEEE, 2019.

sensitive. In recognition of this, we took precautions to limit personal data collection. The ANE was in charge of registrations and kept all personally identifiable data secure. No other partners had access to these data. The University of Copenhagen created the online questionnaire, making sure that the data collected through this instrument was anonymous. The ANE sent the questionnaire link out to the registered participants. During the event, we ensured that all participants were comfortable with the proceedings being recorded. We also asked our participants to clearly indicate whenever they were saying something that shouldn't be included in notes or transcriptions. During the event, all participants indicated that they wished to have their quotes directly attributed to them rather than kept confidential. Throughout the document, we have respected this wish.

In developing this report, we transcribed all of the data, pseudonymising the transcripts, but keeping a separate indication of participation to enable us to attribute quotes when we used these. After transcription, all video content was deleted to ensure participant privacy. All data were anonymised for data analysis. Once the analysis was completed, we backtracked to identify and attribute quotes we had selected as representative examples to illustrate our points. All of the participants were allowed to comment on the report and those whose quotes were used in the report were asked to approve them before publication. All names, affiliations and attributions in the report are reproduced as requested by the participants themselves.

The framing paper, developed by the University of Copenhagen, presented an overview of existing principles and guidelines for ethics and AI and asked participants to think about the gaps between principles and practice. We designed the hackathon activities around discussions of the gaps between ethical principles and the practice of AI development. Specifically, we grouped participants into domains of interest and discussing specific, real-life cases related to each domain. Discussions on each topic cantered around a particular case selected by the participants and followed a set of guiding questions introduced by the organisers, who acted as facilitators. Below we introduce the group topics and the cases selected by the participants:

> **Group 1 (facilitated by the DataEthics ThinkDo Tank):** Healthcare & Tracking: (Case: social/companion robot)
>
> **Group 2 (facilitated by the ANE):** Employment (Case: workplace surveillance technology for productivity tracking)
>
> **Group 3 (facilitated by the IEEE):** Healthcare and ageing (Case: health data collection to predict healthy ageing – a project funded by Novo Nordisk, as well as some discussion on Amazon Hulu device)
>
> **Group 4 (facilitated by the University of Copenhagen):** Education: (Case: Use of algorithms to predict students' grade based on examples of UK GSCE and A-level and the IB-international baccalaureate controversies)

Each group sought to define the problem from an engineering point of view and to identify what kinds of responsibilities and obligations must be taken on by the engineers. Finally, the groups iden-

tified practices, discussions, guidelines, and knowledge necessary for fulfilling these responsibilities and obligations. We provided the cases to create a non-judgmental atmosphere for discussing ethical challenges in the development of AI systems, thus offering participants an opportunity to bring up their own experience but not requiring this.

The second group exercise used the scaling What If Everyone In The World (WIEITW)[40] tool developed by VIRT-EU project. Using examples discussed in the prior exercises, the groups were asked to consider "what if everyone in the world" were to use a particular product or service agreed from their earlier discussions. Given the outcomes of the exercise, participants were to reflect on 1) what might have been overlooked in the original development process; 2) the principles or guidelines that might need to be in place to avoid this; 3) whose responsibility should it be to make these considerations; 4) if problems are identified, who should be held accountable, in what way, and to whom? These questions allowed participants to not only locate gaps in accountability and governance but also provided them with a space to express their concerns openly.

For the final exercise, participants prepared a presentation to summarise their group discussions for everyone in a joint plenum. In these presentations, participants defined the problems from an engineering point of view, identified the responsibilities and obligations that engineers could or should take on, identified practices, discussions, guidelines and knowledge necessary to fulfil these, discussed what form accountability might take and what is required for this. Following the presentations, engineers were free to comment on and discuss each of the group topics. The final phase of the event culminated in an extensive plenum discussion that led participants to agree on common issues that had emerged throughout the event.

Participants jointly discussed and agreed upon the following proposals that formed the foundation for this report:

- **Documentation is key** – design decisions in AI development must be documented in detail, potentially taking inspiration from the field of risk management.
- **There is a need to develop a framework for large-scale testing of AI effects**, beginning with public tests of AI systems, and moving towards real-time validation and monitoring.
- **Governance frameworks for decisions in AI development need to be clarified**, including the questions of post-market surveillance of product or system performance.
- **Certification of AI ethics expertise** would be helpful to support professionalism in AI development teams.
- **Distributed responsibility should be a goal**, resulting in a clear definition of roles and responsibilities as well as clear incentive structures for taking in to account broader ethical concerns in the development of AI systems.
- **Spaces for discussion of ethics** are lacking and very necessary both internally in companies and externally, provided by independent organisations. Looking to policy en-

40 "Paper Tools". VIRT-EU, 2019.

suring whistleblower protection and ombudsman position within companies, as well as participation from professional organisations.

- **One solution is to look to the existing EU RRI framework** and to ensure multidisciplinarity in AI system development team composition. The RRI framework can provide systematic processes for engagement with stakeholders and ensuring that problems are better defined.
- **The challenges of AI systems point to a general lack in engineering education.** We need to ensure that technical disciplines are empowered to identify ethical problems, which requires broadening technical education programs to include societal concerns.
- **Engineers advocate for public transparency** of adherence to standards and ethical principles for AI-driven products and services to enable learning from each other's mistakes and to foster a no-blame culture.

References

Agrawal, Ajay, Joshua Gans, and Avi Goldfarb. "How AI Will Change the Way We Make Decisions". *Harvard Business Review*, 26 July 2017. https://hbr.org/2017/07/how-ai-will-change-the-way-we-make-decisions

Aitken, Mhairi, Ehsan Toreini, Peter Carmichael, Kovila Coopamootoo, Karen Elliott, and Aad van Moorsel. "Establishing a Social Licence for Financial Technology: Reflections on the Role of the Private Sector in Pursuing Ethical Data Practices". *Big Data & Society 7*, no. 1 (1 January 2020): 1–15. https://doi.org/10.1177/2053951720908892

Ananny, Mike, and Kate Crawford. "Seeing without Knowing: Limitations of the Transparency Ideal and Its Application to Algorithmic Accountability." *New Media & Society 20*, no. 3 (1 March 2018): 973–89. https://doi.org/10.1177/1461444816676645

Beaudouin, Valérie, Isabelle Bloch, David Bounie, Stéphan Clémençon, Florence d'Alché-Buc, James Eagan, Winston Maxwell, Pavlo Mozharovskyi, and Jayneel Parekh. "Flexible and Context-Specific AI Explainability: A Multidisciplinary Approach." *SSRN Electronic Journal*, (12 March 2020): 1-65. https://doi.org/10.2139/ssrn.3559477

"Data Ethics Principles". DataEthics ThinkDoTank, 2017. https://dataethics.eu/data-ethics-principles/

"Ethically Aligned Design" (first edition). IEEE, 2019. https://standards.ieee.org/content/dam/ieee-standards/standards/web/documents/other/ead1e.pdf

European Commission. "Algorithms and Democracy - AlgorithmWatch Online Policy Dialogue, 30 October 2020." European Commission, October 30, 2020. https://ec.europa.eu/commission/commissioners/2019-2024/vestager/announcements/algorithms-and-democracy-algorithmwatch-online-policy-dialogue-30-october-2020_en

Fiesler, Casey, Natalie Garrett, and Nathan Beard. "What Do We Teach When We Teach Tech Ethics? A Syllabi Analysis". *In Proceedings of the 51st ACM Technical Symposium on Computer Science Education*, 289–295. SIGCSE '20. New York, NY, USA: Association for Computing Machinery, 2020. https://doi.org/10.1145/3328778.3366825

Hagendorff, Thilo. "The Ethics of AI Ethics: An Evaluation of Guidelines." *Minds & Machines 30* (February 1, 2020): 99-120. https://doi-org.ep.fjernadgang.kb.dk/10.1007/s11023-020-09517-8

High-level Expert Group on Artificial Intelligence. "Ethics Guidelines for Trustworthy AI." 2019. https://ec.europa.eu/futurium/en/ai-alliance-consultation/guidelines/1

High-level Expert Group on Artificial Intelligence. "Sectoral Considerations on the Policy and Investment Recommendations for Trustworthy Artificial Intelligence." 2019. https://futurium.ec.europa.eu/en/european-ai-alliance/document/ai-hleg-sectoral-considerations-policy-and-investment-recommendations-trustworthy-ai

IEEE Ethics in Action in Autonomous and Intelligent Systems. https://ethicsinaction.ieee.org

"IEEE-SA - Working Group Areas", 2020. https://standards.ieee.org/initiatives/artificial-intelligence-systems/index.html

'Iron Ring'. In *Wikipedia*, 30 June 2020. https://en.wikipedia.org/w/index.php?title=Iron_Ring&oldid=965316153.

Jobin, Anna, Marcello Ienca, and Effy Vayena. "The Global Landscape of AI Ethics Guidelines." *Nature Machine Intelligence 1*, no. 9 (September 2019): 389–99. https://doi.org/10.1038/s42256-019-0088-2.

Kerr, Aphra, Marguerite Barry, and John D Kelleher. "Expectations of Artificial Intelligence and the Performativity of Ethics: Implications for Communication Governance." *Big Data & Society 7*, no. 1 (1 January 2020): 1–12. https://doi.org/10.1177/2053951720915939

Metcalf, Jacob, Emanuel Moss, and danah boyd. "Owning Ethics: Corporate Logics, Silicon Valley, and the Institutionalization of Ethics". *Social Research: An International Quarterly 86*, no. 2 (28 August 2019): 449–76.

Mittelstadt, Brent. "Principles Alone Cannot Guarantee Ethical AI." *Nature Machine Intelligence 1*, no. 11 (November 2019): 501–7. https://doi.org/10.1038/s42256-019-0114-4

Mittelstadt, Brent Daniel, Patrick Allo, Mariarosaria Taddeo, Sandra Wachter, and Luciano Floridi. 'The Ethics of Algorithms: Mapping the Debate'. *Big Data & Society 3*, no. 2 (1 November 2016): 1-21. https://doi.org/10.1177/2053951716679679

Munoko, Ivy, Helen L. Brown-Liburd, and Miklos Vasarhelyi. "The Ethical Implications of Using Artificial Intelligence in Auditing." *Journal of Business Ethics*, 167 (January 8, 2020): 209–234. https://doi.org/10.1007/s10551-019-04407-1

"Nordic Engineers' Stand on Artificial Intelligence and Ethics Report". ANE, 2018. https://ipaper.ipaper-cms.dk/IDA/ane/report/

"Paper Tools". VIRT-EU, 2019. https://www.virteuproject.eu/servicepackage/paper-tools/

Schiff, Daniel, Justin Biddle, Jason Borenstein, and Kelly Laas. "What's Next for AI Ethics, Policy, and Governance? A Global Overview." *In Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, 153–58. New York: Association for Computing Machinery, 2020. https://doi.org/10.1145/3375627.3375804

Simon, Judith. "Distributed Epistemic Responsibility in a Hyperconnected Era". In *The Onlife Manifesto*, 145–59, 2015. https://doi.org/10.1007/978-3-319-04093-6_17

Tupasela, Aaro, Karoliina Snell, and Heta Tarkkala. "The Nordic Data Imaginary." *Big Data & Society 7*, no. 1 (1 January 2020): 1-13. https://doi.org/10.1177/2053951720907107

"VIRT-EU service package". VIRT-EU, 2019. https://www.virteuproject.eu/servicepackage/

January 2021

IEEE SA
STANDARDS ASSOCIATION

ane | ASSOCIATION OF NORDIC ENGINEERS

UNIVERSITY OF COPENHAGEN

DATAETHICS