02.10.2020

Opinion on research ethics and artificial intelligence

The National Committee for Research Ethics in Science and Technology (NENT) hereby submits an opinion on research ethics and artificial intelligence (AI).

It is a stated goal of all superpowers to become leaders in AI, and this field of research is rapidly developing. AI is already having an impact on most areas of society. At the same time, prominent researchers and business leaders have expressed concern about this development, particularly in relation to self-learning systems that not only replace routine actions, but radically change and expand the human sphere of action. The further development of AI technology and the uncertainty associated with the consequences it will have for people and society requires reflection that researchers must also relate to. This forms the background for NENT's opinion.

Introduction

This opinion identifies and describes the special challenges manifested in connection with AI research, and the research ethics issues this gives rise to. The *Guidelines for research ethics in science and technology* (NENT 2016) emphasise the independent responsibility for the role research plays in the development of society. This opinion specifically details how research's social responsibility should be understood in light of the challenges associated with AI research.

In accordance with *The Research Ethics Act* (the Act on the Organization of Research Ethical Work of April 28, 2017) researchers and research institutions have an independent responsibility to ensure that the research they carry out is ethically responsible. NENT is a professionally autonomous body that advises researchers and authorities on research ethics issues. The ambition of this opinion is more specifically to facilitate high - quality and responsible AI research in Norway. The main target group is researchers, research institutions and other contributors who define the guiding principles of or are involved in AI research. The opinion should be viewed in the context of other parts of the Norwegian National Research Ethics Committees' efforts to identify research ethics challenges associated with digitalisation and big data (NESH, A Guide to Internet Research Ethics, 2019; FEK, Big Data Report, *to be published* in 2020).

Summary

1. Human dignity

Researchers and research institutions must ensure that AI systems are structured in a way that respect human dignity, the right of individuals to self-determination and democratic rights. Similarly, researchers have a responsibility to consider the expected or potential impact on individuals, animals, the environment and society, and must facilitate fair and ethical use of the systems.

2. Responsibility

Researchers who develop and design AI systems are in a position to guide the decisions that the systems make and the actions they perform. The scientific communities therefore have particular responsibility for these decisions and actions. When researchers conduct

commissioned research or plan to commercialise research results, they should work with external stakeholders to assess the risk associated with further use of the research.

3. Transparency

Transparency, i.e. the ability to identify the sources of the data used and generated by the systems, as well as how the systems make decisions, is critical to ensuring fairness and trust when decisions are made automatically. Where transparency is lacking, researchers should point this out and justify it. AI research should aim to produce 'glass boxes', i.e. systems that can be inspected.

4. Research dissemination

Ensuring a balanced discussion on the risks and possibilities associated with AI can be a challenge. As a society, we should not be naive; we need to be aware of potential risks and possibilities. AI falling into the wrong hands is one such risk. Researchers should contribute to an informed social discourse in which discussions are based on realistic assumptions. Researchers have a particular responsibility to present a balanced picture of risks and possibilities, since they have the most knowledge on how far the development has progressed.

5. Uncertainty

From a research ethics perspective, it is essential to communicate the uncertainty associated with research. The development of AI is characterised by fundamental uncertainty and unpredictability. NENT therefore sees a need for systematic studies of the risks associated with the development of AI. Authorities and others who fund research should facilitate interdisciplinarity in research, thereby acknowledging its unpredictability, and minimise uncertainty where possible. Ethics should be introduced as a separate subject in curricula aimed at future developers of AI.

6. Broad involvement

Researchers also have a responsibility to communicate the risks resulting from their research findings. What risks and opportunities are emphasised, may also depend on the ethical perspective and the values and interests applied. Those who are most affected by the decisions made must be given a voice in the decision-making process. Authorities and

research institutions should facilitate broad public involvement regarding the purpose of research, the direction of research programs and the use of research.

7. Privacy and respect for individuals

Basic privacy principles enshrined in the data protection legislation must be followed. From a research ethics perspective, obtaining consent is one of the main rules for using personal data in research. Even if anonymized data is used in the analyses, compilations with other data can still reveal sensitive information or reveal individuals, thus constituting personal data. Obtaining and using data that includes personal data may challenge the requirement for informed consent. When collecting and compiling large volumes of data, there is a considerable risk of personal data being used for purposes unfamiliar to us (because the purpose is also unknown to the researcher at the time of the data collection), and which we may not want.

When assessing research ethics related to information and consent, researchers have a responsibility to assess the data's accessibility in the public sphere, the sensitivity of the data, the vulnerability of the participants and the interaction and consequences of the research (NESH 2019).

8. Quality

In AI research, there may be reasons to ask critical questions about the quality, truthfulness and relevance of the data, because the data sources are not always known, and metadata can be incomplete or uncertain. Material bias, analysis tool characteristics, and human interpretations increase the potential for erroneous inference. This can lead to uncertainty in interpretations and decisions based on AI. In order to ensure verifiability and quality, researchers and research institutions should therefore facilitate open and widely available data sources.

9. Fair access to data

As a general rule, research, including data and results, should be made available to all. NENT sees a risk that much of the research efforts in AI avoid the transparency requirements that apply to research otherwise (as stipulated in the FAIR principles), citing the need for competitive secrecy. Authorities and research institutions should facilitate public access to data. They should ensure openness concerning ownership of technology, infrastructure and

data, what research areas are prioritised and why, and who may be expected to benefit from the research.

The opinion is structured as follows: following a brief account of the method NENT has used, we detail the characteristics of present-day AI research. We then examine the challenges posed by AI in society and the research sector, and the implications for research ethics that this entails.

Method

NENT has been in contact with relevant AI research communities through a consultation round from June to August 2018 and a workshop in February 2019. The purpose was to map out the key possibilities and research ethics challenges identified by the Norwegian research communities. NENT requested responses to the following:

1. Do you conduct research that you would say is within the field of AI, and which research communities are involved in this?
2. What do you think are positive possibilities for AI? Do you see any worrying aspects of the development of AI?
3. What research ethics issues and challenges (including issues relating to consequences for society) exist in AI research, including your own?
4. What should researchers and research institutions take responsibility for in the sustainable development of the field, i.e. a development that promotes innovation and knowledge, whilst ensuring the attention to research ethical issues?

NENT received 13 responses and, together with the workshop organised by the committee, this has provided an important backdrop for the mapping of what the research communities themselves consider to be the key challenges of AI. In addition, NENT has looked at what has been done in this area in Norway and abroad. Some of the most important international documents to date, and which NENT has reviewed, include:

- The Asilomar AI Principles, The Future of Life Institute, 2017
- The General Principles in *Ethically Aligned Design (V2)*, IEEE, 2017
- Report on Robotics Ethics, COMEST, 2017
- Towards a Digital Ethics, EDPS, 2017
- Code of Ethics and Professional Conduct, ACM (2018)
- The Ethical Principles in *Statement on Artificial Intelligence,* EGE, 2018

- Guidelines for trustworthy AI, High-Level Expert Group on AI, European Commission, 2019.
- Principles on AI, OECD, 2019

The scope of the documents differs; some target the area of AI as a whole, while others address related and partly overlapping areas, such as autonomous and intelligent systems (IEEE), robotics (COMEST) and digital technology in general (EDPS). Almost all documents have been the subject of broad-based discussion rounds with participants from research, industry, political bodies and other stakeholders. In the business sector, companies such as IBM, Microsoft and Google's DeepMind have developed their own ethical guidelines and joined forces to develop broad initiatives such as 'Partnership on AI' and 'OpenAI'.

In Norway the Norwegian Data Protection Authority has published a report on AI and privacy, while the Norwegian Board of Technology has presented the report 'Artificial Intelligence: Opportunities, Challenges and a Plan for Norway', which also addresses ethical aspects of AI (both published in 2018). The government has decided to draw up a national strategy for AI, where ethical perspectives will also be central. This is in line with developments in most European countries, where such strategies have already been developed or will be developed by the end of 2020.

NENT believes it is important to note the growing number of reports on ethics and AI, which indicates that there is a strong awareness of the need for ethical reflection. The research ethics aspects of AI, on the other hand, are less developed, and NENT has identified a need to explore these implications of the development of the technology. In NENT's assessments of AI research, research ethics guidelines, in particular the *Guidelines for Research Ethics in Science and Technology (NENT 2016)*, provide a framework. Ethical considerations in AI may relate to issues that arise during the actual development of the technology, and also to issues that arise when it is used. According to the definition in the guidelines, research ethics also includes the latter, i.e. ethical reflection is required beyond the research process itself.

Characteristics of AI

AI has existed as a research field in computer science and informatics for around 60 years, and the aim of the research is to make AI a reality. Broadly defined, AI involves the development of technology that enable computers to be integrated into technological systems

in ways that make them behave 'intelligently', i.e. capable of solving cognitive and physical tasks previously reserved for humans. This is done through computer programs that use data and algorithms to train or optimise a system to give a desired response – either in a development phase or after the system is implemented. One such example is the speech recognition function on mobile phones; the system gradually becomes more effective as the user corrects errors.

A core and distinctive characteristic of AI is therefore that such systems *emulate, replace and expand human intelligent action, and human decision-making and cognition*. Experimental technology capable of simulating human emotions also exists. The degree of complexity of the human actions that are replaced or automated can vary considerably. At one end of the scale we find '*deterministic systems*' that replace routine actions, and at the other end we have *'cognitive' or 'autonomous systems'* that replace actions we associate with human cognition, reasoning and learning (COMEST, 2017, p.7). This distinction also includes different degrees of automated systems, i.e. systems that can operate 'independently' or 'autonomously', without human intervention. This scale ranges from systems that are remotely controlled by an operator to fully autonomous systems that make all decisions themselves, based on a task they have been given by an operator. The first type of system can largely be pre-programmed, while the second group of systems will usually need to continue learning while in use. Another related distinction is between specific and general AI. On the one hand, AI can perform fairly simple services, such as customising recommendations by various web services, playing chess or recognising faces. On the other hand, however, there may be systems capable of performing many different tasks. General AI in its most developed form is often referred to as superintelligence, often associated with a notion of fully developed 'conscious' machines. While the development of specific AI has been rapid in recent years, few advances have been made in general AI. Whether we will be able to develop general AI and superintelligence at all, and if so when, is the subject of much debate.

Another important characteristic of AI is countless *areas of application and a huge potential for change*. Some of the development is the result of human interaction, and some relates to interaction between the technical systems. The technology is thus also characterised by *unpredictability*; it is difficult or impossible to predict the impact of the technology on individuals, society and the environment. In many countries, the implementation of the

technology is already well underway in areas such as healthcare, the judiciary, transport and communications, while other areas are under development. For example, in the field of health today, AI is better at interpreting images of potential melanoma than many dermatologists. At the University of Agder and the University of Oslo, research is being conducted into the use of AI in psychological healthcare. In the field of health, there is assumed to be countless applications, for example in complex surgical procedures and patient care. There is also an expectation that AI will have the same kind of major impact on other sectors, such as the labour market, the economy, politics and culture. AI may also be able to influence us as individuals, for example as we have seen with the importance of the mobile phone for interpersonal relationships and for how we think and feel.

The development of specific intelligence in recent years points to a third characteristic of AI, namely the *generation of big data* that may contain personal data. AI in the form of machine learning (particularly 'deep learning') is driven by big data and computing power. At the same time, AI is a source of new big data.

Research ethics challenges

The following review of nine research ethics challenges can be structured into three groups that reflect the characteristics of AI as described above. Different research fields and applications of AI naturally raise different challenges, and the research ethics norms that are discussed may be under intense pressure in some areas, but in other areas will not be significantly affected. The points are a summary of the aspects that NENT, in collaboration with the scientific communities, has found special reason to focus on in connection with AI research. Supplementary reflections are needed for each individual research project in order to expand on these points.

A) Responsibility for the development and use of autonomous systems

The first set of considerations is related to the goal of AI technology to emulate, replace and expand human intelligent action, and human decision-making and cognition.

1**. Human dignity**

Human dignity entails understanding human beings as ends in themselves and never as a means alone. This limits the definitions and categorisations of humans based on algorithms and autonomous systems. The development and use of AI have a fundamental effect on

human dignity; for example, through the development of smart aids in everyday life, AI can help promote individual self-realisation and human dignity. On the other hand, AI can also constitute a threat to human dignity. One such example is the use of AI to monitor the population within a system of social control and sanctioning, such as that used in China. The development of algorithms for use in 'profiling', i.e. techniques used to analyse, predict and potentially influence future preferences and behaviour patterns is another example. When individuals are not treated as ends in themselves, but as aggregates of data collected for the purpose of, for example, optimising the administrative interaction with them, it could be argued that this is not compatible with respect for human dignity (EDPS 2017, pp. 16-17). In recent years, we have also seen several examples of AI systems being used to manipulate democratic processes, such as the US presidential election in 2016 and Brexit. It is important that the use of AI does not become a threat to democratic rights (EGE 2018, pp. 17-18).

AI research can be developed or used to promote individuals' self-determination, human dignity and democratic rights, but its choice of themes, in relation to any research participants and in the dissemination and use of results, can also threaten these values. Research ethics contain requirements to prevent the misuse of research as well as prevent participation in such activity:

> Where scientific and technological development can be misused to undermine the right of self-determination and human dignity and the democratic rights of individuals, researchers must strive to prevent and refrain from taking part in any such misuse of research. Researchers have an independent responsibility to ensure that research benefits society, directly or indirectly, and to minimise risk (NENT 2016, guideline 1).

This requires good procedures that can ensure that research ethics are considered right from the start of the research process. 'Ethics by design' is a term that refers to the need for a proactive approach to ensuring a high standard of responsible AI research. It is based on the more well-established concept of 'data protection by design', which we find in, for example, the data protection legislation. 'Ethics by design' is a broader concept that refers to the need to structure AI systems in a way that safeguards human dignity and protects privacy, including the expected or potential impacts on individuals and society, and the need to facilitate fair and ethical use of such systems. 'Ethics by design' thus also involves an assessment of the context of the system being developed.

2. Responsibility

Issues related to human decision-making control and allocation of responsibility are fundamental in the context of AI, and such issues are even more relevant when developing and using adaptive and autonomous systems. In general, it can be argued that the more adaptive and autonomous an AI system is, the more difficult it will be to control it and to assign responsibility.

When a Boeing 737 crashed in Ethiopia in 2018 and 157 people lost their lives, it was because the pilots were unable to override a serious bug in the MCAS anti-braking software. This plane crash raises questions about whether human control over the systems was possible, who is responsible when decisions are automated, and whether a computer program can be responsible for an accident.

When the movements of a machine are controlled by a computer program, it is possible to allocate the responsibility for the outcomes. However, in the case of autonomous systems with a deep learning capacity, behavioural or decision-making processes cannot be programmed in the same way as for deterministic systems. In the debates about autonomous weapons systems and autonomous cars, the concept of 'meaningful human control' has been a central theme in the discussions on who is responsible. The principle is formulated as a prerequisite for legitimacy, and implies that it is humans, not machines or their algorithms, who ultimately must have control and be morally responsible. 'Meaningful' often refers to whether a human being will have sufficient time to intervene and override the machine. In a strict sense, 'human control' can mean that a human operator monitors the system and makes all critical decisions. In a weaker sense, it means that the system is designed to function reliably and predictably, without involving a human being in each decision.

More specifically, this raises the research ethics issue of what researchers can and should have control over and take responsibility for in the development and use of adaptive and autonomous systems. A distinction must be made between the responsibility for AI research on the one hand and the responsibility for further use of the research results on the other. Researchers who develop and design more or less autonomous AI systems are in a position to guide the decisions and actions taken by the systems. The scientific communities therefore have a particular responsibility.

Researchers also have a shared responsibility for the use of research (see NENT 2016, guidelines 1-3 and 8-9). In connection with commissioned research or the planned commercialisation of research results, researchers should therefore work with external actors to assess the risks associated with the further use of the research.

3. Transparency

The term 'black box problem' refers to the various challenges that can make AI systems and algorithms so complex that we do not understand how they arrived at their given answer. The term thus implies a lack of openness with regard to an essential component of the decision-making process; we may know about the data entered and know the answer, but we are unable to determine how the data gave rise to the answer. The emergence of big data increases this problem; the volume of data that is used can be so vast that maintaining an overview of it is impossible.

The black box problem can also be about a lack of openness surrounding the conditions and parameters on which the work of the machines is based. In the development of autonomous systems, mathematical techniques are used to transfer value selection to algorithms. *DeepMind*, for example, has developed algorithms by building on basic assumptions from rational choice theory, which have some way to go before achieving general acceptance.

In the responses received by NENT, *openness* was a principle that almost everyone highlighted in their formulation of research ethics challenges in AI.

> Many of the algorithms used in AI are poorly understood, and the applications of AI appear as products of a 'black box'. As we do not fully understand these algorithms, the applications of AI solutions may lead to unforeseen side effects. These side effects may potentially be dangerous, e.g., if AI-based networks are used for medical decisions (Simula).

The lack of openness can give rise to biased decisions being made and certain values and perspectives being omitted without us being aware of it, which in turn may lead to a lack of confidence in the decisions made. In this context, it makes sense to distinguish between two types of 'black boxes': *First*, this may be an *involuntary* black box, where the lack

of openness is because the nature of the model is such that it cannot be scrutinised. *Second*, it may be a matter of voluntary shielding for security reasons, or of commercial players not believing they are served by publishing which algorithms they use. In both of these 'black box' cases, the problem is that the machines hide why they make the choices they do, and that those responsible are also unable to explain the reasons for those decisions. This is particularly problematic when the algorithms increasingly make choices that have consequences for individuals and society, for example in the judiciary, the financial sector or the education sector. The Norwegian Tax Administration uses, for instance, predictive analytics to select which tax returns should be checked for possible tax fraud.

In addition to openness, other, partly overlapping principles, such as transparency or explicability, are also consistently emphasised as central components of responsible AI development. From a research ethics perspective, openness is partly about being open and explicit about the choice of data sources, development processes and stakeholders. *Scrutability* refers more specifically to the ability to describe how decisions are made by the systems, as well as the origin of the data used and generated by the system. Under the personal data legislation, scrutability is also crucial to ensuring openness and confidence in automated decisions, such as 'profiling', where automated machines analyse or predict certain characteristics of individuals or groups (such as aspects of their finances, health and behaviour).

The performance of a system that has an 'involuntary black box' is often better than that of a more scrutable/transparent system, which will result in a trade-off between quality and transparency. However, it is not necessarily a question of having to choose one over the other. Researchers should in any case identify and justify such trade-offs, and AI research should aim to produce 'glass boxes', i.e. systems that can be scrutinised. However, placing a spotlight on understanding the mechanisms of a tool can take the focus away from how and in what context it is used. As well as requiring an explanation for any operation performed by a machine, NENT believes that responsibility should also be taken for the assumptions, choices and uncertainties associated with a system (see section 5 below).

4. Research dissemination

Attempts to identify long-term consequences of AI research and its use have great uncertainty attached to them, which makes them seem speculative. In the responses to NENT, several

parties point out that, in the public sphere, discussions of AI are often of a dystopian nature. The problems seem to be fabricated, however, as they are often linked to the development of general AI and fully autonomous systems, while current developments are mainly related to specific AI. On the one hand, it can be asserted that it is unlikely that the most pessimistic predictions will come true, and that attention should therefore be directed towards the advantages and disadvantages of existing tools and data. On the other hand, the long-term consequences may cause great damage, which is why we should have a long-term perspective. The risk often mentioned in connection with AI is "singularity", which refers to the point where AI reaches a human level of understanding and no longer requires human interaction. Google's Director of Engineering Ray Kurzweil (2017) has claimed that we will reach this point by 2029. He also believes that existing technology represents the beginning of this point. He talks about our growing dependency on our phones, where the next step will be to connect technology directly to our brain. This is a view that is shared by several leading voices in the field. At the same time, there is a lack of consensus, including among researchers, on whether general AI at the human level is possible at all, and if so, *when*.

Researchers must contribute to an informed social discourse, so that the public debate is based on realistic assumptions. However, it is difficult to have a balanced discussion about the risks and possibilities associated with AI. Discussions of AI in the public arena are occasionally characterised by a 'moral panic' which highlights scenarios associated with the possibility of a superintelligence. On the other hand, these possibilities may be exaggerated, while the dangers inherent in the technology are under-communicated by parties looking for research and development funding. As a society, we should not be naive, and we should be aware of possible risks and possibilities, for example the risk of AI falling into the wrong hands. Researchers have a particular responsibility to present a balanced picture of risks and possibilities, since they have the most knowledge on how far the development has progressed.

B) Societal consequences and the social responsibility of research

The second set of challenges discussed here is related to the innumerable applications of AI, and the enormous potential to create change. The development partly takes place in collaboration with people, and partly through interaction between the technical systems. The

technology is therefore also characterised by unpredictability; it is difficult or impossible to predict the effects of the technology on individuals, society and the environment.

5. Uncertainty

While AI offers great possibilities, there is considerable uncertainty attached to it. No matter how good our intentions are, its use and consequences can be counter-productive or negative because we do not have enough knowledge about how the technology works or how it will be used. Like other enabling technologies, AI is therefore characterised by *unpredictability*; it is difficult or impossible to predict the effects of the technology on individuals, society and the environment. This is an issue that has been mentioned in several of the responses to NENT. The uncertainty associated with AI research is related to several of the following dimensions: a) the development and customisation of systems, including the quality of basic data; b) use of the systems and their consequences for individuals, animals, the environment and society; c) the values that are explicitly or implicitly built into the systems and how they affect outcomes or, in a larger context, individuals, animals, the environment and society; and d) the consequences of *not* developing the technology.

The lack of predictability associated with the development of the technology gives rise to a renowned dilemma that is common to AI and other enabling technologies, referred to as the 'Collingridge dilemma' (1980). The dilemma refers to how development can be difficult to control in an early phase because the full extent of the consequences is often unclear until society has adopted the knowledge and the technology. By then it is often too late for regulation, as it is difficult to hold back technology that has been developed or to withdraw technology that is already in use. The questions associated with unpredictability include the matter of whether we should focus on long-term and future consequences or on more immediate effects, the type of uncertainty, and how to handle this uncertainty. Both the EU and the Research Council of Norway have adopted the Responsible Research and Innovation (RRI) approach in an attempt to address such challenges. The literature on RRI often points at a forward-looking concept of responsibility. It is a matter of taking responsibility at an early stage of the research and ensuring that good decisions can be made going ahead, partly by identifying and assessing possible consequences, and partly by building an apparatus to deal with them.

According to the *Guidelines for Research Ethics in Science and Technology (NENT 2016)*, research must also communicate the degree of uncertainty in the research and evaluate the risk associated with the implications of the work:

> Researchers must clarify the degree of uncertainty in their research and evaluate the risk associated with the research findings.
> Researchers must clarify the degree of certainty and precision that characterises their research results. They must be particularly meticulous about clarifying the relative certainty and validity range of their findings. In addition to presenting knowledge critically and in context, researchers must strive to point out any risk and uncertainty factors that may have a bearing on the interpretation and possible applications of the research findings. Communicating the relative certainty and validity of knowledge is part of a researcher's ethical responsibility and effort to achieve objectivity. Where possible, researchers should also use appropriate methods for demonstrating the uncertainty of the research. Research institutions have an obligation to teach these methods to their employees and students (NENT 2016, guideline 8).

NENT sees a need for systematic studies of the risks associated with the development of AI. It is important that both researchers and political decision-makers acknowledge uncertain but possible consequences, and also *unidentified* unknowns; i.e. future consequences that have yet to be revealed. The authorities and research funding institutions should facilitate interdisciplinarity in research, in order to better acknowledge unpredictability and minimise uncertainty when this is possible.

6. Broad involvement

Many of the risks and possibilities associated with the development of AI are uncertain and it is difficult to identify them, but some can already be pinpointed and confirmed. Several of the challenges associated with AI are also shared by other enabling technologies, like biotechnology and nanotechnology. Such technologies have extensive potential to change society through the opportunities they provide for the establishment of new links between different disciplines and activities. In the same way, the development of AI could provide such systems with innumerable areas of application in society. On the one hand, research can help solve the major challenges facing society in core areas like health, energy, the climate and safety. On the other hand, it may provide grounds for concern regarding the risk of possible misuse and unwanted consequences. Most of the responses to NENT highlight the

wealth of possibilities provided by AI. Many parties also mention possible negative consequences of the development, but the communities are generally optimistic, and the responses do not reflect many of the concerns presented internationally by researchers in recent years.

The University of Bergen writes the following:

> The nature of the theoretical research on AI facilitates applications in many fields. In mathematics and nuclear physics, the knowledge may have applications in, for example, health, but it may also have controversial applications like weapons development or purposes that clearly cause harm to individuals and society.

The Norwegian University of Life Sciences stresses the need for interdisciplinarity and a wide-ranging, inclusive discourse:

> The research should be interdisciplinary because AI may also involve humans and societal consequences, and there may be a need to establish ethical rules for the development of new technology.

In research ethics, there have been attempts to face the challenges associated with major societal consequences. The *Guidelines for Research Ethics in Science and Technology* (NENT 2016) underline the independent social responsibility of researchers.

> Research has an independent responsibility for the role it plays in social developments. Researchers and research institutions must contribute to the collective accumulation of knowledge and to resolving major challenges facing the global community (NENT 2016, guideline 1).

The first guideline implies that researchers must reflect critically on their role in the development of technology and society and explain it. NENT believes that it is important for scientific communities themselves to critically assess the visions behind AI research, and consider what are legitimate and more questionable purposes. The government's AI strategy will probably largely define the national vision for AI, and thus guide the direction of research. In many cases, the party commissioning the research will set the purpose, and

others will often determine its use. In such cases however, considerable responsibility may lie with the researcher, given that it is possible to influence why and how AI systems are developed.

Researchers are also responsible for communicating the risk inherent in research findings. The precautionary principle may apply when managing risk with scientific uncertainty attached to it. This has been worded as follows in the *Guidelines for Research Ethics in Science and Technology (NENT 2016)*:

> Researchers must strive to observe the precautionary principle.
> Where there is plausible, but uncertain knowledge to the effect that a technological application or a development of a research field may lead to ethically unacceptable consequences for health, society or the environment, the researchers in the field in question must strive to contribute knowledge that is relevant for observing the precautionary principle. This means that researchers must work together with other relevant parties in observing the precautionary principle. The precautionary principle is defined here as follows: 'When human activities may lead to morally unacceptable harm that is scientifically plausible but uncertain, actions shall be taken to avoid or diminish that harm.' This principle is important for a large part of science and technology research, and researchers have a shared responsibility for ensuring that evaluations are based on the precautionary principle and contribute to avoiding or diminishing harm (NENT 2016, guideline 9).

The precautionary principle does not apply when there is complete uncertainty, only when there is 'plausible, but uncertain knowledge'. There is great uncertainty in the field of AI, and there are also disagreements as to whether certain negative consequences will materialise in the future, especially in relation to fully autonomous systems. However, many of the consequences of the development of specific AI are already known, and in many cases the risks have been established satisfactorily, triggering the precautionary principle. The precautionary principle means that AI researchers must describe and communicate the risks associated with the development and use of AI in their fields of research. However, the 'ethically unacceptable consequences for health, society or the environment' that are emphasised will vary, depending on the ethical perspective, and the values and interests applied. Pressure from different parties may affect the algorithms, without any professional or political assessment having been made. For example, AI-based surveillance systems may be seen as both a benefit and a risk. From a defence and security perspective, surveillance may

be considered a benefit that prevents crime and warns society about possible dangers, but from a data protection perspective, this type of surveillance may also be considered a threat to personal integrity.

The parties that are affected the most by decisions must also be guaranteed a voice in decision-making processes. The authorities and research institutions should therefore enable broad public participation in a debate on the purpose, direction and use of research.

C) Big data

Together with computing power and algorithms, the development of specific AI has mainly been driven by big data, which can also include personal data. The third set of challenges that NENT believes must be addressed is related to big data in AI research. Big data gives rise to new challenges associated with data protection in research. Big data also raises other questions of research ethics which we will address in the following, including in terms of material bias, data quality, and ownership and access to data.

7. Privacy and respect for individuals

Protecting data that contains personal data may present its own challenges in terms of the development and use of AI. Even though anonymised data is used in analyses, data comparisons may still reveal sensitive data or identify individuals, and thus constitute personal data. Collection and use of data that includes personal data may conflict with the requirement of informed consent. When collecting and compiling large volumes of data, there is a particular risk of personal data being used in ways that are unfamiliar to us (because the purpose is also unknown to the researcher when it is collected), and which may not be what we are looking for.

Respect for individuals and groups involved in research in different ways or directly impacted, are regulated in i.e. the data protection policy (i.e. the EU GDPR and the Norwegian Personal Data Act, which supplements it). This policy is a guiding principle for researchers, but it cannot, in isolation, address the many challenges that researchers face when handling personal data. In Norway, the *Guidelines for Research Ethics in the Social Sciences, Humanities, Law and Theology* (NESH 2016) are the main tool used to elaborate on ethical responsibility towards subjects of research and other parties who are affected by research.

The implementation of the data protection rules in Norwegian law saw the introduction of a number of basic principles that must be followed in order for the processing of personal data to be legal. One of these is the principle of *data minimisation.* From this follows that one should not use more personal data than necessary to fulfil the purpose of the proposed processing must be used. In addition, the data must be adequate and relevant to the processing. People who work with AI may have difficulty limiting the amount of data to process, because development and use of AI usually requires large volumes of data in order to train the systems. In order to assess what is necessary, adequate and relevant, the researcher must be fully aware of the purpose of the processing desired.

Another basic principle of data protection is precisely that the data processing must be *limited to the purpose*. Personal data must not be processed without a legitimate, specific and explicitly stated purpose. A clear and specific purpose is one that is specifically described. As the specified purpose guides the fulfilment of various other data protection principles, vague and wide-ranging descriptions of the purpose are not permitted. For example, the personal data must be deleted when the purpose of the processing has been met. If the specified purpose is development of AI, it may be difficult to fulfil this requirement. However, the requirement of purpose limitation may be met by stating what type of AI will be developed, and the tasks that this system is expected to perform. It will nevertheless be difficult to determine whether the requirements regarding data minimisation and purpose limitation have been met in such situations. In its report on AI and data protection, the Norwegian Data Protection Authority notes that AI development should seek to limit the training data when starting up, and then expand the data set when there is a greater understanding of what is needed. In its statement regarding automated decision-making, the Article 29 Working Party, an advisory group in the EU, points to the importance of the data controller introducing procedures and systems that ensure that the personal data used is always correct and up-to-date.

The intention of the principles behind this regulation is that personal data must only be processed when necessary, in order to limit the encroachment on the privacy of individuals. It is important to keep this in mind when using large volumes of data to develop AI, and to consider whether the volume of data can be limited at no detriment to the purpose stipulated. If the volume of data cannot be limited, it must be possible to explain this choice in order to

show that the data minimisation principle has been met in the same way as the criteria of necessity, adequacy and relevance. In order to reuse the data, the developers must make sure that the secondary use is consistent with the original purpose.

The general rule in research ethics is that personal data must not be collected, processed or shared without informed consent. When data that has been collected for other purposes is reused in new and unexpected areas, the consent must be updated where this is possible. Challenges may also arise when new compilations of data include data that is, in principle, anonymised. When assessing information and consent in terms of research ethics, researchers must assess the public nature of the data, its sensitivity, the vulnerability of the participants, and the interaction and consequences of the research (NESH 2019).

8. Quality

In AI research, there may be particular reasons to ask critical questions about the quality, veracity and relevance of the data, because the data sources are not always known and metadata can be incomplete or uncertain. Material bias, analysis tool characteristics, and human interpretations increase the potential for erroneous inference and biased decisions. This can lead to uncertainty in interpretations and decisions based on AI. In recent years, we have seen several examples of how data can result in unreasonable decisions. When Amazon tried to establish an objective employment process in 2018 using AI, this resulted in decisions with a gender bias, because the data sets favoured men.

> AI must train using real data. Methods like deep learning work best with large volumes of data. This means that the quality of data is paramount, and it is not always good (Norwegian Computing Center).

NENT believes that in order to guarantee verifiability and quality, it is critical that researchers and research institutions make sure that data sources are open and universally available. At the same time, uncertainty factors and the limitations of research must be acknowledged and communicated.

9. Fair access to data

The development of AI technology may give a few individuals, companies or research groups a chance to dominate this arena.

> However, the greatest concern is that AI appears to be dominated by a few companies like Facebook, Google, Amazon and a few others. AI that truly works requires vast volumes of data and computing power. Companies like Facebook have enormous volumes of data and computing power that other players could not possibly match (CAIR, University of Agder).

NENT sees a risk that much of the research on AI will circumvent the requirement of openness that otherwise applies to research, as stipulated in the FAIR-principles (Wilkinson et al. 2016), for example, citing the need for competitive secrecy.

From the perspective of research ethics, it is critical to facilitate universal access to research, including data and results. As stated in the *Guidelines for Research Ethics in Science and Technology (NENT 2016)*, the requirements of openness in research ethics call for research results, methods and data to be shared and made public, both in order to facilitate quality assurance, maintain trust in research, and ensure that the results benefit society (see NENT 2016, guideline 3, 4 and 17).

The lack of data sharing is problematic for several reasons. First, if only researchers in a few privileged companies are able to analyse large data sets, it will be impossible for outsiders to reproduce and evaluate their results. Second, researchers with close ties to private companies may have motivations and interests that will influence what research is given priority, and the results of the research. Companies that perform evaluations or research do so for commercial purposes. Researchers with the appropriate competence and access to the right data can help develop a better foundation of knowledge which in turn can benefit society more widely.

NENT believes that the authorities and research institutions should facilitate universal access to data. They should ensure openness on the ownership of technology, infrastructure and data, the research areas that are prioritised and why, and who may be expected to benefit from the research.

Conclusion

The nine points in this document are meant to provide a starting point for reflection, guidance, and discussion within the research communities. They have also been drawn up for

funders and facilitators of AI research, or those who use AI. Given the rapid pace of development in this field of research and the uncertainty attached to it, this document should be reviewed and revised after a period of time. NENT wants to continue the dialogue with the scientific communities regarding the research ethics challenges associated with AI research, thus facilitating ethically sound and responsible AI research in Norway.

References

European Data Protection Supervisor (EDPS), 2017. *Towards a Digital Ethics*.
European Group on Ethics in Science and New Technologies (EGE), 2018. *Statement on Artificial Intelligence, Robotics and 'Autonomous' Systems'*.

Norwegian Board of Technology, 2018. *Artificial Intelligence: Opportunities, Challenges and a Plan for Norway*.

The National Committee for Research Ethics in Science and Technology (NENT), 2016. *Guidelines for research ethics in science and technology*

The National Committee for Research Ethics in the Social Sciences and the Humanities (NESH), 2016. *Guidelines for Research Ethics in the Social Sciences, Humanities, Law and Theology.*

The National Committee for Research Ethics in the Social Sciences and the Humanities (NESH), 2019. *A Guide to Internet Research Ethics.*

The Norwegian Data Protection Authority, 2018. *Artificial intelligence and privacy.*

The Norwegian Personal Data Act
Lov 15. juni 2018 nr. 38 om behandling av personopplysninger (personopplysningsloven)

The Research Ethics Act
Lov 28. April 2017 nr. 23 om organisering av forskningsetisk arbeid (forskningsetikkloven)

Vissgren, Julie, 2017. «5 spådommer fra Googles fremtidsforsker Ray Kurzweil». InnoMag 17.10.2017: https://www.innomag.no/5-spadommer-fra-googles-fremtidsforsker-ray-kurzweil/

Wilkinson, M. D. et al., 2016. *The FAIR Guiding Principles for scientific datamanagement and stewardship*. Sci. Data 3:160018.

World Commission on the Ethics of Scientific Knowledge and Technology (COMEST), 2017. *Report of COMEST on robotics ethics.*