The Norwegian National
RESEARCH ETHICS
COMMITTEES

Statement on research ethics in artificial intelligence

NENT • The National Committee for Research Ethics in Science and Technology

Statement on research ethics in artificial intelligence
1st edition – November 2019
ISBN: 978-82-7682-100-0

Translation by Samtext.

Front cover:
An unmanned helicopter is being prepared for missions on the USS Coronado, a US battleship. Photo: Flickr/COMSEVENTHFLT, licence CC BY-SA 2.0

STATEMENT ON RESEARCH ETHICS IN ARTIFICIAL INTELLIGENCE

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

All global superpowers have a stated goal of becoming leaders in AI, and this field of research is developing rapidly. AI already affects most areas of our modern society. At the same time, prominent researchers and business leaders have expressed concern about this development, especially with regard to self-learning systems that are not only taking over routine actions, but radically changing and expanding people's scope of action. The further development of AI technology, and the uncertainty associated with what the consequences may be for people and society, require a considered approach, also on the part of researchers. This forms the background for NENT's statement.

Oslo, November 2019

NENT (2018–2021)
Øyvind Mikkelsen (Committee Chair), Michaela Aschan, Ingrid Bay-Larsen, Tone Druglitrø, Ole Andreas Engen, Hanne Pernille Guldbrandsen, Steinar Heldal, Kjellrun Hiis Hauge, Gorm Idar Johansen, Cecile Marie Mejdell, Rune Nydal, Jørn Paus, Ketil Skogen, Jim Tørresen, Lise Øvreås, and Helene Ingierd (Head of Secretariat).

Introduction

This report points out and describes the special challenges that arise in connection with research on AI, and the questions these challenges trigger in relation to research ethics. *Research ethics guidelines for science and technology* emphasise the independent responsibility of the research for the role it has in societal development, and this statement elaborates in particular on how the social responsibility of the research should be understood in light of the challenges raised by AI.

In accordance with the Act Concerning the Organisation of Work on Ethics and Integrity in Research (Research Ethics Act), researchers and research institutions have an independent responsibility to ensure that the research they conduct is ethically sound. NENT is a professionally independent body that provides advice to researchers and authorities on issues related to research ethics. More specifically, with this statement, NENT aims to facilitate constructive and responsible AI research in Norway. The statement is primarily directed at researchers, research institutions and other actors who set the premises for or are involved in AI research. The statement should be seen in connection with other aspects of the work conducted by the Norwegian National Research Ethics Committees to identify research ethical challenges in relation to digitalisation and big data (NESH, A Guide to Internet Research Ethics 2019; FEK, Report on Big Data, *to be published in* 2020).

Summary

1. Safeguard human dignity
Researchers and research institutions must ensure that AI systems are structured in a way that safeguards individuals' self-determination, human dignity and democratic rights. Likewise, researchers must consider the expected or possible impact on individuals, animals, the environment and society, and must facilitate the fair and ethical use of AI systems.

2. Assign responsibility
Researchers who develop and design AI systems can provide guidance on the decisions the systems make and the actions they perform. Consequently, academic communities have a special responsibility. For commissioned research or planned commercialisation of research results, researchers should collaborate with external actors to assess the risk of further use of their research.

3. Inspectability
Inspectability, i.e. the opportunity to identify the sources of the data used and generated by the systems, as well as how the systems make decisions, is crucial to ensure fairness and confidence when decisions are made automatically. Researchers should point out and justify any lack of inspectability, and AI research should aim to produce "glass boxes", i.e. systems that can be inspected.

4. Dissemination of research
It can be quite a challenge to ensure a balanced discussion about the risks and opportunities presented by AI. As a society, we should avoid naivety, while also being aware of possible opportunities and risks, such as AI falling into the wrong hands. Researchers should contribute to an informed public debate, so that society's assessments can be based on realistic assumptions. Researchers have a special responsibility for presenting a balanced view of risks and opportunities, as they have the best knowledge of how far development has come.

5. Acknowledge uncertainty
From the point of view of research ethics, it is essential to assess and 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 this research, thereby acknowledging its unpredictability and minimising uncertainty where possible. Ethics should be included as a subject in the education of future AI developers.

6. Ensure broad involvement
Researchers also have a responsibility to communicate the risks that derive from their research findings. Which risks and opportunities related to the technology are emphasised may also depend on the ethical perspective and the values and interests on which they are based. Those who will be most affected by the decisions that are made must be guaranteed a voice in decision-making processes. Authorities and research institutions should facilitate the broad involvement of citizens in discussions on the purpose of the research, the structure of the research programmes and the application of the research.

7. Ensure data protection and consideration of individuals
Fundamental principles of data protection, enshrined in data protection legislation, must be followed. From the point of view of research ethics, consent is a main rule when personal data is used in research. Even if anonymised data is used in analyses, comparing such data with other data might still reveal sensitive information or identify individuals. This anonymised data could therefore still end up constituting personal data. Collecting and using data that includes personal data may challenge the requirement for informed consent. When collecting and compiling large amounts of data, there is a particular risk that personal data may be used in ways of which we are unaware (because the purpose is also unknown to the researcher at the time of collection) and which we may not want.
In their research ethical assessments of information and consent, researchers have a responsibility to assess the degree of public access to the information, the sensitivity of the information, the vulnerability of those affected and the

research's impact and consequences (NESH 2019).

8. Quality assurance
In AI research, there may be particular reason to ask critical questions about the quality, truthfulness and relevance of the data, because we do not always know the sources of the data, and metadata may be absent or uncertain. Skewness in the material, the properties of the analysis tool and human interpretations all increase the chances of drawing erroneous conclusions. This provides a basis for uncertainty in relation to interpretations and decisions based on AI. In order to ensure verifiability and quality, researchers and research institutions should therefore facilitate making data sources open and publicly available.

9. Fair access to data
From the point of view of research ethics, it is essential to ensure that research, including data and results, is generally made available to everyone. In NENT's view, there is a risk of large parts of the research into AI evading the requirements for transparency that otherwise apply to research (as they are laid down in e.g. the FAIR principles), for example with reference to the need to keep a competitive advantage secret. Governments and research institutions should facilitate public access to data. They should ensure transparency about who will have ownership of technology, infrastructure and data, which research areas are being prioritised and why, and who can be expected to benefit from the research.

The present statement has the following structure: After a brief account of the method used by NENT, we elaborate on what characterises current AI research. Then, we ask what challenges AI represents for society and how research is conducted, and the research ethical implications of these challenges.

Method

NENT has been in dialogue with relevant academic environments involved in AI research through a consultative process conducted in June–August 2018 and at a workshop in February 2019. The purpose was to identify what Norwegian researchers see as the key opportunities and research ethical challenges.
NENT requested input on the following questions:
1. Do you conduct research that you would say is within the area of AI, and which research environments are involved in this?
2. What do you consider to be positive opportunities for AI? Do you see any worrying aspects in the development of AI?
3. What questions and challenges related to research ethics (including questions regarding the consequences for society) apply to AI research, including your own?
4. What responsibility should researchers and research institutes take to ensure the sustainable development of the field, i.e. development that promotes innovation and knowledge, while also ensuring that research ethical considerations are taken into account?

NENT received 13 responses, and together with the workshop arranged by the committee, these responses have constituted an important backdrop for identifying what the researchers themselves consider to be key challenges related to AI. In addition, NENT has investigated what has already been done in this area, both nationally and internationally. Some of the most important documents that have so far been published internationally, 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 documents have somewhat different areas of focus; some focus on the field of AI as a whole, while others deal with adjacent and partly overlapping fields, such as autonomous and intelligent systems (IEEE), robotics (COMEST) and digital technology in general (EDPS). Virtually all of the documents above have been the subject of broad rounds of input, with participants from the research community, industry, political bodies and other stakeholders. In the business world, companies such as IBM, Microsoft and Google's Deep Mind 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 the ethical aspects of AI (both published in 2018). The Norwegian government has decided to develop a national strategy for AI, where ethical perspectives will also be in focus. 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 increasing number of reports on ethics and AI, which indicates that there is a strong awareness of the need for ethical reflection on this technology. The research ethical aspects related to AI, on the other hand, have received little attention, and NENT sees a need to elaborate on these implications in relation to the development of this technology. In NENT's assessments of AI research, guidelines for research ethics, in particular *Guidelines for Research Ethics in Science and Technology*, constituted a framework. Ethical assessments related to AI can be partly about questions that arise during the actual development of the technology and partly about questions that arise from its further use. As defined in the guidelines, research ethics also includes the latter issue, i.e. it requires an ethical reflection beyond the research process itself.

Characteristics of AI

Research has been conducted in the field of AI for around 60 years, within computer technology and computer science, with the aim of making AI a reality. Broadly defined, AI comprises techniques designed to incorporate computers into technological systems in ways that make them behave "intelligently", i.e. that they become able to solve cognitive and physical tasks that have previously been reserved for humans. This is done through computer programs that use data and algorithms to train or optimise a system to produce a desired response – either in a development phase or after the system has been taken into use. One example is speech recognition on mobile phones; the system continuously improves as the user corrects errors.

A key and distinctive feature of AI is thus that such systems *mimic, replace and extend intelligent human action, human decision-making and judgment*. This technology also has the potential to identify and simulate human emotions in ways that make us feel that the machine has human qualities. At one end of the scale, we find "*deterministic systems*", which can take over routine actions; at the other end, we have "cognitive" or "complex and completely autonomous" systems, which can take over actions we associate with human judgement, reasoning and learning (COMEST, 2017, p. 7). This distinction also applies to different degrees of automated systems, i.e. systems that can operate "independently" or "autonomously", without human intervention. This scale extends from systems that are remotely controlled by an operator, to completely autonomous systems that make all decisions themselves, based on a task they have been given by an operator. The first type of systems will largely be possible to be pre-programmed, while the second group of systems will most often need to continue learning while in use. The degree of complexity concerns what a system must be able to perceive and perform. Another related distinction is between specific and general AI. On the one hand, AI can perform fairly simple services, such as customising recommendations provided by various online services, playing chess or recognising faces. On the other hand, however, there may be systems that are capable of performing many different tasks. General AI in its most developed form is usually referred to as superintelligence, and is often associated with a notion of fully developed, "conscious" machines. While

there has been rapid development in specific AI in recent years, little progress has been made in general AI. Whether we will be able to develop general AI and superintelligence at all, and if so, when it will happen, is a controversial question.

Another important feature is that AI has *countless applications and a huge potential to bring about change*. Development takes place partly in interaction with humans, and partly in interaction between the technical systems. The technology is thus also characterised by *unpredictability*; it is difficult or impossible to predict what effects this technology will have on individuals, society and the environment. In many countries, implementation of this technology is already well underway in areas such as healthcare, the judicial system, transport and communications, while other areas are under development. In the area of health, for instance, AI can today assess images of possible mole cancer better than many dermatologists. At the University of Agder and the University of Oslo, research is being done on how AI can be used in psychological health care. It is believed that there are countless applications related to health, such as in complex surgical procedures and patient care. The expectation is that AI will have a similar, major impact on other sectors, such as the labour market, the economy, politics and culture. AI will likely also influence us as individuals, similar to what we have seen in connection with the importance of mobile phones for interpersonal relationships and for how we think and feel.

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

Research ethical challenges

The following review of nine research ethical challenges can be structured in three blocks that reflect the characteristics of AI we have described above. Different research fields and applications of AI naturally raise different challenges, and the research ethical norms that are discussed may be under

strong pressure in some areas, but may not be significantly affected in others. The points are a summary of what NENT, in collaboration with academic environments, has found special reason to draw attention to in connection with AI research, and they must be elaborated with supplementary considerations for each individual research project.

A) Responsibility for the development and use of autonomous systems

The first set of assessments is related to the ambition of AI technology to mimic, replace and extend intelligent human action, and human decisions and assessments.

1. Safeguard human dignity

Human dignity means that human beings should be treated as an end in themselves and not as a means to something else. This sets limits for definitions and categorisations of people on the basis of algorithms and autonomous systems. The development and use of AI can affect human dignity in a fundamental way. On the one hand, AI can contribute to promoting individuals' self-realisation and human dignity, through the development of smart aids to assist them in their daily lives, for example. On the other hand, AI can also threaten human dignity. One example is the use of AI for surveillance within a system designed to exercise social control and sanction the population, such as in China. The development of algorithms for use in "profiling", i.e. techniques used to analyse, predict and possibly influence future preferences and behavioural patterns, is another example. When individuals are not treated as an end in themselves, but as aggregates of data collected to optimise administrative interaction with them, for example, it becomes questionable whether this can be reconciled with respect for human dignity (EDPS 2017, p. 16–17). In recent years, we have also seen several examples of AI systems being used to manipulate democratic processes, such as the 2016 US presidential election and Brexit. It is important to prevent the use of AI from becoming a threat to democratic rights (EGE 2018, p. 17–18).

AI research can be developed or used to promote individuals' self-determination, human dignity and democratic rights, but depending on the choice of research topics, in relation to potential research participants and in the dissemination and application of results, it can also threaten these values.

In research ethics, there is a requirement to prevent and refrain from taking part in the misuse of research:

> 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 (Guideline 1).

This requires good procedures that can ensure that research ethical assessments are made from the outset of the research process. "Ethics by design" is a concept that refers to the need for a proactive approach for ensuring a high standard of responsible AI research. It is based on the more well-established concept of "Data protection by design and by default", which we find in data protection legislation, for example. "Ethics by design" is broader, and points out that AI systems must be structured in a way that safeguards human dignity and privacy, including expected or possible effects on individuals and society, and facilitating the fair and ethical use of such systems. Ethics by design thus also involves an assessment of the context of the system being developed.

2. Assign responsibility

Issues related to the possibility of human control and assigning responsibility are fundamental in the context of AI, and such issues are even more relevant in the development and use of 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 the more difficult it will be to assign responsibility.

When a Boeing 737 crashed in Ethiopia in 2018 and 157 people lost their lives, it was because the pilots on board were unable to override a serious error in the software MCAS, an anti-brake system. This plane crash raises questions related to whether human control over the systems was possible, who is responsible when decision-making processes 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 may be possible to assign responsibility. However, in the case of autonomous systems with deep learning capacity, behavioural or decision-making processes cannot be programmed in the same way as for deterministic systems. In the debates on autonomous weapon systems and self-driving cars, the concept of "meaningful human control" has been central to the question of who is responsible. The principle is formulated as a prerequisite for legitimacy, and implies that it is people, not machines or their algorithms, who must ultimately have control and be morally responsible. "Meaningful" often refers to whether a person will have sufficient time to intervene and override the machine. "Human control" can, in a strict sense, mean that a human operator monitors the system and makes all critical decisions. In a milder sense, it indicates that the system is designed so that it works reliably and predictably, without a human being involved in every single decision.

More specifically, a question related to research ethics arises pertaining to what researchers can and should have control over and take responsibility for in the development and use of adaptive and autonomous systems. There must be a distinction between the responsibility for AI research on the one hand and the responsibility for the further use of research results on the other. Researchers who develop and design more or less autonomous AI systems can provide guidance on the decisions the systems make and the actions they perform. Academic environments therefore have a special responsibility. Researchers also have a co-responsibility for the use of the research (cf. Guidelines for Research Ethics in Science and Technology, 1–3 and 8–9). In connection with commissioned research or the planned commercialisation of research results, researchers should therefore collaborate with external stakeholders to assess the risks of the further use of their research.

3. Inspectability

The concept of the "black box problem" refers to the various challenges associated with the fact that AI systems and algorithms can be so complicated that we do not understand how they arrive at their answers. The "black box problem" thus implies a lack of transparency regarding an essential component of the decision-making process; we may know the data that has been entered – and we know the answer – but we cannot determine how the data led to the answer. The emergence of big data is exacerbating this

problem; the amount of data employed can be so massive that we have no chance of gaining a proper overview of the data.

The black box problem can also relate to a lack of transparency regarding the conditions and frameworks on which the machines base their output. In the development of autonomous systems, mathematical techniques are used to allow algorithms to make value choices. *Deep Mind* has, for example, developed algorithms based on basic assumptions from rational choice theory, which has received a significant amount of criticism.

In the consultative input provided to NENT, *transparency* was a principle that nearly everyone highlighted in their discussions of research ethical challenges in relation to 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).

A lack of transparency can give rise to discriminatory decisions and to certain values and perspectives being omitted without us being aware of it, which in turn can lead to a lack of confidence in the decisions that are made. In this context, it makes sense to distinguish between two types of "black boxes": *The first one* can be an *involuntarily* black box, where the lack of transparency is due to the model being of such a nature that it cannot be inspected. *The second one* may be a matter of voluntary shielding for security reasons or because commercial actors do not see the benefit of making public which algorithms they use. Common to both types of "black box" is that the machines hide why they make choices and those responsible cannot explain the background for these decisions. This is particularly problematic when the algorithms increasingly make choices that have consequences for individuals and society, such as in the judicial system, the financial sector or the education sector. For instance, the Norwegian Tax Administration uses predictive analysis to select which tax returns should be checked for possible cheating.

In addition to transparency, other, partly overlapping principles such as interpretability and explainability are also generally emphasised as the main concerns with respect to the responsible development of AI. From the point of view of research ethics, transparency means, among other things, being open and explicit about the choice of data sources, development processes and stakeholders. *Inspectability* more specifically denotes 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. According to the data protection legislation, inspectability is also crucial for ensuring transparency and confidence in automated decisions, such as "profiling", where machines automatically analyse or predict conditions in individuals or groups (e.g. related to conditions such as finances, health and behaviour).

A system that includes an "unintentional black box" will often be able to provide better performance than a more inspectable/transparent system, which will lead to a need to find a balance between quality and transparency. However, it is not necessarily a question of having to choose one over the other. Researchers should, in all cases, make visible and justify such decisions, and AI research should aim to produce "glass boxes", i.e. systems that can be inspected. At the same time as it is desirable to be transparent about the movements a machine makes, NENT believes that researchers have a responsibility to account for the assumptions, choices and uncertainties associated with a system (see section 5 below).

4. Dissemination of research

Attempts to identify the long-term consequences of AI research and its application are encumbered by great uncertainty, and may therefore seem speculative. In the consultative input provided to NENT, several respondents emphasise that AI is often portrayed in a dystopian light. These issues are perceived as contrived, as they are often related to the development of general AI and fully autonomous systems, while current development is mainly focused on specific AI. On the one hand, it can be argued that the probability of the most pessimistic predictions coming true is low, and that the main focus should therefore be on considering the opportunities and disadvantages of existing tools and data. On the other hand, the possible harm associated with long-term consequences is consierable, and this

suggests that we should have a long-term perspective. The risk that is often mentioned in connection with AI relates to "the singularity", which refers to the point in the development of civilization where AI reaches a human level of understanding, and is no longer dependent on human interaction. Google's Director of Engineering Ray Kurzweil has claimed that we will reach this point by 2029. He also believes that, with current technology, we are on the cusp of this point. He refers to the addiction we are developing to our phones, where the next step will be to connect technology directly to our brains (https://www.innomag.no/5-spadommer-fra-googles-fremtidsforsker-ray-kurzweil/ [article in Norwegian only]). This view is shared by several prominent figures in the field. At the same time, there is disagreement, also among researchers, as to whether human-level general AI is possible at all, and if so, *when*.

Researchers should contribute to informed public debate, so that society's assessments can be based on realistic assumptions. However, it can be quite a challenge to achieve a balanced discussion about the risks and opportunities associated with AI. The public portrayal of AI can sometimes have the character of a kind of "moral panic" that highlights scenarios related to the possibility of a superintelligence. On the other hand, these opportunities may be exaggerated, at the same time as the risks that come with technology may be under-communicated by those seeking funding for development and research. As a society, we should avoid naivety and be aware of possible opportunities and risks, such as AI falling into the wrong hands. Researchers have a special responsibility for presenting a balanced view of risks and opportunities, as they have the best knowledge of how far development has come.

B) Societal consequences and the social responsibility of research

The second set of challenges we will discuss is related to the fact that AI has countless areas of application and an enormous potential to generate change. Development takes place partly in interaction with us, and partly in interaction between the technical systems. The technology is thus also characterised by unpredictability; it is difficult or impossible to predict what effects this technology will have on individuals, society and the environment.

5. Recognise uncertainty

While AI offers great opportunities, we are facing considerable uncertainty. No matter how good our intentions, the use or consequences of AI could prove to be counterproductive or negative, as we do not have full knowledge of how the technology works or how it will be used. Like other enabling technologies, AI is thus characterised by *unpredictability*; it is difficult or impossible to predict what effects this technology will have on individuals, society and the environment. Several of the consultative inputs provided to NENT also touched on this issue. The uncertainty related to AI research is linked to the following dimensions: a) the development and adaptation of the systems, including the quality of raw data; b) the 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 – seen in a larger context – how they will impact individuals, animals, the environment and society; and d) the consequences of not developing the technology.

The unpredictability of technology development gives rise to a much-discussed challenge that AI shares with other enabling technologies, and which is referred to as the "Collingridge dilemma". This dilemma refers to how development can be difficult to control at an early stage, because the full extent of the consequences is often unclear before society has adopted the knowledge and technology. Then, it is often too late to regulate, as it has been proven that it is difficult to hold back technology that has been developed or to withdraw technology that has already been taken into use. The questions that are raised in connection with unpredictability include whether we should focus on long-term and future consequences or on more immediate effects; what kind of uncertainty is relevant, and how can we deal with this uncertainty? Both the EU and the Research Council of Norway have launched "Responsible Research and Innovation" (RRI) in an attempt to meet such challenges. In the literature on RRI, the answer has often been to point to a forward-looking concept of responsibility. It is about taking responsibility early in a research process and ensuring that good choices can be made in the further process, partly by anticipating and assessing possible consequences and partly by building an apparatus to mitigate them.

The *Guidelines for Research Ethics in Science and Technology* emphasise that researchers also have a responsibility to convey uncertainty in their own research and assess the risks associated with the implications of their own activities:

> 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 (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 policy makers recognise uncertain but possible consequences, and also *unknown* unknowns, i.e. future consequences we do not yet know. Authorities and research-funded institutions should facilitate interdisciplinary research, in order to better recognise unpredictability and minimize uncertainty where possible.

6. Ensure broad involvement

Many of the opportunities and risks associated with the development of AI are uncertain and difficult to identify, but some can already be identified and their probability assessed. Several of the challenges associated with AI are also relevant for other enabling technologies, such as biotechnology and nanotechnology. A common feature of these technologies is their broad potential to change society through the opportunities they provide to establish new connections between different disciplines and activities. In the same way, AI development could provide such systems with countless areas

of application in society. On the one hand, research can contribute to solving major societal challenges in core areas such as health, energy, climate and security. On the other hand, they may give rise to concerns about the risk of possible abuse and undesirable consequences. Most of the consultative input provided to NENT emphasises the great opportunities associated with AI. However, many respondents also point to possible negative consequences of AI development, but overall, the environments appear to be optimistic, and their input reflects to a lesser extent the concerns that have been raised internationally by multiple researchers in recent years.

The University of Bergen writes the following:

> Theoretical research within AI is of such a nature that it could have applications in many fields. As with mathematics and nuclear physics, knowledge can have applications in e.g. health, but may also contribute to controversial applications such as weapons development or for use for purposes that obviously cause harm to individuals and society.

The Norwegian University of Life Sciences (NMBU) emphasises the need for an interdisciplinary approach and a broad and engaging debate:

> The research should be interdisciplinary because AI can also relate to people and consequences for society, and there may be a need to establish ethical rules for the development of new technology.

In research ethics, there have been attempts to meet the challenges associated with societal consequences. The *Guidelines for Research Ethics in Science and Technology* emphasise the researchers' independent social responsibility:

> 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 (Guideline 1).

The first guideline entails that researchers must reflect critically on and

account for their own role in the development of technology and society. NENT believes it is important that academic communities themselves also critically assess the visions behind AI research and what should be deemed legitimate and less legitimate purposes. The government's AI strategy will probably go a long way in defining Norway's national vision for AI, and thus lay down guidelines for the direction of research. In many cases, the purpose will largely be predetermined by a client, and the further use will often be determined by others. In this space, researchers may still have a significant responsibility, to the extent that they have an opportunity to influence why and how AI systems are developed.

Researchers also have a responsibility to communicate the risks that derive from their research findings. The precautionary principle may be relevant with regard to the management of risk in relation to scientific uncertainty. This is formulated as follows in the *Guidelines for Research Ethics in Science and Technology*:

> 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 (Guideline 9).

The precautionary principle does not apply in cases of full uncertainty, only where there is "plausible but uncertain knowledge". In the field of AI, there is great uncertainty, and there is also disagreement about whether certain

negative consequences will actually materialise in the future, especially with regard to fully autonomous systems. However, the development of specific AI has many known consequences, and in many cases, the risks have been sufficiently substantiated to actualise the precautionary principle. The precautionary principle entails that AI researchers must describe and communicate the risk associated with the development and use of AI in their field of research. However, exactly which "ethically unacceptable consequences for health, society, or the environment" are emphasised, will vary according to the applied ethical perspective, values and interests. Pressure from various stakeholders can affect the algorithms, without this being the subject of an academic or political assessment. For instance, surveillance systems based on AI can be perceived both as a possible benefit and as a risk. From a defence and security perspective, surveillance could be considered a benefit that prevents crime and warns society of possible dangers, but from a point of view that emphasises privacy, such surveillance could also be considered a threat to the integrity of individuals.

Those who are most affected by the decisions that are made must also be guaranteed a voice in decision-making processes. Authorities and research institutions should therefore make it possible for citizens to be broadly involved in a debate about what the purpose of research should be, the structure of research initiatives and the application of research.

C) Big data

Together with computing power and algorithms, the development of specific AI has largely been driven by big data, which can also include personal data. The third set of challenges to which NENT believes it is important to draw attention is related to big data within AI research. Big data gives rise to new challenges related to data protection and protection of individuals in connection with research. Big data also raises other issues related to research ethics, which we will address below, including issues associated with biases in the data material, data quality and ownership of and access to data.

7. Ensure data protection and consideration of individuals

Protecting data that contains personal data can present its own challenges in the development and use of AI. Even if anonymised data is used in analyses,

it will still be possible to make comparisons with other data to reveal sensitive information or identify individuals; this thus also constitutes personal data. Collecting and using data that includes personal data may challenge the requirement for informed consent. When collecting and compiling large amounts of data, there is a particular risk that personal data may be used in ways of which we are unaware (because the purpose is also unknown to the researcher at the time of collection) and which we may not want.

Consideration of individuals and groups, who in various ways are involved in or directly affected by the research, is regulated by data protection rules and legislation (i.e. EU regulations and the supplementary Norwegian Personal Data Act). The data protection rules and legislation provide important guidelines for researchers, but cannot by themselves provide answers to the many challenges that researchers will face in connection with handling personal data. In Norway, the *Guidelines for Research Ethics in the Social Sciences, Humanities, Law and Theology* (NESH 2016) comprise the central tool for elaborating on the ethical responsibility of researchers and others affected by the research.
The implementation of data protection rules in Norwegian legislation introduced a number of basic principles that must be observed to ensure the legality of the processing of personal data. One of these is the principle of *data minimisation*. From this principle, it follows that one must not use more personal data than is necessary to fulfil the purpose of the intended processing, and the data must also be adequate for and relevant to the processing. For those who work with AI, however, it can be challenging to limit the amount of information that is processed because the development and use of AI usually requires large amounts of data to train the systems. In order to assess what is necessary, adequate and relevant, the researcher must have a clear idea of the purpose of the intended processing.

Another basic principle of data protection is precisely that data processing must be *purpose limited*. The processing of personal data cannot take place without there being a legitimate, specific and explicitly stated purpose. That the purpose must be clear and specific entails that it must be specifically described. Because this purpose is instructive with regard to the fulfilment of a number of the other data protection principles, vague and all-encompassing

statements of purpose are not permitted. For example, personal data must be deleted when the purpose of the processing has been fulfilled. However, if the purpose is stated to be the development of AI, it will be difficult to meet this requirement. Nevertheless, the requirement for purpose limitation can probably be met by specifying what kind of AI is to be developed, and what tasks it is assumed that this system will be able to perform. In any case, it may be challenging to determine whether the requirements for data minimisation and purpose limitation are met in such contexts. In its report on AI and data protection, the Norwegian Data Protection Authority notes that, when developing AI, one should seek to limit the training data at start-up, and then expand the data set when one knows to a greater extent what one needs. The Article 29 Working Party, an advisory group within the EU, points out in its statement on automated decisions the importance of the data controller establishing procedures and systems which ensure that the personal data used is correct and up-to-date at all times.

The object of these principles behind the regulations is to ensure that the processing of personal data only takes place when necessary, in order to limit encroachment on the individual data subject's privacy. When using large amounts of data to develop AI, it is important to keep this in mind and to assess whether the amount of data can be limited without compromising the stated purpose. If the amount of data cannot be limited, this choice must be justified and explained in order to demonstrate compliance with the principle of data minimisation, in the same way as with the requirements for necessity, adequacy and relevance. If the data is to be reused, the developers must ensure that this use is in accordance with the original purpose.

The main rule in research ethics is that personal data must not be collected, processed or shared without informed consent. When data collected for other purposes is reused in new and unanticipated areas, consent should be updated where possible. Challenges may also arise when data that is initially anonymised is compiled in new ways. In their research ethical assessments of information and consent, researchers have a responsibility to assess the degree of public access to the information, the sensitivity of the information, the vulnerability of those affected and the research's interaction and consequences (NESH 2019).

8. Quality assurance
In connection with AI research, there may be particular reason to ask critical questions about the quality, truthfulness and relevance of the data, because we do not always know the sources of the data, and because metadata may be absent or uncertain. Skewness in the material, properties of the analysis tool and human interpretations all increase the chances of logical fallacies and discriminatory decisions. This provides a basis for uncertainty in relation to interpretations and decisions based on AI. In recent years, we have seen several examples of how data can give rise to unreasonable decisions. In 2018, when Amazon tried to establish an objective recruitment process using AI, this turned out to produce discriminatory decisions based on gender bias because the datasets favoured men.

> To train AI, one must employ real data. Techniques such as deep learning work best with a lot of data. Then we are at the mercy of data quality, which is not always good (Norwegian Computing Center).

NENT believes that, in order to ensure verifiability and quality, it is essential that researchers and research institutions ensure that data sources are open and publicly available. At the same time, the uncertainty factors and limitations of research should be recognised and communicated.

9. Fair access to data
The development of AI technology may give a few individuals, companies or research groups the opportunity to dominate this field.

> The biggest concern is that AI seems to be becoming dominated by a few individual players such as Facebook, Google, Amazon and some others. In order for AI to function really well, it is dependent on large amounts of data and computing power. Companies such as Facebook have massive amounts of data and computing power that other actors cannot possibly match (CAIR, UiA).

In NENT's view, there is a risk that large parts of the research related to AI will disregard the requirements for transparency that apply to research otherwise, as they are laid down in the FAIR principles, by citing a need for

secrecy to protect competitive advantages, for example.
From the point of view of research ethics, it is essential to ensure that research, including data and results, is generally made available to everyone. As formulated in the *Guidelines for Research Ethics in Science and Technology*, the requirement for transparency means that research results, methods and data should be shared and published, not only to facilitate quality assurance, but also to maintain confidence in the research and ensure that the results benefit society (See Guidelines 3, 4 and 17).
Lack of data sharing is problematic for several reasons. Firstly, if only researchers in a few privileged companies have the opportunity to analyse large data sets, it will be impossible for outsiders to reproduce and evaluate their results. Secondly, researchers who are closely associated with private companies may have motivations and interests that could influence both which research is prioritised and the results of the research. Companies that conduct evaluations or research do so with a commercial goal in mind. Researchers who have the appropriate expertise and access to the right data can help to produce a better knowledge base, which in turn can benefit society more broadly.

NENT believes that governments and research institutions should facilitate public access to data. They should ensure transparency about who will have ownership of technology, infrastructure and data, which research areas are being prioritised and why, and who can be expected to benefit from the research.

Conclusion

The nine points in this statement are intended to serve as a starting point for reflection, guidance and discussion in research environments. They are also intended for stakeholders who fund and facilitate AI research, or who use AI. Because the development of this field of research is characterised by a fast pace and high uncertainty, this statement should be reassessed and revised regularly. NENT would like to continue the dialogue with academic environments about the research ethical challenges presented by AI research, and thus facilitate ethically sound and responsible AI research in Norway.

The Norwegian National Research Ethics Committees
Kongens gate 14, NO-0153 Oslo, Norway
Tel.: +47 23 31 83 00
www.forskningsetikk.no
ISBN: 978-82-7682-100-0