ane | association of nordic engineers

Nordic engineers' stand on Artificial Intelligence and Ethics

Policy recommendations and guidelines

Produced by the IT University of Copenhagen in cooperation with ANE.

IT UNIVERSITY OF COPENHAGEN

ETHOSLAB

Table of Content

Executive Summary

The Global Risks Report 2017, published by the World Economic Forum, lists artificial intelligence (AI) and robotics as the third of twelve key emerging technologies. These emerging autonomous technologies are disrupting established business models, and changing society in ways that are not always easy to anticipate. The European Commission has stated that "the way we approach AI will define the world we live in", and this is supported by the proliferation of discussion and concerns among policy makers, business interests and general public alike. While advances in machine intelligence provide unprecedented opportunities, large-scale data collection to support such opportunities represents significant causes for concern. To mitigate the risks, a regulatory framework is needed that includes ethical standards, normative expectations, assessment of responsibility, and accountability for actions. Questions such as who should take moral, ethical and legal responsibility for artificial intelligence technologies need to be prioritized on the political agenda. The Nordic countries are known for low levels of corruption, high levels of involvement of civil society in policy making and a commitment to ethical treatment of consumers and of the labor force. As such, the Nordic countries are positioned well to be frontrunners in setting the agenda for how to address the issues of ethics in AI development and implementation.

Engineering plays an essential role in building, sustaining, and improving the quality of life for individuals in contemporary societies. In this way engineers are at the forefront of developing autonomous systems and adding machine intelligence to existing mechanisms and processes. The many standards and codes of conduct agree that one of the major responsibilities of engineers is to promote positive outcomes for society, and to limit harm. However, in a rapidly changing world, what comes to constitute a positive outcome, and what could potentially cause harm have become much more difficult to recognize. Current guidelines and standards often do not fully address the problems that engineers face and the responsibilities they must take on in working with AI. As the stakes rise so does the need for addressing the ethics of engineering in practice more directly.

The Association of Nordic Engineers (ANE) membership is composed of national unions, all of which have their own versions of guidelines and codes of conduct for engineers to use in their work today. However, recent developments in AI and machine learning have outpaced most of the existing ethical guidelines and frameworks for best practices. In this time of global digital data economy and an increasingly fast pace of technological change, ANE sees the development of an overall position for all Nordic engineers on what constitutes ethical conduct with respect to AI as a crucial step forward.

On September 25th, 2018, the ANE in cooperation with the IT University of Copenhagen organized an ethics hackathon entitled "Nordic engineers' stand on the EU future AI and ethics framework", in order to gather engineers from five Nordic countries to collaboratively develop a joint position based on practical experience and in conversation with current debates on AI and ethics. The resulting policy document, recommendations and guidelines were produced using the output of the hackathon, and it reflects the collective view of the Nordic engineers on AI and ethics.

> "While engineers and their organizations will need to shoulder much of the growing responsibilities in the design and implementation of AI systems, the relevant governing bodies of the Nordic countries and at EU level must acknowledge their own responsibilities and opportunities for action."

RECOMMENDATIONS FOR GOVERNMENT BODIES TO ADDRESS ISSUES OF AI AND ETHICS

Where specific implementations of particular ethical engineering conduct in practice is best left to companies and the engineers themselves, issues such as the necessary changes in education, implementation of new forms of legislation and regulation remain the purview of governance activities at the national and regional level. As such, we present a set of policy recommendations for consideration.

POLICY RECOMMENDATIONS

1. There is a need to anchor discussions on the political level and to advance the public understanding on AI. This could be accomplished through the creation of a platform - a meeting space that would engage decision makers, business, academia, civil society and professionals including engineers to come up with stable and transparent solutions for AI through joint discussions.

2. Education for ethical considerations and guidelines is often insufficient in the technical disciplines and throughout work-life. This needs to be addressed through changes in educational goals and priorities for technical subjects as well as through provision of relevant opportunities for lifelong learning.

3. Development of an appeal process with governmental oversight is crucial. Such a process must enable individuals and organizations to address the AI behaviour and decisions that they find potentially harmful.

4. There is a need for shaping regulation and legislation to govern issues related to AI that formalises relevant responsibility and defines accountabilities.

5. Engineers, policy makers, civil society and the general public need spaces for sustaining a living dialogue around issues of AI and ethics. These need to be facilitated and supported through funding and other forms of support.

GUIDELINES FOR ENGINEERS AND THEIR ORGANIZATIONS

The guidelines below have organically emerged from discussions with engineers as well as from an overview of other ongoing efforts to address the issues of AI and ethics. These are not exclusively for individual engineers to follow, because ethical development of AI will not come about only as a result of individuals taking on particular types of ethical responsibility. There are plenty of guidelines for what constitutes ethical conduct for engineers and some of the guidelines below can be taken on board by individuals and organizations alike as additions to those that are already in existence in the Nordic countries. However, many of the guidelines are oriented towards organizational practices rather than individual responsibility, because efforts towards ethical practices need strong institutional backing to be effective and therefore organizational commitment is a requirement for addressing ethics in AI.

We present these guidelines with an understanding that their implementation will require effort and commitment on the part of the individual engineers and of their organizations together.

GUIDELINES OF ETHICAL CONDUCT FOR AI DEVELOPMENT AND IMPLEMENTATION

1. Create spaces for discussion of the issues around AI and ethics. These need to be facilitated and supported by both workplaces and civil society organizations.

2. Invest into and develop tools that enable ethical discussions, questions and decision making throughout the design process and not only at the beginning and the end.

3. Establish a set of internal standards and checklists tackling ethical issues in AI development such as ensuring meaningful human control.

4. Support and facilitate internal reporting of risk and violations, establishing rules for clear action in response.

5. Establish internal training programs for staff to deepen an understanding of ethics and to develop skills for ethical reflection, debate and recognition of biases.

6. Pay special attention to potential biases encoded in system development, training data and model performance, especially those that may affect the most vulnerable.

7. Develop ways for accepting organizational responsibility for potential harm, for example, by establishing ways to address the harm inflicted on others by AI systems that the organization has built.

8. Establish an internal ethical review process that democratizes company decision-making by involving more internal actors.

9. Work to increase transparency not only in the decisions leading to design and development of AI systems, but also in organizational chains of responsibility.

10. In working towards transparency, maintain awareness that transparency has its own ethical pitfalls and limits.

> “Efforts towards ethical practices need strong institutional backing to be effective and therefore organizational commitment is a requirement for addressing ethics in AI.“

Introduction

Although ideas about AI have been active imaginaries for centuries, early concerted research in this direction began in the US in the wake of World War II, inspired by wartime technological innovations of signal detection, code-breaking and ammunition tracking. At the time the goal was to develop devices that could act in an intelligent and autonomous fashion through a fusion of science and technology. A group of researchers at the Massachusetts Institute of Technology (MIT) had built the first neural network already in the 1950s. The initial goals of this community, however, were so over-optimistic that within several decades the discourse around AI moved back to the discussions of imaginaries, although the work on many of the underlying technical processes and innovations continued apace.1 As a result, although there is significant continuity to the development of technologies that are now termed AI, the recent proliferation of concerns may make it feel as if the problems we are now facing suddenly arose just in the first two decades of the 21st century. For example, the last two decades of the 20th century saw development of expert systems that were intended to provide technical information processes to support human decision-making. Implementations of such systems also raised concerns about incorporating biases, when early recommendations suggested that such systems should only be used in an advisory capacity to human decision makers.2 Although much of the current discussion of AI may still be futuristic, the reality is that at least in the Western world we already have many autonomous systems deployed across many functions of society from hospitals, to government, to executive decision-making.3 SIRI is not the first AI to enter the home although she is perhaps the first that is rather communicative.

As advancements in computing speed, growth of network connectivity and the proliferation of big data drive rapid research and innovation in machine learning, data mining and neural network applications, the concerns these developments raise are increasingly acute. As a result, current discussions of the problems arising from AI must address a set of far more powerful and sophisticated technologies although the basic thrust of the expressed concerns remains similar. How might we develop AI technology that produces a positive impact on society?

1 Agre, "Toward a Critical Technical Practice: Lesson Learned in Trying to Reform AI."

2 Khalil, "Artificial Decision-Making and Artificial Ethics."

3 Crawford and Calo, "There Is a Blind Spot in AI Research."

The advent of myriad AI systems and solutions has had a significant impact on how prediction is conceptualized and managed. If the use of computing systems for prediction used to be a significant financial investment, it is now merely an expected part of any computing system. As prediction becomes a commonplace effort the question remains who might judge the appropriateness of these predictions? No matter the efforts to automate many current processes and practices, many have written that AI is still "made out of people"[4] relying as it often does on human judgement as part of input or even the final stage.[5] This move to use prediction in new areas also opens avenues for new forms of experimentation as prediction and experimentation are inextricably linked. The uses of experimentation in far broader applications across autonomous systems raise novel ethical concerns.[6] How might engineers to develop systems that are "provably safe" even after recursive self-improvement, is there a need for a new approach to safety engineering?[7]

Alongside the issues of bias in computation, another important aspect of AI is the issue of diversity in the workforce that produces and implements these technologies. Many have argued that the glaring lack of diversity in technical occupations is significant problematic and for example, is one reason that hampering efforts to address problems of bias in algorithmic systems.

> "Focusing on the role of human decisions in the creation of technologies is a way to retain responsibility and to care for each other since "technologies don't care"."[8]

Such considerations of responsibility are necessary in the face of technological changes that reconfigure power dynamics in our social structures, leaving pre-existing ethical rules unreliable and our ability to predict the potential consequences of design and implementation limited.

To consider these issues not from the outside but together with engineers in order to retain the focus on the human practice involved in the creation of AI, on September 25th, 2018, the ANE in cooperation with the IT University of Copenhagen organized an ethics hackathon entitled "Nordic engineers' stand on the EU future AI and ethics framework." We gathered engineers from five Nordic countries to collaboratively develop a joint position based on practical experience and in conversation with current debates on AI and ethics. The policy document in front of you, its recommendations and guidelines were produced using the output of the hackathon. As such, this document reflects the collective view of the Nordic engineers on AI and ethics.

[4] Irani, "The Hidden Faces of Automation."

[5] Agrawal, Gans, and Goldfarb, "The Simple Economics of Machine Intelligence."

[6] Bird et al., "Exploring or Exploiting?"

[7] Yampolskiy, "Artificial Intelligence Safety Engineering."

[8] Silverstone, "Proper Distance: Toward an Ethics for Cyberspace."

THE REMAINDER OF THIS REPORT IS STRUCTURED IN THE FOLLOWING MANNER:

Section 1 (The ANE Hackathon) describes the Hackathon itself, its structure and process.

Section 2 (Operational definitions) presents a discussion of the how the ANE membership has agreed to define the operational notions of AI and ethics through their discussions.

Section 3 (Pressing Issues for an Ethics of AI) lists the most prominent issues and concerns that need to be dealt with in order to produce ethical means of working with AI. These issues are transparency, accountability and trust, avoiding harm, and addressing bias.

Section 4 (Opportunities for addressing pressing issues) considers how the issues identified in the prior section might be addressed and what are some of the more practical implications of these actions.

Section 5 (Recommendations and Guidelines) details recommendations for working with issues of ethics in AI for individual engineers, engineering institutions and governments.

The ANE Hackathon

On October 25th, the ANE in cooperation with the IT University of Copenhagen organized a hackathon entitled “Nordic engineers’ stand on the EU future AI and ethics framework”. The goal of the workshop was the development of a set of recommendations and guidelines, which in turn contributed to the production of this report.

The workshop brought together a diverse group of engineers from five Nordic countries. Participants included members from Nordic unions: the Swedish Association of Graduate Engineers, Sveriges Ingenjörer, the Norwegian Association of Engineers and Technologists, NITO, Association of Chartered Engineers in Iceland, VFÍ, the Danish Society of Engineers, IDA and the Association of Academic Engineers and Architects, TEK. A majority of participants were practitioners, while a few were engaged in research on the topic.

The workshop was designed around an initial framing paper, which aimed to provide an overview of current debates regarding AI and ethics. The framing paper was composed by researchers at the IT University of Copenhagen, who surveyed academic literature and previously published ethical guidelines for AI. The framing paper was shared with the participants prior to the workshop.

The workshop activities were centred around discussions presented in the framing paper. Specifically, the participants were divided into five groups and each group was invited to engage with one issue outlined in the paper during the day of workshop. As a first step, all groups were asked to come up with working definitions of AI and ethics. The exercise involved an individual component where participants were invited to reflect on their own professional experience, as well as a group activity where they synthesized their individual reflections into a definition according to consensus within the group. At the end of the task, two groups were asked to present their definitions in a plenary session, and all participants were invited to comment on the presentations.

In the second task, groups were assigned one of the five different issues presented in the framing paper and asked to discuss it from the standpoint of their own professional experience. Anchored in their assigned issue, the participants produced examples where the issue presented itself in the practice of developing AI and discussed the implications. This exercise was again followed by presentations by two of the groups and a plenum discussion.

The final task invited participants to use the outcomes of previous two tasks as a springboard for coming up with practical guidelines that could be used in the development of this policy paper.

Participants discussed ethical ways of working with AI and considered what they themselves would have liked to have when engaging with developing or implementing AI or what they would like to impart to their junior colleagues. Each group then presented their propositions in plenum as part of an extensive final discussion. These guidelines and recommendations form the core of ANE's position on AI and ethics, and are reflected in this document.

Operational Definitions

An essential task in crafting any framework from collective conversation is to establish mutually agreed upon definitions of the major terms in question. In this case, the terms AI and ethics are central to our discussion. The hackathon participants were challenged to discuss their own definitions of these terms, taking their departure from the framing paper, and then to come to an agreement on a jointly shared understanding. The following sections provide definitions that were initially derived from existing literature and then were shaped through conversation with the members of the ANE.

DEFINING ARTIFICIAL INTELLIGENCE (AI)

From Roomba vacuum cleaners to Siri and other mobile phone apps, we are increasingly surrounded by systems that are not only able to understand when they are being addressed, but also respond in ways that are useful. While not exactly “intelligent”, they perform very well in their specific contexts. These systems, from military drones to warehouse robots, to car navigation systems, to robotic assistants for the elderly, exemplify the ever-expanding array of the uses of AI. Even though AI systems have surpassed humans in many specific domains such as chess, there is nearly universal agreement among modern AI professionals that AI falls short of human capabilities in a critical sense.[9] This idea of human capability, and what parts of it machines should emulate are easily visible from how AI is defined by those working with it.

There are no universal definitions of AI in use today, but several are well established. The European Commission statement on Artificial Intelligence in Europe uses the following definition:

> **“**Artificial Intelligence (AI) refers to systems that display intelligent behavior by analyzing their environment and taking actions - with some degree of autonomy - to achieve specific goals.**“**[10]

[9] Bostrom and Yudkowsky, “The Ethics of Artificial Intelligence.”

[10] European Group on Ethics in Science and New Technologies, “Artificial Intelligence, Robotics and ‘Autonomous’ Systems.”

According to the EU Commission,11 some definitions of AI focus on bringing autonomy into artifacts, while other definitions focus on AI as a collection of rapidly converging smart digital technologies that are often interrelated, connected, or fully integrated. This latter group includes classical AI, machine learning algorithms, deep learning and connectionist networks, generative adversarial networks (GANs), mechatronics and robotics. The convergence of these technologies are easily recognisable in innovations such as chatbots, robotic weapon systems, speech and image recognition systems, and self-driving cars.

According to the Institute of Electrical and Electronics Engineers (IEEE) statement on Ethically Aligned Design, "AI describes (typically digital) artifacts that demonstrate any combination of the following capacities: capacity to perceive contexts for action, capacity to act and capacity to associate contexts to actions." The statement goes on to argue that "as the use and impact of autonomous and intelligent systems (A/IS) become pervasive, we need to establish societal and policy guidelines in order for such systems to remain human-centric, serving humanity's values and ethical principles. These systems have to behave in a way that is beneficial to people beyond reaching functional goals and addressing technical problems".12 The above definition of AI used by the IEEE is very broad, the argument that the capacity to act and the action itself must be aligned with humanity's values and principles attempts to delineate the kinds of activities in question here.

Large technology corporations with significant investments in AI have also provided their own definitions. For example, Google provides a seemingly simple definition; "At its heart, AI is computer programming that learns and adapts",13 while IBM does not define the term AI at all, pointing instead to the proliferation of what they term "the A* algorithm" which is "an essential tool for AI, present in every AI teaching book".14 Rather than calling the variety of increasingly autonomous systems in the world "AI", IBM sees a proliferation of the use of AI techniques in system design - be that machine learning, deep learning, or adversarial neural networks (GANs).

During the hackathon, the discussion among the ANE members surfaced many concerns about the use of the term Artificial Intelligence with most participants being unwilling to use the word intelligence due to the complexity of its definition. As one participant explained: "It is not clear what intelligence even is, and which level is required." In the end, however, all agreed that given the renewed recent dominance of the term, it does not make sense to redefine it. Some suggested that: "In a sense AI could be defined as a system that combines automated automation & machine learning with a general notion of context awareness and adaptability." Participants agreed that whatever the definition, AI is not a single type of technology, but a group of technologies all of which display some form of awareness, autonomy and adaptability in automation of tasks and processes. Rather than

[11] European Group on Ethics in Science and New Technologies.

[12] IEEE Standards Association, "Ethically Aligned Design: A Vision for Prioritizing Human Wellbeing with Artificial Intelligence and Autonomous Systems."

[13] Pichai, "AI at Google: Our Principles."

[14] Rossi, "Artificial Intelligence: Potential Benefits and Ethical Considerations."

defining what does and does not qualify as "intelligent", the ANE engineers instead proposed that there is no strict boundary between AI and other computer programs. Instead, they saw a continuum, a broad palette of methods resulting in a variety of technologies defined by levels of adaptability and autonomy. After all, what we consider AI is transforming over time as history has shown already.

> " Many engineers preferred instead the term "machine intelligence" or even "extended intelligence" conveying that the technology in question is more of a set of tools rather than a system on its own. What is at issue here is not man against machine, but how man goes about implementing how the machine thinks." 15

ETHICS

In discussions of technology in general and AI specifically, ethics is the word du jour. Media articles are arguing about ethics, corporations are investing in ethical review committees and inviting civil society organizations to conduct evaluations and research. Such proliferation of ethics discourse around technology has been critiqued as a way for tech business interests to get around regulation, using references to ethics as a form of soft regulation and as a way to showcase goodness to the public.16 The question of course is not whether to be ethical but what is meant by ethics in discussions of AI technologies.

Modern writing on ethical concerns with regard to technology leverages a range of different ethical frameworks. By and large, however, these concerns broadly fall into two general approaches of consequentialist and utilitarian ethics. Much of the ethical assessment of emerging technologies concerns the question of what is good and bad about the products, services and processes that they may bring about, and what is right and wrong about ways in which these may be used.17 Some explorations, such as, for example, discussions of self-driving cars have specifically focused on utilitarian concerns of minimizing harm and maximizing benefits for all affected, while grappling with the difficulties of how to define harm or benefit and how to identify boundaries around who ought to be included in such a calculus.18

In general terms, ethics concerns the frameworks and principles that define individual ability to have a good life and to clearly conceptualize individual rights, obligations and responsibilities.

15 Burrell, "How the Machine 'Thinks.'"

16 Wagner, "Ethics as an Escape from Regulation."

17 Brey, "Anticipating Ethical Issues in Emerging IT."

18 Howard and Borenstein, "The Ugly Truth About Ourselves and Our Robot Creations."

As the Norwegian Society of Engineers and Technologists explains: “Ethics does not give us any recommendations or orders. Instead, it gives us practical tools to distinguish between good and bad reasons, thus making wise decisions.”19 Despite the lack of precision in the definition, however, many of the ethical guidelines that were reviewed provide some sort of practical recommendations combined with some discussions of general principles. We discuss a few of these below.

The British Standard BS8611:2016, titled “Robots and Robotic Devices: Guide to the Ethical Design and Application of Robots and Robotic Systems”, defines ethics simply as “a common understanding of principles that constrain and guide human behavior.”20 The IEEE statement on Ethically Aligned Design, on the other hand, does not define ethics beyond stating that autonomous and intelligent systems “have to behave in a way that is beneficial to people beyond reaching functional goals and addressing technical problems.”21

> “ Discussions of ethics are discussions of obligations engineers must take on in their work and practice. Ethical obligations have two dimensions: professional and personal.22 The former codifies decision making and behavior in expert practice, while the latter ensures that individual reflection and action are present when professional codes of conduct fall short.“

Professional ethics delineates how broader ethical standards, such as responsibility, integrity, fairness, transparency and avoidance of harm apply to the particular types of work that engineers do. Being a professional means being part of a moral community of others who share the same responsibilities and being able to draw on the experience of others to navigate similar moral dilemmas, tough decisions or adverse consequences. The personal dimension ensures that individuals are not indifferent to their effect on the lives of others where professional codes of conduct fall short. Personal ethics enables engineers to take responsibility for their own moral choices and consequences in the face of the moral choices made by their employer should these not align.

Professional guides and codes of conduct provide recommendations but are not meant to be checklists or exhaustive accounts of how to be ethical in any given situation an engineer might encounter. These are tools intended only to help engineers learn to judge what is ‘appropriate’ in any given circumstance. In this light, the ANE members have developed their own definitions and terms of engagement.

Throughout the hackathon, ANE members acknowledged that ethics depends on cultural values and changes over time. They questioned whether ethics of AI had to be different from any other ethics and looked to existing guidelines as starting points. Participants agreed that ethics can be thought of

[19] The Norwegian Society of Engineers and Technologists., “Code of Ethics for Engineers and Technologists.”

[20] BS 8611: 2016, “Robots and Robotic Devices: Guide to the Ethical Design and Application of Robots and Robotic Systems.”

[21] IEEE Standards Association, “Ethically Aligned Design: A Vision for Prioritizing Human Wellbeing with Artificial Intelligence and Autonomous Systems.”

[22] Vallor et al., “An Introduction to Software Engineering Ethics.”

as guiding principles for human behavior. As one participant commented: “Making things right is hard, as is making the right things. Ethics comes into play throughout the process.” Participants agreed that ethical principles for the development of autonomous systems were needed if only as constraints and guidelines, determined by what can be considered morally, culturally and socially acceptable. The discussion was more focused on who got to decide what is acceptable and under what conditions.

While some moral principles are not universal and change over time, others are more consistent, such as those enshrined in the UN Universal Declaration of Human Rights (UDHR). A recent report from the Data & Society think tank argues that the UDHR could help chart the course to guide AI development.23 The demands for respecting human dignity, upholding nondiscrimination and equality and protecting freedom of expression are ideals that all could agree with in principle, but how might such demands translate into practical decision making was less clear. The hackathon participants questioned however, who must be held responsible for upholding or violating these principles in the design of AI. Technical companies and organizations make a point of ensuring legal compliance with current data protection legislation and other regulations relevant for a particular area of activity. Given the rapid evolution of AI technologies, however, legal compliance may not be enough. Who must set the boundaries and guidelines that go beyond legal compliance in any project or organizational structure? The question of where do responsibilities lie and who must be held to account for the potential adverse outcomes does not have a clear answer. The engineers might be principled and ethical but this may not be enough in large distributed projects where the full architecture is difficult to comprehend for people who are working on different aspects of the same system. Clearly the ethical discussions are not only the responsibility of individual engineers. The hackathon participants were very clear that discussions about the ethics of the technologies being developed and built must happen at different levels within organizations.

[23] Latonero, “Governing Artificial Intelligence.”

Pressing Issues for an Ethics of AI

The excitement about AI is building, as evidenced by three expansive reports produced in 2016 by the White House Office of Science and Technology Policy (OSTP), the House of of Commons' Science and Technology Committee in the UK and the European Parliament's Committee on Legal Affairs respectively. These reports lay out a vision of what to do in order to prepare for the future of broad implementations of AI into every aspect of the modern society. Cath et al.[24] compare these three reports and their treatment of the fundamental question that is posed by AI, that of its ethical, social, economic and political impact. The three reports lay out their visions of "a good AI society"[25] as well as their expectations of the kind of regulatory role each government would be willing to take on, consistent with respective approaches to governance. Where the OSTP envisions self-regulation in the tech section, the EU report advocates development of new institutional arrangements and legal structures for addressing possible risks while supporting research and development.

Throughout, the reports call for more research and development in AI in order to take advantage of its potential but warn that efforts must be made to ensure transparency, accountability and alignment in human values in the design of these technologies. There is a strong emphasis not only on minimizing bias in the developed AI systems, but also on ensuring diversity in the workforce as well as considerations of how educational systems may need to be reformed to address the mounting needs and pressing concerns.[26] The potential problems with broad scale implementation of AI systems identified in the three reports are reminiscent to the same issues debated at length across academic, media and policy discussions.

In this section, we present the issues raised by existing and newly developed international codes of conduct and statements of concern that deal with ethics in AI, software, or digital technologies in general (a list of relevant reports is included in the appendix) and those that were debated by the ANE hackathon participants. The issues are grouped under five partially overlapping categories of responsibility towards society, transparency, accountability & trust, avoiding harm and addressing bias.

[24] Cath et al., "Artificial Intelligence and the 'Good Society': The US, EU, and UK Approach."

[25] Cath et al.

[26] Cath et al.

RESPONSIBILITY TOWARDS SOCIETY

Engineers as a profession have a long history of discussing obligations and responsibilities. For example, at the turn of the last century, Canadian engineers instituted a ceremony known as "The Ritual of the Calling of an Engineer" where young engineers are conferred an iron ring that they must wear on their little finger throughout their professional career to remind them of their obligations and responsibilities.27 While not a marker of qualification, the ring has a strong symbolic meaning of the power of engineers and that it must be used for good. As professionals, ANE members have a responsibility towards society arising from their key role in the design, development, and production of technologies. The role technologies play in society, and how technologies organise society, are important concerns for ANE members.

Starting with the question, "around which values do we want to organise our societies?" the ANE workshop participants discussed the role of technologies, specifically the role of technologies that make use of AI methods, in the organisation of societies. A commonly accepted assumption, minimising negative consequences of technologies, serves as a good starting point, but ultimately remains too broad for practice. More specific questions, such as "Who benefits from the development of AI? Are those only a few individuals, specific groups, or a larger population?" can guide us towards more concrete answers when tackling the issue of responsibility towards society.

The engineers taking part in the ANE hackathon took as their own responsibilities diverse concerns from ensuring a positive impact on their societies to safeguarding of democratic processes. As an example, they discussed the problem of the interference in the 2016 US presidential election and the responsibility of engineers to develop systems that can prevent this. Yet at the same time they were aware of the limits of both existing conceptualizations and connected to roles and power. Many commented on disparate access to power in organizations where engineers can not make the same kinds of of decisions as manager. In other cases, engineers had a hard time estimating unintended consequences of their work. For example, a participant referenced Airbnb: the idea was good (renting a room for cheap) but in reality it made the prices for apartments in popular cities increase out of reach for local residents who were seeking to either rent or buy properties. Thus, systems can be misused, and this is more problematic in connection with the way concentrations of money and power are distributed, which in turn might cause impasses in innovation. Therefore, questions of how to identify problems and how to determine whether proposed solutions can cause new problems were acute. Most importantly, many participants in the hackathon asked: What are we optimizing for and for whom? This is a crucial question given the focus on optimisation in technical development and innovation. Assessing the societal impact of technologies on society might be daunting, but knowing how to ask the right questions and making an effort to see technologies in a broader context is paramount.

27 "Background | The Iron Ring."

TRANSPARENCY

AI technologies are not easy to understand for those not involved in their design and development. Even those with considerable knowledge on the subject often find it difficult to understand how software and devices with AI-content produce their output as their algorithms remain opaque. This opacity complicates efforts to determine how decisions were made, whether there are errors, how these might have occurred. This makes explaining the underlying logic of a particular system to a larger group, whether that is fellow professionals or the broader public that is affected by its operation, very difficult. As system learn to perform tasks in an increasingly autonomous fashion, that is without human operator supervision, they may produce outcomes not envisioned by the original designers. One of the ways many have proposed to address this issue is through ensuring that how autonomous systems operate must be transparent to all of the relevant stakeholders. The IEEE Vision for Ethically Aligned Design28 notes that "the term transparency also addresses the concepts of traceability, explicability, and interpretability." For many of the existing documents that address ethics in AI, transparency is also essential for consent; no one can consent meaningfully when they do not understand the implications of consenting. Hence, the question is: How can the highly intricate inner workings of the systems engineers develop be reconciled with the pressing need to explain how they function to others?

Although transparency is often seen as a solution to addressing many of the ethical issues in the functioning of AI systems, it is not a complete solution. In fact, as a solution transparency has many limitations. After all, just because something is transparent about its processes does not mean it is understandable or something that can be acted upon. In fact, there are situations where full transparency can result in significant harm.29 Although companies often invoke notions of the importance of protecting trade secrets as an argument against transparency, what is made visible, to whom and for what purpose are questions that must be considered carefully. Furthermore, efforts towards transparency can often produce so much information that what is important can be made obscure in the deluge unintentionally. How much must be made visible, when and to whom are not simple questions to be answered given the complexity of AI systems. Finally, attempts at transparency do not necessarily result in building trust.30 These concerns suggest that, while transparency is a worthwhile goal, its applications require considerations of potential pitfalls as well.

Participants in the workshop were more than willing to engage with the issue of transparency, and gave it due consideration. One participant brought up the oft-debated concerns about opacity of algorithmic decision-making processes, noting that the demand for transparency seems to expose a fault in our current decision-making process, because not even the decision-making process of humans is transparent. After all, human decision making can be quite mysterious as well, when

28 IEEE Standards Association, "Ethically Aligned Design: A Vision for Prioritizing Human Wellbeing with Artificial Intelligence and Autonomous Systems."

29 Ananny and Crawford, "Seeing without Knowing."

30 Albu and Flyverbom, "Organizational Transparency."

for example time of day can have an effect on how judges make decisions about parole. While some participants proposed that approaches to design might need to include something like "Transparency by Design" throughout the development process, others pointed to the need for independent verification as a means to address the problem.

However, the ANE hackathon participants also acknowledged that efforts towards transparency in the design of AI technologies is only part of the solution. It is also important to have transparency in how decisions are made throughout the organizations responsible for building these technologies - visible internally as well as externally as a way to foster greater levels of both accountability and trust.

ACCOUNTABILITY & TRUST

In all technological development, questions of accountability and trust are deeply connected to the structures of the organizations that produce particular technologies. Consider a bridge built by the state or a social media platform built by a private corporation. Gaining the trust of those who have to live with these technologies is closely connected to establishing chains of accountability within and outside the organization responsible for them. Where transparency may be one aspect of fostering accountability, it does not necessarily ensure the development of trustful relationships between the technologies in question and their stakeholders.31

Concerns with accountability and trust in autonomous systems is a mainstay of current ethical discussions with respect to AI, mentioned in the EU statement on Artificial Intelligence, Robotics and 'Autonomous' Systems, the IEEE vision for Ethically Aligned Design, all three government statements on the future of AI32 and many others. There is considerable agreement here that in the event that an AI system acts in a way that we do not anticipate or understand, claiming ignorance cannot absolve engineers of the ethical responsibility for the outcome. It is clear that designers and developers must remain accountable for the outcomes of their own work. How does this accountability intersect with the goals of the organizations that produce these technologies? In the case of autonomous systems, "who is responsible when things do things?".33

Participants in the ANE hackathon noted repeatedly that "accountability goes hand in hand with transparency." However, "gaining trust and ensuring accountability is not something separate but instead goes together," and trust can be "very quickly lost if misused". Despite the imaginaries of newly designed systems entering into our world, the reality is that many current AI systems are far from perfect, often outdated, and their outcomes sometimes unpredictable. When design

31 Ananny and Crawford, "Seeing without Knowing."

32 Cath et al., "Artificial Intelligence and the 'Good Society': The US, EU, and UK Approach."

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

requirements for AI implementations are often specified by either customers or companies, how is any engineer to gain a position where their opinion on accountability or trust might have influence? Accountability for personal actions and for maintaining professional conduct is always relevant, but how far can accountability for the actions of any autonomous system be taken? Some engineers suggested that it is important to establish processes inside organizations so that accountability could be assigned reasonably; for example, 'checkpoints' could be established at different stages of the application's lifecycle (planning, implementation, deployment) where responsibility is taken by engineers and/or users to a degree that takes into consideration the contingencies of that particular stage of development. The important point here is that responsibility cannot only be in the hands of the designer or the engineer, but it has to be distributed across the process and its stakeholders. This requires that all stakeholders build awareness of possible issues, potentially including the use of risk assessment and verification tools.

> " Many engineers insisted that in the work with AI it is important to adhere to the golden rule when imagining the intended users and keeping in mind that most stakeholders may not be able to make a decision that is quite as informed as the developer's."

AVOIDING HARM

Harm and efforts to avoid it is deeply connected to the engineer's responsibility towards society. The principle of avoiding harm is paramount and central to many codes of conduct, but it requires specification to be applied in practice. Some documents define avoiding harm from AI-based systems through normative statements such as AI should not be weaponized, or any AI must have an off switch. These serve as good starting points. How we can make sure that AI systems do not make the world less safe? Here it is essential to think about asymmetrical effects of technological development: Vulnerable populations suffer the negative effects of technologies much deeper, and in higher numbers. AI, much like any other technology, can cause physical, psychological, social and/or financial harm. Consider AI-driven bank loan decisions or school districts firing good school teachers because an AI system identified them as ineffective (as has happened in the US). In the light of such issues we ask, do benefits outweigh the risks? And if not, should we develop AI at all? Or should we at least consider slowing down development?

The notion of harm is central to the discussion of the possibilities and risks of AI from the human rights point of view.34 In discussions of risks and harms, the framework of human rights can often provide moral legitimacy to the expressed concerns. Yet how to address the notion of harms in practice is a more difficult question. Some ethicist have argued that the answers lie in turning

34 Latonero, "Governing Artificial Intelligence."

to the framework of virtue ethics, which guides engineers towards cultivating ethical wisdom by paying attention to the moral salience of routine options and decisions.35 However, what constitutes acting virtuously is more difficult to define in practice.

At the ANE hackathon, attention was devoted to cross-border cooperation between private and public sector and internally within organizations, from the CEO to the engineer. Participants argued that avoiding harm required unified policies at an organizational level, which would enable developers to ask themselves the "right questions" before and throughout design and development of applications. Governmental institutions, regulations and standards could also inform the decisions made to avoid misinterpretation of codes and to minimize unintentional harm. Most importantly, however, the hackathon participants called for more spaces and forums for discussion that would enable engineers and their organizations to clearly define the rights and wrongs in AI implementations. If standards are to be followed, they should be defined collectively.

ADDRESSING BIAS

The problem of bias is probably the most common concern with respect to implementations of autonomous systems. Algorithms are increasingly used to guide decisions by human experts, including judges, doctors, and managers. Researchers and policymakers, however, worry that these systems might inadvertently exacerbate societal biases. Some claim that AI is robust against external manipulation, meaning human emotional manipulation and this being an advantage especially in areas where particular pernicious human biases are rife.36 However, the advent of adversarial models (GANs) in particular makes this claim no longer viable given the ability of one form of AI implementation to essentially fool another through forms of manipulation invisible to human observers.37

Concern with biases stems from a democratic commitment to perpetuating just and fair societies. Where biases become embedded and reproduced by AI technologies, some of these may adversely affect particular vulnerable populations in ways that perpetuate pre-existing inequalities. As the reality of such consequences became obvious, engineers have responded with the development of myriad of competing mathematical definitions for what it means for an algorithm to be fair. However, nearly all of the prominent definitions of fairness are limited to formal specifications, which require precise definitions of concepts that are primarily determined socially. Thus such definitions reproduce subtle shortcomings that can lead to serious adverse consequences when used implemented technically as objective solutions (Corbett-Davies & Goel).38 The problem with bias is that not all biases ought to be eradicated. There are plenty of

35 Shilton, "Values and Ethics in Human-Computer Interaction"; Vallor et al., "An Introduction to Software Engineering Ethics."

36 Bostrom and Yudkowsky, "The Ethics of Artificial Intelligence."

37 Fawzi, Fawzi, and Frossard, "Analysis of Classifiers' Robustness to Adversarial Perturbations."

38 Corbett-Davies and Goel, "Defining and Designing Fair Algorithms."

very useful biases that guide human behavior every day. For example, most of us have a bias against grabbing obviously very hot things with bare hands or against jumping off great heights without a parachute. Arguably, such biases are crucial to our survival. There are other biases, however, that we as a society want to guard against. Racial and gender biases are two common examples that are difficult but vitally necessary to address. In developing autonomous systems, careful considerations of assumptions and personal biases is key as these can guide engineer's decisions subtly resulting in systems that codify bias into practice.

One of the most persistent current discussions is the use of biased training data for many AI implementations. Even with AI implementations intended to be deployed broadly in society we see many recurrent problems that evidence problematic biases. For example, in 2016 Microsoft announced that they have developed an AI to judge human beauty and so they will hold a beauty contest judged by robots. Unfortunately, results displayed a suspicious tendency towards equating lighter toned skin with beauty. More recently Amazon had to announce that they will be retiring to an automated human resources system that downgraded any CV with the word "women's (soccer team, debate club, etc.)" on it as not qualified for a technical job. A few weeks earlier Amazon had to deal with a public relations upheaval when the facial recognition system they were marketing toward polic departments misclassified African American US senators as criminals in a test run.39

Although the examples above are unfortunate they are also relatively common where problems with AI and algorithms are concerned. There are many reasons for why automated systems continue to produce such problematic output. Some of the source of these biases is the training datasets that developers used for these models. Typically the data sources used are either gleaned from public sources or capitalize on pre-existing information. The problem of reliance on historical data to build models (using historical data for training datasets) with the resulting biases encoded in these data is that the systems trained such will reproduce these biases with surprising consistency. This, however, is not a new problem, but something that was an open question since the initial uses of statistical analysis for calculations of loan risks in the 40s in the US. Yet these long-standing problems have been made more acute by technological advances. It would be a mistake to think that these issues are present in the context of AI alone – human decision making is just as prone to many of the same biases. However addressing these issues in the context of AI may help our efforts toward more democratic and just societies.

At the hackathon, engineers expressed concerns over the consideration that bias is an inherent yet unintentional property of an automated system, identifying input data and (clustering) algorithms as the main source of bias. As bias is hard or impossible to remove, the proposed solutions below put emphasis on awareness and auditing methods.

39 Levin, "A Beauty Contest Was Judged by AI and the Robots Didn't like Dark Skin"; Lee, "Amazon Scrapped 'sexist AI' Tool"; Singer, "Amazon's Facial Recognition Wrongly Identifies 28 Lawmakers, A.C.L.U. Says."

Revealing biases: In situations where bias is not desired it would be beneficial to make sure any bias is revealed and, if needed, dealt with it. It is important to acknowledge the narrow scope of technical development and outcome. As one of the hackathon particpants noted: "The tech guys (we engineers) in industry do not always know where and how the data is collected and pre-processed and also towards the end result, not whe data on the outcome is utilized as there are sales/marketing departments and also some deployment where the full scope of system is not known." How might engineers be able to detect potential biases in the datasets they use where data provenance is difficult to establish and the full scope of the system is obscured? The larger and more complex the organization, the more acute such problems are likely to become.

Auditing biases: Many participants commented that it ought to be possible to audit systems repeatedly by an external, neutral entity (such as for example a "model testing institute"), tested on carefully selected data and granted approval only in case no major biases are detected. As another engineer commented: "For AI systems used in public governance, public health care, and public education for example, evaluation by an independent entity might be a requirement. Furthermore, model design criteria, the models themselves (unless privacy issues prevent that) and the results of model evaluation should be made public." Many participants clearly distinguished systems that by virtue of being integrated into the core of society must have significantly more oversight from systems that are oriented merely towards advertising or consumer good production and purchasing. In the former case, detecting biases came to seem like a task that is too important to be left to individual engineers. This is in part because identifying and deciding which biases ought not to be allowed in government, healthcare or educational systems is part of governance decisions rather than purely technical ones.

Questions were raised on whether it should be acceptable to produce systems with known biases and what could be done to mitigate such issues. Such questions are clearly issues of ethics and moral judgement where individuals, no matter how virtuous, do need support from their communities to develop a stance. What biases are acceptable and which might the Nordic society want to work to avoid? Just how much effort might be a reasonable investment would of course depend on the priorities and moral objectives of the society in question.

Opportunities for addressing pressing issues

The IEEE statement on Ethically Aligned Design articulates general principles that apply to the development of all types of AI – autonomous and intelligent systems (A/IS) regardless of whether they are physical robots or software systems.

AUTONOMOUS/INTELLIGENT SYSTEMS MUST:

1. Embody the highest ideals of human beneficence as a superset of Human Rights.

2. Prioritize benefits to humanity and the natural environment from the use of A/IS. Note that these should not be at odds — one depends on the other. Prioritizing human well-being does not mean degrading the environment.

3. Mitigate risks and negative impacts, including misuse, as A/IS evolve as socio-technical systems. In particular by ensuring A/IS are accountable and transparent.

These general principles are supported by responsibilities for engineers and other stakeholders that are discussed in many other similar documents. These include the insistence on assessing priorities and ensuring that human interests prevail over those central to institutions and commercial actors. Following the precepts of human-centered computing, many newly developed professional codes of conduct such as the Association of Computing Machinery, include a concern for putting people at the center of technology design and focusing on human-centered design and engineers. Such as focus, some argue is crucial for the development of public trust in AI systems. Trust can be earned over time and via natural interaction modalities, but can be easily undermined through careless data processing or incomprehensible decisions affecting people's lives. Many documents acknowledge that developing and maintaining trust in AI technologies requires a system of best practices that can guide the safe and ethical development and management of AI, a carefully thought out alignment with social norms and values, algorithmic accountability, compliance with existing legislation and policy, and protection of privacy and personal information.

In fact, privacy is seen as one of the central concerns given the capabilities of AI systems to collect and quickly process immense amounts of data. Many AI technologies enable the collection, monitoring, and exchange of personal information quickly, inexpensively, and often without the knowledge of the people affected. Therefore, there is an effort to ensure that computing professionals become conversant in the various definitions and forms of privacy. They should understand the rights and responsibilities associated with the collection and use of personal information. The EU general data protection regulation (EU GDPR) requires the use of data protection by design (DPbD) approaches in the development of any data intensive systems.

Given the extensive discussions of the potential harm that can be caused by AI systems, addressing harm is an obvious topic of concern. Many advocate that extraordinary care should be taken to identify and mitigate potential risks in machine learning systems. A system for which future risks cannot be reliably predicted requires frequent reassessment of risk as the system evolves in use, or it should not be deployed. Any issues that might result in major risk must be reported to appropriate parties. AI systems should include explanation-based collateral systems or roll-back of decisions so direct consequences can be undone.

One of the concerns with autonomous and self-learning algorithms is their use in the development of autonomous weapons. Here the principle of Meaningful Human Control (MHC) over individual attacks is a term coined by the NGO Article 36 in order to express the core element that is challenged by the movement towards greater autonomy in weapons systems. This principle requires deployment of human judgment meaningfully in utilisation of autonomous weapons and other critical systems.

Developing data-intensive AI systems of course can create many opportunities for harm, starting with the problems of privacy concern given the speed and breadth of data collection that is possible with AI systems. This is especially a concern in situations where human rights are a particular issue. As such, special attention should be paid to vulnerable people, such as people who due to their political, economic, social or health reasons are particularly vulnerable to profiling that may adversely affect their self-determination and control or expose them to discrimination or stigmatisation. Paying attention to the vulnerable also involves working actively to reduce bias in the development of self-learning algorithms.

Set legal limits to classification and determination can enable affected publics to be aware that they are dealing with a smart machine. This is crucial especially when dealing with disadvantaged and vulnerable populations. As AI implementations can potentially tip the world towards entrenchment of past and perhaps outdated sentiments given the reliance of these technologies on historical data, addressing fairness and accountability becomes ever more important. This requires formulating new models of fair distribution and benefit sharing in accordance with the economic transformations caused by automation, digitalization and AI. It also requires ensuring accessibility to core AI technologies, and facilitating training in STEM and digital disciplines. Further, these principles call for increased vigilance over processes that undermine social cohesion, give rise to radical individualism, jeopardize, inhibit or influence political decision making, infringe on the freedom of expression and the right to receive and impart information without interference.

While the considerations above are, no doubt, important - what is an engineer to do if they were to observe any of the above mentioned problems? When identified, it should be made possible to report signs of system risks that might result in harm. Leaders should prioritize the mitigation of the risks identified and take steps to reduce potential harm. In cases where these steps are not taken it may be necessary to "blow the whistle" to reduce potential harm. To aid in this, the design of systems should include appropriate opportunities for feedback, relevant explanations, and appeal.

These and other considerations of how to address emergent problems have been extensively discussed in many current documents as well as throughout the ANE hackathon, where participants debated the necessity and feasibility of many of the proposed solutions. These debates formed a basis for an initial set of guidelines for engineering in practice as well recommendations for government response.

ETHICS IN ENGINEERING EDUCATION?

What does it mean to be a responsible engineer? How do engineers come to know what is and is not responsible behavior and what are their responsibilities in the first place? It is clear that the first encounter with these issues must come during education. The practices learned and internalized in educational programs will then continue to evolve throughout professional life. As Google argues, engineers must: "responsibly share AI knowledge by publishing educational materials, best practices, and research that enable more people to develop useful AI applications."40 Beyond enabling more people to develop AI applications, many have raised concerns about enabling people to understand existing AI applications without the need for extensive background in computer science and computational methods. What changes are needed in existing approaches to engineering education at different levels? What is missing and what needs to be addressed? Working engineers are best positioned to begin answering these questions.

Many engineers working with AI whether in startups or in mature companies are struggling because, as one engineer explained to us during the ANE hackathon: "We don't yet know what is expected from the people who design, develop and use the AI systems - accountabilities and responsibilities are not clearly defined." In other words, the decision making about what is "good and responsible behavior" does not yet have real precedents or pre-existing experience to guide it. Even if engineers are attempting to be responsible, what constitutes responsibility in practice remains a complicated question with many unknowns. For example, one group of engineers wondered whether it is at all possible to have an unbiased training dataset and how to spot bias if it is present. They were certainly aware of the importance of considering the training data but were much less certain about what to look for and what might constitute harmful bias.

40 Pichai, "AI at Google: Our Principles."

As one of the solutions, many engineers argued for the importance of coming together in collaboration with their stakeholders to develop ideas about what responsibility means in this context. What relations must be considered, what obligations must be taken on and enacted are important decisions precisely because building new systems requires acknowledgment and renegotiation of interrelations of responsibilities. At the same time, the shifting standards and new regulations continuously shape and structure what sorts of decisions might be made. Who gets to make these decisions and whose values might guide these are also pertinent questions. In a globalized economy, the notion of "good" does not work as a local concept and yet "good" is always contextual, so who is responsible for moments when "good" pivots and takes on negative consequences?41 Such discussions lead to many engineers considering what changes need to happen in engineering education from the very beginning so that the necessary conversations begin earlier and perhaps young engineers can develop more ethically informed practices.

Questions about AI and education extend beyond educating engineers specifically. The Finnish ministry of economic affairs and employment states that: "There is a need for an artificial intelligence literacy, that is, the basic understanding of how things will function in the age of artificial intelligence." The question here is what kind of literacy is necessary more broadly in society, what basic concepts must everyone know and is it possible to achieve this?

INSTITUTIONAL RESPONSIBILITIES

In the discussion of responsibilities who must take these on? Trust in government and expectations of ethical behavior from corporate actors are particularly strong features of the Nordic context. Beyond individual responsibilities who must take on the new responsibilities and what might these be? What ought to be the role of professional organizations such as the ANE or national trade unions with respect to supporting the efforts of engineers in acting responsibly? What are the obligations of workplaces where engineers perform their duties? What might these entities need to change and how? What are the obligations of governments with respect to ethics and AI?

41 Shklovski, "Responsibility in IoT: What Does It Mean to 'Do Good'?"

Recommendations and Guidelines

While engineers and their organizations will need to shoulder much of the growing responsibilities in the design and implementation of AI systems, the relevant governing bodies of the Nordic countries and at EU level must acknowledge their own responsibilities and opportunities for action. Where specific implementations of particular ethical engineering conduct in practice is best left to companies and the engineers themselves, issues such as the necessary changes in education, implementation of new forms of legislation and regulation remains the purview of governance activities at the national, and regional level. As such, we present a set of policy recommendations to consider.

POLICY RECOMMENDATIONS

1. There is a need to anchor discussions on the political level and to advance the public understanding on AI. This could be accomplished through the creation of a platform - a meeting space that would engage decision makers, business, academia, civil society and professionals including engineers to come up with stable and transparent solutions for AI through joint discussions.

It is clear that addressing potential issues of broad implementation of AI technologies demands government action and oversight. However, the particular problems that autonomous systems pose involve significant technical components and require high levels of technical expertise in order to develop solutions and regulatory proposals that can support and foster innovation while addressing potential concerns. The question here is how do we exploit AI technologies for their usefulness while avoiding exploitation of its users. Participatory governance approaches are deeply embedded into the fabric of the Nordic culture and offer avenues for engaging diverse forms of expertise of necessary depth as part of the government deliberation processes. However, developing new forms of such engagement will require political will and financial investment.

2. Education for ethical considerations and guidelines is often insufficient in the technical disciplines and throughout work-life. This needs to be addressed through changes in educational goals and priorities for technical subjects as well as through provision of relevant opportunities for lifelong learning.

Discussions of ethical issues in implementations of AI systems require a sophisticated vocabulary and at least familiarity with existing ethical frameworks and their limitations. Efforts to augment or even reform technical education is already happening at different levels from rudimentary introductions of ethics content modules into technical courses to the development of new workshops and courses. Much of this development is happening either through grass-roots efforts or with the support of civil society and commercial actors. For these changes to become systemic however, it is clear that government support and oversight are crucial.

3. Development of an appeal process with governmental oversight is crucial. Such a process must enable individuals and organizations to address the AI behaviour and decisions that they find potentially harmful.
One of the biggest concerns with respect to AI technologies is that if things do go wrong (as often already have), how might the people affected be able to act in response in ways that respect their agency and afford them dignity. Responsible organizations must work to establish clear chains of responsibility and accountability throughout the life of any technical system and to support engagement with the affected publics. However, such processes need not only government blessing and support, but also some structured oversight to ensure trust and clarity of consequences.

4. There is a need for shaping regulation and legislation to govern issues related to AI that formalises relevant responsibility and defines accountabilities.
It is clear that those that design and develop technologies must be held responsible for their decisions and actions, but this can only be upheld if we recognize that both individual engineers and the organizations that they are part of, are embedded in the social, political and economic systems of societies. In the end, it is crucial to formalise responsibility and to define who is accountable when things do things and negative consequences arise.

5. Engineers, policy makers, civil society and the general public need spaces for sustaining a living dialogue around issues of AI and ethics. These need to be facilitated and supported through funding and other forms of support.
The need for deliberation about what constitutes ethics with respect to AI and how to determine the rights and wrongs of the outcomes of AI implementations is acute. Such deliberation spaces should provide opportunities for professionals and decision-makers from different backgrounds and with different expertise to meet and debate. Supporting such deliberation and dialogue must not fall exclusively on the shoulders of the relevant stakeholders themselves but requires sustained political support and government investment to be sustained.

GUIDELINES FOR ENGINEERS AND THEIR INSTITUTIONS

While this document is intended to speak directly to engineers themselves, we must acknowledge that two things are necessary:
1. Individual engineers must have the education and training to be able to take on their responsibilities.

2. Individual engineers must have support of the organizations and institutions they work with and for in order to be able to take on the responsibilities and emergent issues effectively.

The guidelines below have organically emerged from discussions with engineers as well as from an overview of other efforts to address the issues of AI and ethics. These are not exclusively for individual engineers to follow, because ethical development of AI will not come about only as a result of individuals taking on particular types of ethical responsibility. There are plenty of guidelines for what constitutes ethical conduct for engineers and some of the guidelines below can be taken on board by individuals and organizations alike as additions to those that are already in existence in the Nordic countries. However, many of the guidelines are oriented towards organizational practices rather than individual responsibility, because efforts towards ethical practices need strong institutional backing to be effective and therefore organizational commitment is a requirement for addressing ethics in AI. We present these guidelines with an understanding that their implementation will require effort and commitment on the part of the individual engineers and of their organizations together.

GUIDELINES OF ETHICAL CONDUCT FOR AI DEVELOPMENT AND IMPLEMENTATION

1. Create spaces for discussion of the issues around AI and ethics. These need to be facilitated and supported by both workplaces and civil society organizations.

As the deliberation processes in determining the definitions of AI and ethics illustrate, discussions that intend to address issues of AI and ethics need to be encouraged, and given enough space and time to develop before they can be used in practice. Although the framing of ANE's hackathon has been centred on ANE members, it is essential that these debates are made possible in other locations and configurations, among experts, stakeholders, as well as much more broadly, as members of society.

2. Invest into and develop tools that enable ethical discussions, questions and decision making throughout the design process and not only at the beginning and the end.

The task of making ethical decisions is does not only happen at the beginning and end of a process, and it is not merely an extra requirement to be fulfilled. This is because ethical issues may involve questioning the very basis of the completed work or the produced artifact, in essence rendering the whole project incompatible with any possible ethical framing. Ethical discussion and evaluation in other words is not something that can be merely bolted on at the beginning or the end of a project as "check" to make sure whatever is produced qualifies as "ethical". Rather, an ethical approach must be integrated into design and development as a method for guiding the project throughout, and not as a set of deliverables to be fulfilled.

3. Establish a set of internal standards and checklists tackling ethical issues in AI development such as ensuring meaningful human control.

While it is important for ethical considerations to be presented throughout design and development, in practice it is difficult to achieve that ethical issues always remain as a priority when the project conditions change due to external factors. A set of internal standards and checklists to tackle ethical issues can help alleviate the challenge of always remaining engaged to ethical principles, as it provides an easy-to-use tool for framing issues in ways relevant to the project or task at hand. For example, a checklist that includes an item noting the importance of ensuring meaningful human control allows the participants in the project to repeatedly pose the question in light of new features added to the project. Although meaningful human control as a concept has been colonized for considerations of autonomous weapons, here we use it as a much more broad idea of ensuring some level of human control in engagements with any AI system.

4. Support and facilitate internal reporting of risk and violations and clear action in response.

Ethical guidelines themselves can become a formality, a box ticking exercise that does not translate to change within the project. To prevent such outcomes, and to allow for proiect participants to raise concerns they encounter within their projects, pathways for reporting risk and violations need to be established within institutions, alongside clear actions and consequences when violations occur.

5. Establish internal training programs for staff to deepen an understanding of ethics and to develop skills for ethical reflection, debate and recognition of biases.

Internal training programmes can allow for participants to test their ideas and to provide them with opportunities to form their own modes of ethical engagement. For example, where biases may be unavoidable they can be managed with training that enables people to recognize and address these. Such initiatives also demonstrate a willingness on the part of the institution that their employees spend time and resources in developing their thinking around ethical issues. Ethical and moral reasoning require training and usable frameworks as well.

6. Pay special attention to potential biases encoded in system development, training data and model performance, especially those that may affect the most vulnerable.

Given the attention currently being paid to the importance of training data used in the development of any AI system that relies on algorithmic data processing, it is crucial to ensure that these considerations are addressed in practice. Learning to think not only in terms of averages but also in terms of edge cases would help to consider the impact on the most vulnerable given design specifications. This in turn can lead to creative decisions and better solutions.

7. Develop ways for accepting organizational responsibility for potential harm, for example, by establishing ways to address the harm inflicted on others by the AI systems that the organization has produced.

How might those affected by AI systems respond should things go not according to plan? If things do go wrong when an AI system is implemented and people come to harm (whether or not they engage with the system directly), the question is who must take the responsibility for negative outcomes. At the moment, it is not clear how people who are negatively affected might need to act, whom should they contact and who might respond. Such uncertainty foster mistrust and doubts about the utility of AI systems, resulting in push-back rather than acceptance. Establishing a clear chain of responsibility and accountability throughout the life of any technological system is crucial to maintaining trust.

8. Establish an internal ethical review process that democratizes company decision-making by involving more internal actors.

An internal ethical review not only provides a stable structure for how ongoing initiatives within an organisation are evaluated, but also allows concerned employees, who are not always able to voice their concerns, to participate in decision-making. This deliberation process can also allow for the circulation of ideas and expertise throughout the organisation.

9. Work to increase transparency not only in the decisions leading to design and development of AI systems, but also in organizational chains of responsibility.

By making the organisational chain of responsibility visible, organisations would display that they are committed to establishing accountability mechanisms in the face of potential harms. This is not to say that increasing transparency in decision making, design, or development is undesirable, but rather that the two processes are complementary, and the lack of either may adversely affect trust in the organisation.

10. In working towards transparency, maintain awareness that transparency has its own ethical pitfalls and limits.

While transparency is a worthy goal for organisations that design and develop AI technologies, it cannot be the only means by which ethical engagements are formed. Transparency is only one part of the equation, and this report has deliberated many other concerns that must also be accounted for. Another issue with transparency is that when it is not accompanied by mechanisms of accountability, for example when algorithms are employed to make discriminatory decisions, it can become exceptionally difficult to affect meaningful change.

References

Agrawal, Ajay, Joshua Gans, and Avi Goldfarb. “The Simple Economics of Machine Intelligence.” Harvard Business Review, November 17, 2016. https://hbr.org/2016/11/the-simple-economics-of-machine-intelligence.

Agre, Philip. “Toward a Critical Technical Practice: Lesson Learned in Trying to Reform AI.” In Social Science, Technical Systems, and Cooperative Work: Bridging the Great Divide, edited by Geoffrey C. Bowker, Lee Gasser, Susan Leigh Star, and Bill Turner, 131–58. Erlbaum, 1997.

Albu, Oana Brindusa, and Mikkel Flyverbom. “Organizational Transparency: Conceptualizations, Conditions, and Consequences.” Business & Society, July 13, 2016. https://doi.org/10.1177/0007650316659851.

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 (March 1, 2018): 973–89. https://doi.org/10.1177/1461444816676645.

“Background | The Iron Ring.” Accessed November 7, 2018. http://ironring.ca/background.php.

Bird, Sarah, Solon Barocas, Kate Crawford, Fernando Diaz, and Hanna Wallach. “Exploring or Exploiting? Social and Ethical Implications of Autonomous Experimentation in AI.” SSRN Scholarly Paper. Rochester, NY: Social Science Research Network, October 2, 2016. https://papers.ssrn.com/abstract=2846909.

Bostrom, Nick, and Eliezer Yudkowsky. “The Ethics of Artificial Intelligence.” In The Cambridge Handbook of Artificial Intelligence, by Keith Frankish and William M. Ramsey, 316–34. Cambridge University Press, 2014.

Brey, Philip A. E. “Anticipating Ethical Issues in Emerging IT.” Ethics and Information Technology 14, no. 4 (2012): 305–17. https://doi.org/10.1007/s10676-012-9293-y.

BS 8611: 2016. “Robots and Robotic Devices: Guide to the Ethical Design and Application of Robots and Robotic Systems.” London: British Standards Institution., n.d. https://shop.bsigroup.com/ProductDetail?pid=000000000030320089.

Burrell, Jenna. “How the Machine ‘Thinks’: Understanding Opacity in Machine Learning Algorithms.” Big Data & Society 3, no. 1 (June 1, 2016): 2053951715622512. https://doi.org/10.1177/2053951715622512.

Cath, Corinne, Sandra Wachter, Brent Mittelstadt, Mariarosaria Taddeo, and Luciano Floridi. “Artificial Intelligence and the ‘Good Society’: The US, EU, and UK Approach.” Science and Engineering Ethics 24, no. 2 (2018): 505–528.

Corbett-Davies, Sam, and Sharad Goel. “Defining and Designing Fair Algorithms.” Fair ML. Accessed November 7, 2018. https://policylab.stanford.edu/fairML/.

Crawford, Kate, and Ryan Calo. “There Is a Blind Spot in AI Research.” Nature News 538, no. 7625 (October 20, 2016): 311. https://doi.org/10.1038/538311a.

European Group on Ethics in Science and New Technologies. “Artificial Intelligence, Robotics and ‘Autonomous’ Systems.” Directorate-General for Research and Innovation., 2018. https://ec.europa.eu/research/ege/pdf/ege_ai_statement_2018.pdf.

Fawzi, Alhussein, Omar Fawzi, and Pascal Frossard. “Analysis of Classifiers’ Robustness to Adversarial Perturbations.” Machine Learning 107, no. 3 (March 1, 2018): 481–508. https://doi.org/10.1007/s10994-017-5663-3.

Howard, Ayanna, and Jason Borenstein. “The Ugly Truth About Ourselves and Our Robot Creations: The Problem of Bias and Social Inequity.” Science and Engineering Ethics 24, no. 5 (October 1, 2018): 1521–36. https://doi.org/10.1007/s11948-017-9975-2.

IEEE Standards Association. “Ethically Aligned Design: A Vision for Prioritizing Human Wellbeing with Artificial Intelligence and Autonomous Systems.,” 2016. https://standards.ieee.org/industry-connections/ec/auto-sys-form.html.

Irani, Lilly. “The Hidden Faces of Automation.” XRDS: Crossroads, The ACM Magazine for Students 23, no. 2 (December 15, 2016): 34–37. https://doi.org/10.1145/3014390.

Khalil, Omar E. M. “Artificial Decision-Making and Artificial Ethics: A Management Concern.” Journal of Business Ethics 12, no. 4 (April 1, 1993): 313–21. https://doi.org/10.1007/BF01666535.

Latonero, Mark. “Governing Artificial Intelligence: UPHOLDING HUMAN RIGHTS & DIGNITY,” n.d.

Lee, Dave. “Amazon Scrapped ‘sexist AI’ Tool,” October 10, 2018, sec. Technology. https://www.bbc.com/news/technology-45809919.

Levin, Sam. “A Beauty Contest Was Judged by AI and the Robots Didn’t like Dark Skin.” The Guardian, September 8, 2016, sec. Technology. https://www.theguardian.com/technology/2016/sep/08/artificial-intelligence-beauty-contest-doesnt-like-black-people.

Pichai, Sundar. “AI at Google: Our Principles,” 2018, 7.

Rossi, Francesca. “Artificial Intelligence: Potential Benefits and Ethical Considerations.” Europe: European Parliament, 2016. http://www.europarl.europa.eu/RegData/etudes/BRIE/2016/571380/IPOL_BRI(2016)571380_EN.pdf.

Shilton, Katie. “Values and Ethics in Human-Computer Interaction.” Foundations and Trends® Human–Computer Interaction 12, no. 2 (2018): 107–171.

Shklovski, Irina. “Responsibility in IoT: What Does It Mean to ‘Do Good’?” ThingsCon: The State of Responsible IoT (blog), 2018. https://medium.com/the-state-of-responsible-iot-2018/responsibility-in-iot-what-does-it-mean-to-do-good-dd31bff2691a.

Silverstone, Roger. “Proper Distance: Toward an Ethics for Cyberspace.” In Digital Media Revisited: Theoretical and Conceptual Innovations in Digital Domains, edited by Gunnar Liestøl, Andrew Morrison, and Terje Rasmussen, 469–90. MIT Press, 2004.

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.

Singer, Natasha. “Amazon’s Facial Recognition Wrongly Identifies 28 Lawmakers, A.C.L.U. Says.” The New York Times, July 27, 2018, sec. Technology. https://www.nytimes.com/2018/07/26/technology/amazon-aclu-facial-recognition-congress.html.

The Norwegian Society of Engineers and Technologists. “Code of Ethics for Engineers and Technologists.,” (n.d). https://www.nito.no/organisasjon/om-nito/etikk-i-nito/.

Vallor, S., A. Narayanan, B. Regnell, C. Jones, and R. B. Skipper. “An Introduction to Software Engineering Ethics.” Ethics & International Affairs 25, no. 03 (2013).

Wagner, Ben. “Ethics as an Escape from Regulation: From Ethics-Washing to Ethics-Shopping?,” 2018.

Yampolskiy, Roman V. “Artificial Intelligence Safety Engineering: Why Machine Ethics Is a Wrong Approach.” In Philosophy and Theory of Artificial Intelligence, edited by Vincent C. Müller, 389–96. Studies in Applied Philosophy, Epistemology and Rational Ethics. Berlin, Heidelberg: Springer Berlin Heidelberg, 2013. https://doi.org/10.1007/978-3-642-31674-6_29.

Further reading

ACM CODE OF ETHICS AND PROFESSIONAL CONDUCT

Gotterbarn, D. W., Brinkman, B., Flick, C., Kirkpatrick, M. S., Miller, K., Vazansky, K., & Wolf, M. J. (2018). ACM Code of Ethics and Professional Conduct. Retrieved from: https://www.acm.org/code-of-ethics

The code is designed to inspire and guide the ethical conduct of all computing professionals, including current and aspiring practitioners, instructors, students, influencers, and anyone who uses computing technology in an impactful way. It includes principles formulated as statements of responsibility, based on the understanding that the public good is always the primary consideration, and guidelines, which provide explanations to assist computing professionals in understanding and applying the principle.

BRITISH STANDARDS INSTITUTION'S GUIDE TO THE ETHICAL DESIGN AND APPLICATION OF ROBOTS AND ROBOTIC SYSTEMS

BS 8611: 2016. (2016). Robots and robotic devices: Guide to the ethical design and application of robots and robotic systems. London: British Standards Institution. Retrieved from: https://shop.bsigroup.com/ProductDetail?pid=000000000030320089

BS 8611 gives guidelines for the identification of potential ethical harm arising from the growing number of robots and autonomous systems being used in everyday life.

The standard also provides additional guidelines to eliminate or reduce the risks associated with these ethical hazards to an acceptable level. The standard covers safe design, protective measures and information for the design and application of robots.

IEEE ETHICALLY ALIGNED DESIGN

IEEE Standards Association. (2016). Ethically Aligned Design: A Vision for Prioritizing Human Wellbeing with Artificial Intelligence and Autonomous Systems. Retrieved from: https://standards.ieee.org/industry-connections/ec/auto-sys-form.html

The discussion document from IEEE Initiative for Ethical Considerations in Artificial Intelligence and Autonomous Systems represents the collective input of global thought leaders in the fields of AI, robotics, law and ethics, philosophy, and policy from the realms of academia, science, government, and corporate sectors, providing insights and recommendations and a key reference for the work of AI/AS technologists in the coming years.

STANDARDIZING ETHICAL DESIGN FOR ARTIFICIAL INTELLIGENCE AND AUTONOMOUS SYSTEMS

Bryson, J., & Winfield, A. (2017). Standardizing ethical design for artificial intelligence and autonomous systems. Computer, 50(5), 116-119. Retrieved from: https://ieeexplore.ieee.org/document/7924235/

Part of IEEE Initiative for Ethical Considerations in Artificial Intelligence and Autonomous Systems, the article sets out to identify issues in the development of autonomous systems, focusing on transparency. The aim to develop a standard that sets out measurable, testable levels of transparency to assess an Autonomous System objectively and determine compliance. The standard will provide AS designers with a toolkit for self-assessing transparency as well as recommendations for how to address shortcomings or transparency hazards

NATIONAL SOCIETY OF PROFESSIONAL ENGINEERS, CODE OF ETHICS FOR ENGINEERS

NSPE Executive Committee. (2007). NSPE code of ethics for engineers. National Society of Professional Engineers. Retrieved from: https://www.nspe.org/resources/ethics/code-ethics

NSPE Code of ethics holds that the services provided by engineers require honesty, impartiality, fairness, and equity, and must be dedicated to the protection of the public health, safety, and welfare.

GOOGLE’S STATEMENT ON AI AND ETHICS

Pichai, S. (2018). AI at Google: our principles. Retrieved from: https://www.blog.google/technology/ai/ai-principles/

Google’s CEO Sundar Pichai lays down the principles for the company’s future development of AI, calling for thoughtful leadership in the area, scientifically rigorous and multidisciplinary approaches, and knowledge sharing

ARTIFICIAL INTELLIGENCE: POTENTIAL BENEFITS AND ETHICAL CONSIDERATIONS

Rossi, F. (2016). Artificial intelligence: Potential benefits and ethical considerations. Europe: European Parliament. Retrieved from http://www.europarl.europa.eu/RegData/etudes/BRIE/2016/571380/IPOL_BRI(2016)571380_EN.pdf

IBM’s briefing for the European Parliament’s policy department Citizens’ Rights and Constitutional Affairs focuses on the societal benefits of AI and the need to make sure that they follow the same ethical principles, moral values, professional codes, and social norms that we humans would follow in the same scenario. Research and educational efforts, as well as carefully designed regulations, must be put in place to achieve this goal.

EUROPEAN GROUP ON ETHICS IN SCIENCE AND NEW TECHNOLOGIES - STATEMENT ON ARTIFICIAL INTELLIGENCE, ROBOTICS AND 'AUTONOMOUS' SYSTEMS

European Group on Ethics in Science and New Technologies. (2018). Artificial Intelligence, Robotics and 'Autonomous' Systems. European Commission. Directorate-General for Research and Innovation. Retrieved from: https://ec.europa.eu/research/ege/pdf/ege_ai_statement_2018.pdf

EGE statement on artificial intelligence (AI), robotics and autonomous systems criticises the current 'patchwork of disparate initiatives' in Europe that try to tackle the social, legal and ethical questions that AI has generated, calling instead for the establishment of a structured framework.

OTHER REFERENCES IN THIS DOCUMENT:

Committee on Professional Ethics of the American Statistical Association. (2018). Ethical Guidelines for Statistical Practice. Retrieved from:
http://www.amstat.org/ASA/Your-Career/Ethical-Guidelines-for-Statistical-Practice.aspx

FINLAND'S AGE OF ARTIFICIAL INTELLIGENCE

Steering Group of the Artificial Intelligence Programme. (2017). Finland's age of artificial intelligence. Turning Finland into a leading country in the application of artificial intelligence. Objective and recommendations for measures. Ministry of Economic Affairs and Employment, Helsinki. Retrieved from:
http://julkaisut.valtioneuvosto.fi/bitstream/handle/10024/160391/TEMrap_47_2017_verkkojulkaisu.pdf

The Swedish Association of Graduate Engineers. (n.d). Code of honour. Retreived from:
https://www.sverigesingenjorer.se/Om-forbundet/Sa-tycker-vi/hederskodex/

The Norwegian Society of Engineers and Technologists. (n.d.). Code of ethics for engineers and technologists. Retrieved from https://www.nito.no/organisasjon/om-nito/etikk-i-nito/

SIRI Commission. (n.d.) AI scenarier. Etiske overvejelser & anbefalinger. Retrieved from:
https://ida.dk/sites/default/files/ai_-_etik_-_sirikommissionen_rapport.pdf

Association of Chartered Engineers in Iceland (VFÍ). (n.d.). Code of ethics. Retrieved from:
https://www.vfi.is/um-vfi/log-og-reglur/

Icelandic Institute for Intelligent Machines. (2018). Ethics Policy.
http://www.iiim.is/2015/08/ethics-policy/

CONTACT INFORMATION

Inese Podgaiska
ANE Secretary General
Phone: +4529743960
E-mail: ipo@ida.dk

Irina Shklovski
Associate Professor, IT University
Phone: +45 7218 5363
E-mail: irsh@itu.dk