ericsson.com/
IndustryLab

AI – ethics inside?

Challenges and opportunities for the future

October 2021

Contents

03 Key takeaways
04 Setting the direction for AI development
06 The hidden consequences of AI
09 Unbiased data is a pipe dream
10 The paradox of trust
11 AI ethics for all
12 AI guidelines are not enough
15 Looking to the future

Methodology and background

The information in this report is based on contemporary subject literature, consumer and white-collar employee insights from the Ericsson Consumer & IndustryLab analytical platform and information gathered through telepresence interviews with 10 artificial intelligence (AI) experts. These included representatives from industries, unions, governmental agencies and academia in the European Union (EU). The interviews were conducted between January and March 2021.

The first in a series of reports, this edition aims to introduce AI ethics and spark an interest in learning more about more about how AI needs to be human-centered and aligned with both societal values and ethical principles. Upcoming editions will dig deeper into AI challenges and opportunities across different industries.

All IndustryLab reports can be found at: www.ericsson.com/industrylab

About Consumer & IndustryLab

Ericsson Consumer & IndustryLab explores the future of technology for consumers, enterprises and a sustainable society. We deliver world-class market research, actionable insights, and design concepts to drive innovation and sustainable business development. We provide a scientific, fact-based analysis regarding environmental, social, and economic impacts and opportunities of ICT.

Our knowledge is gained from global consumer, enterprise and sustainability research programs, including collaborations with leading customers, industry partners, universities and research institutions. Our research programs cover in-depth studies and over 100,000 interviews with consumers, working people and decision-makers each year, in 30 countries – statistically representing the views of 1.1 billion people.

With thanks to the insights of (in alphabetic order):

Victor Bernhardtz
Ombudsman for Digital Labour Markets, Unionen

Jacob Dexe
Research Institutes of Sweden

Virginia Dignum
Professor of Ethical and Social Artificial Intelligence at Umeå University, member of EC HLEG on AI

Andreas Henningsson
Domain Architect – Cognitive Data Analysis & Innovation, Swedish Social Insurance Agency

Isaiah Hull
Senior Economist, Central Bank of Sweden

Magnus Kjellberg
Section leader Digital R&DI, Sahlgrenska University Hospital

Sofia Löfstrand
Senior Project Manager, Drive Sweden

Inese Podgaiska
Secretary General, Association of Nordic Engineers, ANE

Amritpal Singh
Founder and CEO, Viscando

Mikko Viitaila
Regional Technology Officer for Microsoft Western Europe

Key takeaways

01 Lack of focus on unintended consequences of AI usage
Much of the AI research being conducted today focuses on ways to mitigate intended misuse of the technology. However, the experts interviewed generally agreed that the unintended consequences of normal deployment could be equally damaging and therefore need more attention.

02 The impossible challenge with bias in AI
Left unchecked, biased data has the potential to further increase gender inequalities or fuel continued racial injustice. Indeed, several of the experts agree that striving for unbiased AI is virtually impossible, though they added that these effects could be minimized.

03 The paradox of trust
AI is often used to support human decision-making, so it is important to build up trust in these systems, but without fostering an overreliance on the technology.

04 AI ethics is important for everyone
AI ethics is not just important for data scientists – it's crucial for everyone. AI will soon permeate many facets of people's work and personal lives.

05 Today's AI guidelines are not enough
To better support AI development for business, industry and society, today's guidelines, rules and regulations need to be improved to cover more than consumer-focused variables.

06 "Ethics by design" is key for acceptance
While technology itself can never be ethical, it can be ethically aligned. An ethical dimension should run parallel to areas like security and privacy within the design process – addressing these problems early will save time and money in the long run.

Setting the direction for AI development

AI systems are quickly becoming a staple in our daily lives, but they also bring a series of overt and covert challenges.

From healthcare to entertainment, advanced data analytics can be found in most parts of society. If done correctly, analyzing large sets of data enables both current and future generations to benefit from wide-scale economic growth and positive social change. In fact, two-thirds of information and communication technology (ICT) decision-makers strongly agree that data-driven processes and cultures will be key to their business success by 2030.[1]

More advanced forms of data analytics allow machines to learn from data, or even adjust their own approach accordingly – this is AI. In terms of adoption and market penetration, AI technology is predicted to have a profound impact on people's lives. This report includes technologies such as machine learning, machine reasoning, natural language processing and neural networks in the context of AI.

AI as a tool

As a general-purpose technology, AI can be applied almost anywhere and in any context. It has also proven to be a powerful tool in addressing societal and sustainability challenges, such as:

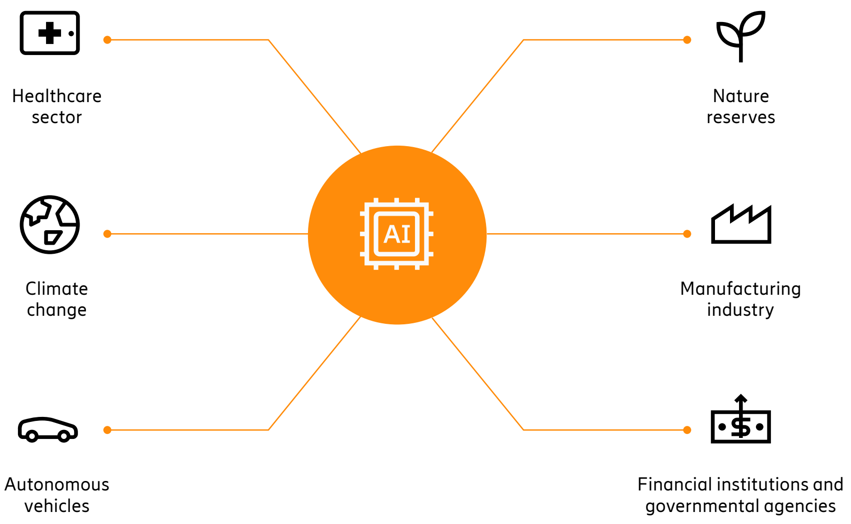
- diagnosis of diseases within the healthcare sector
- protection of threatened wildlife in remote nature reserves
- effective and energy-efficient processes within the manufacturing industry
- equal and fair treatment in contacts with financial institutions and governmental agencies
- avoidance of fatal traffic accidents by enabling autonomous vehicles and an intelligent transport infrastructure
- new, innovative solutions that address climate change

Exploring potential side effects

While AI can be a powerful tool supporting the development of a just, safe and sustainable society, it can also be applied for purposes that are destructive or harmful. Recent developments in AI have uncovered several negative side effects, many of which are accidental or unintended. For example, users' personal data has been exposed, unfair judgements have been carried out by opaque decision-making computer systems, and there is fierce debate about who is liable when an autonomous system malfunctions.[2]

As will be elaborated on later, these problems often stem from a human misunderstanding of how algorithms process different types of information, biases in the underlying data used for training models, or a general underestimation of the power of AI.

Figure 1: AI is a potentially powerful societal and sustainability tool



[1] www.ericsson.com/en/reports-and-papers/industrylab/reports/future-of-enterprises-dematerialization-path-to-profitability-and-sustainability
[2] Cathy O'Neil, 'Weapons of Math Destruction', Crown books, 2016

These challenges can cause harm to both individuals and organizations, resulting in an erosion of trust in AI systems and the underlying technology they are built upon.

Given the generalizability, built-in potency, and widespread adoption of AI, many of the experts interviewed agree that it will be crucial to set a clear direction for AI development and implementation. They advocate an agenda that enables governments, corporations, and societies to leverage the opportunities and positive effects of AI, while avoiding its misuse and potential unintended negative effects.[3]

Building a new kind of trust

More than other technologies, AI research has focused on building trust with qualities such as reliability, resilience, and privacy. Over time, researchers have learned how to address each of these topics by creating methodologies to mitigate potential negative effects and build trustworthiness. This can be done, for example, by implementing security by design[4] and privacy by design.[5]
Early research findings in the field of trustworthy AI have identified a versatile set of domains that need to be explored further.[6] Based on these, the work of Virginia Dignum, Professor of Ethical and Social Artificial Intelligence at Umeå University,[7] and the points outlined by António Guterres, UN Secretary-General, in his strategy on new technologies,[8] a first step would be to deepen the understanding in each of the following focus areas in relation to AI systems:

- control over technology
- understanding how (and why) algorithms reason
- privacy and integrity
- fair treatment
- physical safety
- unintended adverse impact
- intentional misuse
- personal freedom
- cognitive bias of developers and bias in learning algorithms
- management of data and consent
- accountability and liability of technology

In the following chapters, these areas will form the backdrop to the discussion on AI ethics, and its moral implications will be considered.

> "AI has the possibility to overcome some of the problems that might arise from discrimination that could happen in person but wouldn't happen from a properly calibrated machine learning model."
>
> **Isaiah Hull**, Senior Economist, Central Bank of Sweden

> "Trustworthy AI is a learning journey. It's not about what AI could do, it's about what it should do."
>
> **Mikko Viitaila**, Microsoft Regional Technology Officer for Microsoft Western Europe

[3] link.springer.com/article/10.1007/s11023-018-9482-5
[4] whatis.techtarget.com/definition/security-by-design
[5] https://ec.europa.eu/info/publications/white-paper-artificial-intelligence-european-approach-excellence-and-trust_en
[6] https://nordicengineers.org/wp-content/uploads/2021/01/addressing-ethical-dilemmas-in-ai-listening-to-the-engineers.pdf
[7] V. Dignum, Responsible AI, Springer (2019)
[8] https://www.un.org/en/newtechnologies/

The hidden consequences of AI

Even if the intentions with a new technology are morally sound, the results could be widespread and damaging.

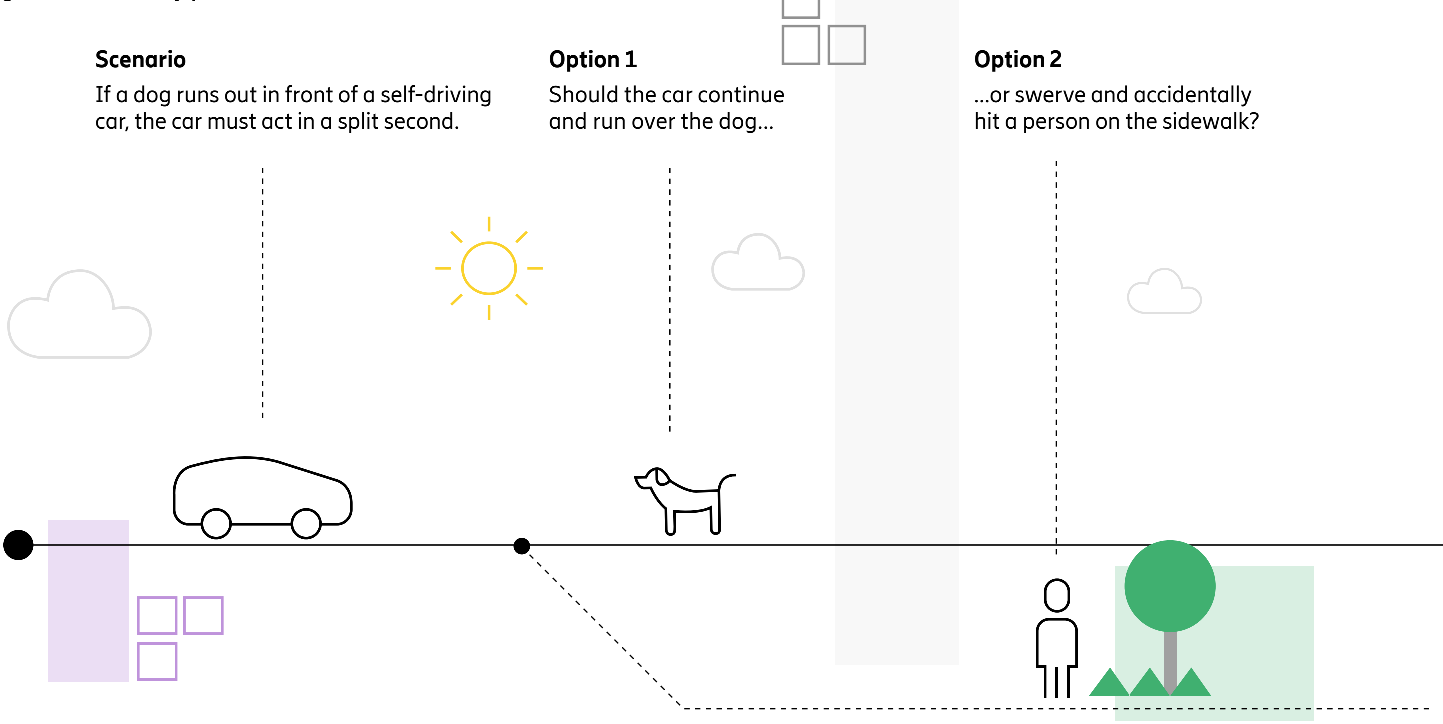
Purpose and usage
Several of the experts interviewed agree that it is important to consider the impact of AI, both from a purpose and usage perspective. The creator has a responsibility to explicitly design an AI for moral purposes, thus intent becomes pertinent. It is also necessary to consider the process by which its purpose is achieved. In terms of usage, it is important to understand how AI engages with its users and society, how it is used and its overall effect.

However, even when the purpose and use are benign, unintended effects can still occur if the AI is used in an unexpected way. These unintended consequences are, by their nature, difficult to foresee.

Unintended consequences
There is currently a distinct lack of focus on unintended consequences of AI usage, according to the experts interviewed. They generally agreed that, in recent times, a lot of AI research has focused on ways to mitigate intended misuse. But they also stressed that the unintended effects of normal deployment and usage could be equally damaging for the trustworthiness of AI-enabled systems, and therefore need more attention.

For example, the development of an AI application that tracks users and reminds them of when they need to exercise (as preventive health care) could also be used by an insurance company to identify high-risk customers (those that do not exercise). This unanticipated use of the collected personal data might not be obvious to the creator of the AI application, so the approach to AI development needs to include an understanding of technical, social and regulatory variables. More examples of the unexpected effects of AI will follow. These include the risk of unfair treatment, the threat to physical safety and the risks induced by overreliance on technology.

Figure 2: The trolley problem


Scenario
If a dog runs out in front of a self-driving car, the car must act in a split second.
Option 1
Should the car continue and run over the dog...
Option 2
...or swerve and accidentally hit a person on the sidewalk?

Negative effects do not always imply foul play

The use of AI is advancing and will enable highly sophisticated systems that would never have previously been possible.

In terms of AI ethics, autonomous vehicles exemplify one of the more widely discussed philosophical dilemmas – the trolley problem. Indeed, the team under Iyad Rahwan, formerly in Media Arts & Science, MIT, cited the trolley problem in the Moral Machine[9] experiment in 2018, as shown in Figure 2.

This example highlights the uncomfortable question of whether the AI in a self-driving car should be allowed to make these kinds of decisions at all?

An AI system does not appear out of thin air – a person has programmed it. Therefore, that person has directed the algorithm controlling the car's steering wheel on which option to select. This moral puzzle is difficult to solve because, depending on where in the world the event takes place, people may have a different opinion on the preferred outcome[10] – if there is one at all.

Real-world events can introduce complex AI opportunities

Another example that has been discussed in the media more recently concerns surveillance.[11] In the fight against the COVID-19 pandemic, governments could introduce smart heat-measuring cameras, facial recognition and aerial drones, for example, to spot individuals displaying symptoms and then contact anyone they've recently seen. An obvious, and likely unpopular downside of this policy is that all citizens would feel monitored, and their actions judged. It could be argued that this infringes upon one's integrity, privacy and basic human rights.[12] But it could equally be said that, for a greater good (such as tracing COVID-19 infections), individuals should sometimes accept an invasion of their privacy.[13]

Some of the experts pose that a pandemic could be seen as a just cause to survey people despite the risks of misuse, but under what other circumstances would this data collection be acceptable? A similar example concerns data-driven workplace analytics. With an increase in home working, there is a need to secure confidential data on personal computers. If an AI system scans a workplace for security threats, the same data could also be used to judge the efficiency of the employee.

If humans plan to implement AI in this way, it is vital that negative externalities are considered. It may indeed be unrealistic to expect that all possible outcomes will be predicted, so an additional layer of responsibility for AI developers is to appreciate that their system might be flawed even if its purpose is morally sound.

> "[We] assume that trust is built by [technology that] behaves coherently and is correct within defined limitations."
>
> **Amritpal Singh**,
> Founder and CEO, Viscando

> "The virus was a good eye-opener. It was also a good eye-opener in the perspective of how health issues can be misused to roll out different technologies."
>
> **Inese Podgaiska**,
> Secretary General, Association of Nordic Engineers, ANE

> "The single main challenge concerns are, I would say, the collection, management and regulation of, and possibly the dissemination of, data generated by employees as they work with tools that are now so plugged into the Internet of Things that they register data while the normal work continues."
>
> **Victor Bernhardtz**, Ombudsman for Digital Labour Markets, Unionen

[9] www.nature.com/articles/d41586-018-07135-0
[10] www.pnas.org/content/117/5/2332
[11] www.cnbc.com/2020/10/06/ecj-limits-government-spying-on-citizens-mobile-and-internet-data-.html
[12] www.un.org/en/about-us/universal-declaration-of-human-rights
[13] www.nature.com/articles/d41586-020-01578-0

Foreseeable unintended effects

The areas discussed on the previous page highlight some of the challenges that need to be addressed' in order to mitigate foreseeable, unintended effects of AI-enabled systems. Some of these unintended effects stem from training a system on only the most obvious situations and solutions, thereby omitting data on rarer use cases. Human reasoning includes the ability to handle new situations based on past experience. The tacit knowledge humans use in these situations is very hard to transfer to AI systems so, unless systems are trained in even the least likely outcomes, unexpected results will surface.

Other unintended effects might be triggered by the presence of new technology, similar to how social networks changed the way people communicate and stay in touch, or how the power of the internet changed the way products are bought and sold. At the time, these were all unforeseeable effects. While most of the changes mentioned have been positive, there are negative aspects to be aware of – ones that only surface once the system is up and running. This means that when trying to understand unknown or even unforeseeable effects of an AI system, it is also important to acknowledge that the original use case, when designing the system, might itself change once deployed.

> "We [Microsoft] have the notion of sensitive use cases – if AI is involved, is the use case something that might impose risks of harm to an individual, infringement on human rights or denial of consequential services?"
>
> **Mikko Viitaila**,
> Regional Technology Officer for Microsoft Western Europe

Risk vs. impact

As well as noting the risk of negative outcomes when using AI, there is also a need to consider the severity of the impact caused by a malfunction. A glitch in a recommendation service will likely have trivial consequences when compared to the failure of a system controlling an unmanned aerial vehicle operating over a densely populated city.

A recent whitepaper from the European Commission (EC) suggests that the area of operation for an AI system should be a factor in determining the minimum level of requirements for developing the system.[14] So-called "high-risk" areas refer to sectors like healthcare, transport and energy, where it is more likely that a malfunctioning system would lead to substantial damage, or even pose a risk to human life. Some technology components might also be defined as high risk regardless of the sector concerned. One such technology is biometric identification (such as face recognition), which has sparked fierce debate – particularly when deciding how it should be regulated and whether it should be restricted.[15,16]

Another example of this can be found in Microsoft's decision to consider sensitive use cases for AI systems within their implementation of responsible AI.[17]

The experts interviewed for this report broadly agreed that there is a need for continuous development of reliable, robust, accurate and secure systems, where relevant ethical challenges are considered throughout the design and development stages of any AI system.

While a previous Ericsson Consumer & IndustryLab study concluded that consumers expect that a range of organizations, products and services will be hacked or become infected in the near future,[18] AI systems can be designed to mitigate the threat of predictable intentional misuse and adversarial attacks.

This type of protection can be achieved by utilizing security and privacy by design throughout the development process.[19] By planning ahead and addressing any robustness issues early on, a level of resilience can be achieved that guards against both predicable intentional misuse and foreseeable changes in the operating environment of the deployed AI system, be it a natural disaster or sudden changes in resource consumption.

[14] ec.europa.eu/info/publications/white-paper-artificial-intelligence-european-approach-excellence-and-trust_en
[15] www.technologyreview.com/2020/06/26/1004500/a-new-us-bill-would-ban-the-police-use-of-facial-recognition
[16] www.ibm.com/blogs/policy/facial-recognition-sunset-racial-justice-reforms/?mhsrc=ibmsearch_a&mhq=facial%20recognition
[17] www.microsoft.com/en-us/ai/our-approach
[18] www.ericsson.com/en/reports-and-papers/consumerlab/reports/10-hot-consumer-trends-for-2016
[19] www.ericsson.com/en/reports-and-papers/white-papers/privacy-in-mobile-networks

Unbiased data is a pipe dream

AI systems are evolving rapidly but they still rely on human input and instruction, which inevitably introduces inherent biases.

AI systems feed on data and, as referenced by many of the experts interviewed, it is increasingly difficult to maintain control over the flow of data when it is used and/or generated by these same systems. Consequently, it is important to be aware of the possibility of skewed results from AI systems depending on the data that is used.

If the input to an algorithm is incorrect, the system might not completely fail, but the result could amplify any underlying bias or, even worse, endanger the physical security of humans or property. Bias in itself could lead to increased gender gaps or further fueling of racial injustice,[20] to name two possible examples. In fact, more than one-third of white-collar employees and managers think it is likely that AI processes will not treat them fairly by 2030.[21]

Perhaps even worse, several of the experts stated that striving for unbiased AI is virtually impossible, due to the inherent biases of humans as a species. Human decisions are often based on incomplete data, reinforced by previous experiences. In this process, data is unconsciously extrapolated, and knowledge is presumed, which risks introducing bias into the decision-making process. This is part of the human condition.[22]

Human biases must be minimized

Any AI system that uses real-world data, where human decision-making has been involved, is susceptible to adopting human biases. But, while it might not be possible to fully eliminate biased data, there are ways to minimize the effects. For instance, when the AI is at risk of inheriting human bias, expectations need to be managed and any faltering output mitigated by analyzing bias and through careful selection of relevant data sets, as well as ensuring diversity within involved development organizations.

"When we talk about AI [from the perspective of our agency], the driving force today is not about efficiency. It is about equal treatment for all… We know that we should treat all citizens equally."

Andreas Henningsson,
Domain Architect – Cognitive Data Analysis & Innovation,
Swedish Social Insurance Agency

"[Making sure the training data isn't biased] is one of the absolutely core questions… Many simply use the data that happens to be available, perhaps without knowing how and for what purpose the data originally was collected."

Amritpal Singh,
Founder and CEO, Viscando

[20] www.theguardian.com/technology/2016/sep/08/artificial-intelligence-beauty-contest-doesnt-like-black-people
[21] www.ericsson.com/en/reports-and-papers/industrylab/reports/the-dematerialized-office
[22] humanhow.com/list-of-cognitive-biases-with-examples/

The paradox of trust

A balance must be struck between building trust in AI technology while avoiding an overreliance.

According to Merriam-Webster's dictionary, trust is when people rely on someone or something to be truthful and/or accurate.[23] Something trustworthy is deserving of confidence. Technological advancements often start with a need to build trust amongst users – be it an industry, a citizen or a government agency. Experiments at Stanford University have confirmed that a user's opinion of how well technology performs is the key determinant of their trust in said technology.[24]

For AI, this also includes the system's ability to resist adversarial attacks and to uphold the integrity of data it handles. In essence, trust is about meeting the expectations that are set on the AI system – by all stakeholders.

Challenges for building trust in AI

An area often overlooked in the discussion of AI and trust is how AI differs from other technologies. The primary difference is that AI is regularly used to support human decision-making, such as providing predictions in diverse fields like economics, policy-making and healthcare.[25] However, people using the system might find it difficult to trust the results, particularly if they would have personally come to a contrasting conclusion without the help of the AI. The process is also complicated by the fact that AI systems sometimes provide no reasoning when coming to their conclusions – this is known as the AI black box problem.[26] Lacking a clear explanation of why and how an AI came to a certain conclusion is likely to negatively impact trust, especially when the AI delivers unexpected results.[27]

Another challenge is when there is a lack of evidence that an AI made a correct prediction or judgment. For example, in predictive policing,[28] the police force use AI to calculate where the next crime is likely to occur. When learning that there is an increased risk of a crime taking place in a certain area, the police can increase their presence in the area to discourage opportunists. However, if the police presence leads to the absence of a crime, there is also a lack of evidence that the AI's prediction helped to avoid crime.

Trust is key to adoption risk

Without trust, AI systems run the risk of being underused. This, in turn, could lead to significant loss of potential value creation in businesses and society as a whole. Equally, overuse and overreliance on AI could be just as damaging. As users become increasingly comfortable with AI technology, they run the risk of becoming dependent on the system. This reliance on intelligent machines for decision-making could create difficulties in questioning AI, even when its output is not understood. In fact, no less than one-third of consumers think AI-powered virtual assistants will lead people to forget how to make their own decisions, and just as many believe critical thinking will disappear due to the overuse of AI.[29]

The effects of underuse and overuse differs, but the consequences for us could be equally severe. This is the paradox of trust that needs to be addressed; to build trust in technology, avoid overreliance and ensure AI evolves as a usable tool for consumers, businesses, and society alike.

> "Trust calls for transparency, responsibility and reliability. It can be upheld by ensuring clear responsibilities for information as well as the understandability, information security and data protection of digital products and services throughout their life cycles."
>
> AI strategy of Finland [30]

> "How will we get doctors to use this [AI technology]? We know that it works well, we have the evidence, but it might not be enough to give them a black box that produces a result. Instead, it needs transparent and explainable AI."
>
> **Magnus Kjellberg**,
> Section leader Digital R&DI,
> Sahlgrenska University Hospital

[23] www.merriam-webster.com/dictionary/trust
[24] news.stanford.edu/2020/12/08/studying-trust-autonomous-products/
[25] Ajay Agrawal, Joshua Gans, and Avi Goldfarb (2018), Prediction Machines:The Simple Economics of Artificial Intelligence, Harvard Business Review Press
[26] www.nature.com/news/can-we-open-the-black-box-of-ai-1.20731
[27] www.wired.com/2016/03/googles-ai-viewed-move-no-human-understand/
[28] www.ojp.gov/pdffiles1/nij/230414.pdf
[29] www.ericsson.com/en/reports-and-papers/consumerlab/reports/10-hot-consumer-trends-2019
[30] vm.fi/documents/10623/7768305/VM_Tiepo_selonteko_070219_ENG_WEB.pdf

AI ethics for all

Philosophers have spent millennia pondering ethics, rarely agreeing on right and wrong. Yet now, humans must teach societal and ethical principles to machines.

Ethics is the study of what is morally right or wrong, and the answers are not always globally applicable. While in some cultures, youth and strength are most highly valued, wisdom and old age might be considered more valuable in others.

The field of ethics harks all the way back to Ancient Greece and Aristotle's teachings to strive for well-being and happiness. Aristotle is often compared to the more contemporary Immanuel Kant, who asked "what would happen if everyone did this", as well as the utilitarian approach which argued we should perform the action that generates the greatest good for the greatest number. These are but a few of the moral influences that guide and define our societal values today.

Before digital technology was embedded into society, discussions around the impact of technology solutions mostly focused on its purpose and intended use, ensuring it was robust, efficient and effective. An AI system today can make decisions based on the data it is provided with and the way it has been programmed. It even learns over time, sometimes creating new and unexpected results.

As AI stands to permeate both people's work and home lives, it is no wonder that several of the interviewed experts agree that AI ethics is not just important to data scientists – it's crucial for everyone. Indeed, almost half of white-collar employees and managers believe that AI-powered facial recognition will be used ubiquitously by 2030 and that the concept of privacy will therefore no longer exist for employees by 2030.[31]

Sharing the load

The responsibility for implementing ethical AI nearly always falls on the data scientist, which might initially make sense since they have knowledge about how the AI operates. However, as the interviewed experts agree, it is practically impossible for one individual to be aware of all the ethical implications of how the AI may be used. To map potential challenges and risks calls for a diverse team with varying perspectives. For instance, a person coming from an ethnic minority group might be more aware of the discriminatory boundaries the AI could accidentally cross.

There are many other roles that hold great responsibilities for ethical development and implementation. Take the example of an HR employee whose task is to participate in the procurement of a new AI system that should handle processes including employee data. The employee would need to be aware of the various ways the AI may risk overstepping privacy or integrity boundaries – at least well enough to know when to flag issues to the subcontractor. In short, a wide range of roles need to be involved in AI ethics in addition to the data scientists.

"Ethics is, to some extent, subjective. So, if Company A is going to have an AI that is ethically aligned, then that has to do with the prerequisites for Company A's product. It has nothing to do with what Company B is doing, their customer base, or their industry sector – it is all about Company A!"

Jacob Dexe,
Research Institutes of Sweden

"If you ask an AI expert to solve a problem, he or she does not usually have deep knowledge about the subject matter at hand. So, to make sure that the available data, analysis, and results are reliable, it is necessary to collaborate with the subject matter experts."

Sofia Löfstrand,
Senior Project Manager,
Drive Sweden

[31] www.ericsson.com/en/reports-and-papers/industrylab/reports/the-dematerialized-office

AI guidelines are not enough

It is not enough to draft a piece of one-size-fits-all legislation when dealing with AI – guidelines should be tailored to suit different industries.

Looking at AI systems, a set of instructions make up a tool created for a specific purpose. That AI system can, therefore, not be ethical by itself, or fully responsible for its actions. It is merely a man-made tool that, based on the data it is served, follows instructions.

The responsibility for this to work in a good, foreseeable way lies in the design, creation, training and deployment of the system. This is no trivial matter. In fact, 99 percent of decision-makers have experienced challenges while implementing AI.[32] If the mitigation of the foreseeable, unintended negative effects are not part of the design and development process, or if the data which it is served is not carefully chosen, it might well fail in its execution or behave unexpectedly.

Ethical frameworks for AI

Today, there are more than 160 AI ethics frameworks and normative guidelines published from different actors across the world,[33] most of which were published in the last 3–4 years. They come from international organizations, non-governmental organizations, professional associations, businesses, trade unions and representatives of civil societies, as well as various governmental and intergovernmental organizations (such as the UN and EU). When comparing these with the research into AI, most do include the four basic bioethics principles: do good, don't do harm, promote autonomy and be just and fair, with the addition of a fifth principle surrounding the ability to explain and to be accountable.[34] These principles also act as the backdrop to the work the EC carried out within their high-level expert group (HLEG) on AI. The EC published "Ethics Guidelines for Trustworthy Artificial Intelligence" in late 2019.[35] These guidelines outline seven challenge areas for AI development in their document: human agency and oversight, technical robustness and safety, privacy and data governance, transparency, diversity, non-discrimination and fairness, societal and environmental well-being and accountability.

Even though these guidelines are a good start for any industry to assess the ethical dimension in their development and deployment of AI-enabled products, they are not nearly good enough. Since the global product landscape is so vast, they will sometimes be misaligned with the target use and context for specific products. Representatives from the EC HLEG, speaking on AI, admit that there is a need for every industry to explore which guidelines are relevant and to determine ways to operationalize those in their respective design processes. To better support AI development for business, industry and society, today's guidelines also need to be improved to cover more than consumer-focused aspects, according to several of the interviewed experts.

Deployment and implementation

As AI is introduced and implemented throughout an organization, it becomes intertwined with many different processes and activities. It starts to involve itself in many different roles within the organization, from experts and decision-makers, to salespeople and operational staff.

As previously mentioned, the way an AI system is deployed is equally as important as development. Getting the right subject matter experts, as well as the intended target user group, involved at an early stage has proven to be an important success factor in the past.

> "[Ethical principles and frameworks] need to be used – they have to be legitimate, and you should feel that they can be translated to fit your work. To just have the principles means absolutely nothing unless they have been internalized."
>
> **Jacob Dexe**,
> Research Institutes of Sweden

> "[Today's EC HLEG AI guidelines] don't cover everything. So, for example, when we talk about AI systems that are embedded in telecom switches or used for determining network capacity, those are not the type of applications that the EC group were concerned about."
>
> **Virginia Dignum**,
> Professor of Ethical and Social Artificial Intelligence at Umeå University, member of EC HLEG on AI

In Sweden for example, the city of Trelleborg implemented AI to handle social security issues by ensuring that AI researchers, as well as users, were involved during the implementation.

The handlers were very pleased with the AI system as it enabled them to focus more of their attention on other activities.

The city of Kungsbacka used the same system but without involving researchers or users, which caused social security handlers to object and resign.[36] Figure 3 shows seven challenge areas within the EC's ethical AI guidelines in detail.

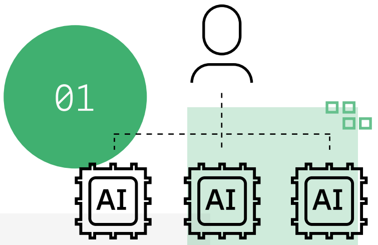
[32] www.ericsson.com/en/reports-and-papers/industrylab/reports/adopting-ai-in-organizations

[33] inventory.algorithmwatch.org

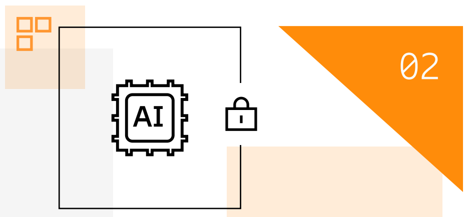
[34] An Ethical Framework for a Good AI Society: Opportunities, Risks, Principles, and Recommendations, L. Floridi et al, Minds and Machines (December 2018)

[35] https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai

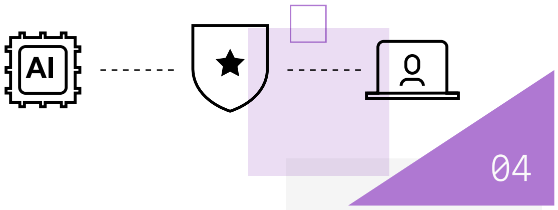
[36] www.arbetsvarlden.se/succemodellen-i-trelleborg-moter-motstand-man-kopierar/(Swedish language)



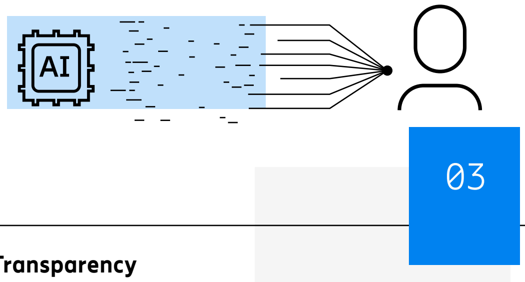
Human agency and oversight
AI systems should empower human beings, allowing them to make informed decisions and fostering their fundamental rights. At the same time, proper oversight mechanisms need to be ensured, which can be achieved through human-in-the-loop, human-on-the-loop and human-in-command approaches.



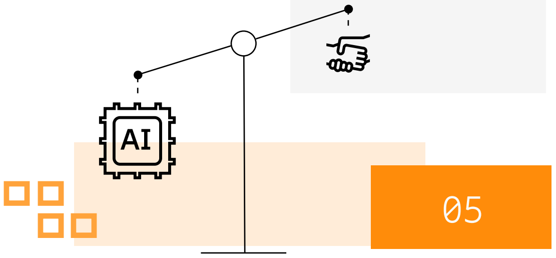
Technical robustness and safety
AI systems need to be resilient and secure. They need to be safe, ensuring a fallback plan in case something goes wrong, as well as being accurate, reliable and reproducible. That is the only way to ensure unintended harm can be minimized and prevented.



Privacy and data governance
Besides ensuring full respect for privacy and data protection, adequate data governance mechanisms must also be ensured, taking into account the quality and integrity of the data, and ensuring legitimized access to data.

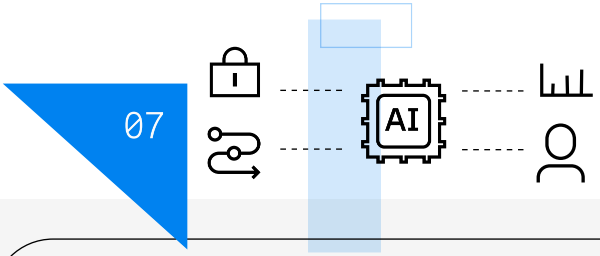


Transparency
The data, system and AI business models should be transparent. Traceability mechanisms can help achieve this. Moreover, AI systems and their decisions should be explained in a manner adapted to the stakeholder concerned. Humans need to be aware that they are interacting with an AI system, and must be informed of the system's capabilities and limitations.



Diversity, non-discrimination and fairness
Unfair bias must be avoided, as it could have multiple negative implications, from the marginalization of vulnerable groups, to the exacerbation of prejudice and discrimination. Fostering diversity, AI systems should be accessible to all, regardless of any disability, and involve relevant stakeholders throughout their entire life cycle.

Societal and environmental well-being
AI systems should benefit all human beings, including future generations. Therefore it must be ensured that they are sustainable. Moreover, they should take into account the environment, including other living beings, and their social and societal impact should be carefully considered.



Accountability
Mechanisms should be put in place to ensure responsibility and accountability for AI systems and their outcomes. Auditability – which enables the assessment of algorithms, data and design processes – plays a key role therein, especially in critical applications. Moreover, adequate and accessible redress should be ensured.

Figure 3: Creating trustworthy AI as laid out in the EC Ethical Guidelines

To conclude, AI can never be fully ethical – but it can be ethically aligned. The AI experts interviewed agree that an ethical dimension needs to be incorporated to run in parallel with areas like security and privacy, throughout the entire design process. Addressing potential issues at the start will save both cost and effort in the long run. This means that "ethics by design" will be key for acceptance.

The EC has addressed the notion of trustworthy AI development in a number of white papers and reports, notably defining three characteristics of trustworthy AI in its ethical guidelines. According to the EC, systems that use AI should be:

- lawful – respecting all applicable laws and regulations
- ethical – respecting ethical principles and values
- robust – both from a technical perspective and considering its social environment

Looking forward to future regulatory frameworks

In April 2021, the EC published a proposal for a regulation laying down harmonized rules on artificial intelligence (Artificial Intelligence Act).[37] This proposal is a legal framework that will address several of the challenges that are discussed in this report, to promote AI development. It focuses on high-risk AI systems and certification, and will apply to all providers and users of AI systems in the EU. It is anticipated that the new legal framework will be refined during discussions between European stakeholders until at least 2023.

[37] https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=CELEX:52021PC0206&from=EN

Looking to the future

In the future, AI could bring countless benefits to global society and the economy, helping to create a more inclusive, sustainable world.

The experts surveyed agree that developers, decision-makers and policy-makers need to ensure that AI is worthy of the trust society will place on it. Furthermore, it is imperative that the work to identify mitigation strategies for potential negative effects is intensified.

Developing AI based on "ethics by design" is about adhering to a consistent, responsible decision-making framework that keeps in mind considerations about the people who are impacted.

As mentioned earlier in this report, there is a set of principles to refer to when considering the ethics of AI. While these principles can be translated into requirements to be used in AI development, further refinement will be needed to better fit different industries, products and purposes.

Confidence in AI will unlock trust

The broad consensus among the interviewed experts was that certification will have an important role going forward. This is also part of the new legal framework proposed by the EC. To be able to state with confidence that a certain AI system is trustworthy by design, a further development of tools will be needed. Examples of such tools could be a combination of methodology, questionnaires, and software that are used throughout the development cycle to assess, for instance, the explicability of an algorithm or the unbiases of training data sets. They could also be used during the deployment of an AI implementation to evaluate if the appropriate considerations of its impact have been taken. In this way, the principles at hand can be operationalized – like the way areas such as security and privacy have already been addressed today – by design.

AI to empower the telecoms industry

The cited progressive steps can naturally also be applied to the telecoms industry. Its ambitions with future AI-enabled ICT systems can only be successfully realized if mobile networks, cloud solutions and other ICT infrastructure are trustworthy and performing according to expectations – even in the face of attacks, faults or other disturbances.[38]

The telecoms industry must also show evidence of transparency to meet these expectations. In addition, the importance of conforming to well-understood principles that define ethical behavior and that can be easily explained must also be recognized.[39]

Trust is at the core of solving this, and it spans all parts of business, not just the technology. So, everyone – not only data scientists – needs to care and be aware of the inherent risks. The future is AI-powered, so being sure that it works the way it is intended is, and will continue to be, vital.

> "If we summarize most of the mistakes and insights [in the development of AI systems], the most important lesson is to remember that AI is just a tool to reach something, it is not a goal in itself."
>
> **Andreas Henningsson**, Domain Architect – Cognitive Data Analysis & Innovation, Swedish Social Insurance Agency

> "I don't think that the certification of companies would help us, but certification of products and services they produce, yes. Like "ethically aligned" products, would be a good starting point and it would also be a competitive advantage for companies, as it will increase trust in them."
>
> **Inese Podgaiska**, Secretary General, Association of Nordic Engineers, ANE

> "One [area of importance] is the issue of transparency, and I don't mean that its transparency in terms of making transparent algorithms or non-black box algorithms. But much more transparency of the auditability possibilities – to ensure that your processes are auditable."
>
> **Virginia Dignum**, Professor of Ethical and Social Artificial Intelligence at Umeå University, member of EC HLEG on AI

[38] www.ericsson.com/en/blog/2021/5/cognitive-networks
[39] www.ericsson.com/en/reports-and-papers/white-papers/building-trustworthiness-into-future-mobile-networks

About Ericsson

Ericsson enables communications service providers to capture the full value of connectivity. The company's portfolio spans Networks, Digital Services, Managed Services, and Emerging Business and is designed to help our customers go digital, increase efficiency and find new revenue streams. Ericsson's investments in innovation have delivered the benefits of telephony and mobile broadband to billions of people around the world. The Ericsson stock is listed on Nasdaq Stockholm and on Nasdaq New York.

www.ericsson.com

Ericsson
SE-164 80 Stockholm, Sweden
Telephone +46 10 719 0000
www.ericsson.com

The content of this document is subject to revision without notice due to continued progress in methodology, design and manufacturing. Ericsson shall have no liability for any error or damage of any kind resulting from the use of this document

GFTL-21:000797 Uen