# AI no longer has a plug

Rudy van Belkom

The Netherlands Study Centre
for Technology Trends (STT)

*About ethics in the design process*

# AI NO LONGER HAS A PLUG

*About ethics in the design process*

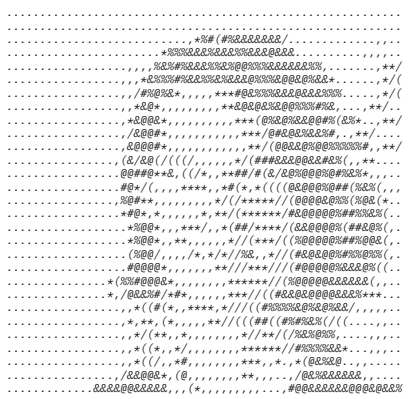
---

Part III in the series 'The future of artificial intelligence (AI)'
*Making choices in and for the future*

---

Rudy van Belkom
The Netherlands Study Centre for Technology Trends (STT)

STT Netherlands Study Centre for Technology Trends



# A handle for the future of AI

*By Maria de Kleijn-Lloyd, Senior Principal, Kearney, Chairperson think tank STT futures exploration AI*

What is the future of AI in the Netherlands? That is a question that is almost impossible to answer. Because: what is AI exactly, which future scenarios are there and who determines ‘what we want’, and on what basis? The third part of the STT trilogy ‘the future of AI’ focuses on that third, normative question. The aim is to generate a broad social discussion about this issue, because AI will touch us all in one way or another: directly, as users of apps, but also indirectly, when other people and organizations use AI, for instance doctors who let a scan get analysed algorithmically to be able to make a diagnosis. This is not science fiction; a lot of it is already possible. Even today, the impact of AI is significant and it is expected that the impact will only grow. That’s why it is good to focus explicitly on the associated ethical and social choices.

> A lot of work is already being done. A high level expert group of the EU, for instance, has described the main ethical principles of AI, like *explainability* and *fairness* in great detail. But that is not enough, because, it is relatively easy to agree when it comes to general principles: of course we want privacy, of course we want fair results. In discussions about a vital infrastructure, these are also known as *feel good principles*. Of course we are in agreement.

Things tend to become more complicated when we try to translate the principles into practical applications, when we are faced with two challenges. Firstly, we need to find a way to apply the principle in practice. For example, what is a transparent algorithm? One of which the entire code – sometimes multiple terabytes – is published, but which can only be understood by a select group of experts? Or one that comes with information written in laymen’s terms regarding the main design choices, source data, operation and side-effects? Secondly, some principles can conflict when put into practice, for instance transparency and privacy. Again, the context is important: medical information is different from Netflix preferences. Who decides which principle takes precedence? We need to take that complicated next step together, because these are choices that the designers of algorithms are already making every day.

> This foreword was written during the *intelligent lockdown* of the corona-crisis, in April of 2020. For those who were worried that the Netherlands would miss out when it comes to digitization and AI: our physical infrastructure (from cables to cloud) turns out to be robust enough and most consumers and businesses were also able to switch to working from home with relative ease. We all develop new digital skills with remarkable speed. But what is perhaps even more interesting is that we also use the media on a massive scale to take part in the dialogue about algorithms and apps. For example the ‘appathon’ that the Ministry of Health organized surrounding the corona app. How do you create that app in such a way that it safeguards the privacy of citizens, cannot be misused and is accurate at the same time? And when we say accurate, does that mean ‘not to miss any corona cases’ (no false negatives) or ‘nobody being quarantined needlessly’ (no false positives)? As such, the current situation, no matter how sad, helps us create clarity regarding a number of ethical choices in AI. With 17 million participants nationwide.

I hope that, with the help of this study and in particular via the interactive online components, we will be able to continue a focused and broad dialogue and translate it into practical handles that will lead to an AI future in the Netherlands that we not only accept but can really embrace as well.

## Research approach

Artificial intelligence (AI) is high on the agenda of politicians, directors, managers and policy-makers. Despite the fact that a lot has been written about AI, there is a lot that we do not know yet about the way the technologies will develop in the future and what impact they may have on society. That is why it is time for a futures research on AI. A major difference with other technologies is that AI is becoming increasingly autonomous, leaving fewer and fewer decisions up to us, which has led to the following central question:

*What is the impact of AI on decision-making in the future?*

Human decision-making is the outcome of various components. In addition to factual knowledge, the perception and ambitions of the people involved in the decision-making process also play a role. The way these components are organized is subject to constant change. In addition, the people taking part in decision-making processes often have diverging ideas about reality. What one person considers to be an irrefutable fact, is questioned by others. In that sense, you could argue that automated decision-making could intrude the necessary objectivity. However, the question remains to what extent AI can take over human decision-making processes. And to what extent we can and are willing to allow that to happen.

> The future is often seen as something that happens to us, when it is in fact something that we, as humans, have the power to influence. The ambition of this futures research is to try and formulate the desired future of AI with a multidisciplinary group of experts, and to examine what we need to realize that desired future. That has led to the following trilogy:





### Part 1: Predicting

When we talk about the future of AI, there appear to be only two flavours: a utopian one and a dystopian one. Often, these discussions begin with ethical questions, while skipping the question whether the technologies can actually produce these scenarios in the future, which is why the focus in part 1 is on the technologies:

> *How does AI relate to human decision-making and how will AI develop in the future?*

To examine that question, we used *technology forecasting*. Based on literature studies and expert interviews, we mapped the most realistic development trajectory of the technologies (scope 0-10 years). For this part, we consulted 40 experts, from AI experts and neuroscientists to psychologists and management experts.

### Part 2: Exploring

The forms in which AI will be deployed in the future depend on the social context. In addition to the utopian and dystopian visions, there are other future visions as well. In part 2, the focus is therefore on the implementation of possible future scenarios:

> *What are the implications of the way AI develops on decision-making in the future and what are the potential future scenarios that can be deployed accordingly?*

To examine that question, we used *scenario planning*. Via creative sessions, several future scenarios were mapped out (scope 10-20 years). For this part, we organized four scenario workshops with 30 experts from different areas.

Part 3: Normative

AI can become the first technology that will determine its own future, in which case it would be more important than ever before to determine the desired conditions for its development. What kind of future do we want? That is why the focus in part 3 is on ethical issues:

> *Which ethical questions play a role in the impact of AI on decision-making in the future and how can we develop ethically responsible AI?*

To answer that question, we used *backcasting*. Based on an *online questionnaire*, the desired future of AI was mapped (scope 20-30 years). We then examined which elements are needed to realize that future. For this part, a questionnaire was distributed among three groups, namely experts, administrators and students, more than 100 of whom filled in the questionnaire.

*Scope*

Usually, the futures research of the Netherlands Study Centre for Technology Trends (STT) have a scope of about 30 years, the aim of which is not to produce concrete statements about a specific year, but to indicate that the explorations look beyond short-term developments. The goal is to overcome the limitations of the current *zeitgeist*. For that, we need to look beyond boundaries and broaden our horizon in such a way that we can adopt a more future-oriented approach.

This futures research builds on the insights from earlier STT studies on data, namely *Dealing with the data flood* (2002) and *Data is power* (2017). Data provides the building blocks for today's AI technologies. The question is whether that continues to be the case in the future or whether the technologies will develop in a different direction.



*» Technology often no longer has a plug, which means you cannot simply unplug it. «*

*Think tank*

During the exploration, we made extensive use of the expertise and experience of administrators and relevant experts, compiling the following multidisciplinary think tank:

| | | |
|---|---|---|
| Marc Burger | Capgemini | CEO |
| Patrick van der Duin | STT | Director |
| Bernard ter Haar | Ministry of the Interior and Kingdom Relations | Special advisor |
| Frank van Harmelen | VU University | Professor Knowledge Representation & Reasoning |
| Fred Herrebout | T-Mobile | Senior Strategy Manager |
| Marijn Janssen | Delft University of Technology | Professor ICT & Governance |
| Maria de Kleijn-Lloyd* | Kearney | Senior Principal |
| Leendert van Maanen | Utrecht University | Assistant professor in Human-centred AI |
| Marieke van Putten | Ministry of the Interior and Kingdom Relations | Senior Innovation Manager |
| Jelmer de Ronde | SURF | Project manager SURFnet |
| Klamer Schutte | TNO | Lead Scientist Intelligent Imaging |
| Maarten Stol | BrainCreators | Principal Scientific Advisor |

* chairperson think tank STT futures research AI

## Reading guide part III

Artificial intelligence (AI) appears to be one of the most frequently discussed technologies at the moment, as well as being one of the least understood technologies. In some areas, it is ‘dumber’ than people think, but in other areas, it is actually ‘smarter’. And although having a computer for a Prime Minister still seems a little far-fetched, AI certainly has an impact on our labour market. *‘AI is here to stay’*. However, these intelligent systems are still often seen as a goal in themselves, without wondering whether AI is the best solution to a given problem. Many people appear to assume that AI is an unstoppable force of nature that we have to put to use somehow. No matter what. We need to realize, however, that AI in itself is neither good nor bad, the question is how it is used by people. So the question is what kind of society we want to be, given all the technological developments. Society will change fundamentally no matter what and AI can help us find the right path. But we do need to find out where it is exactly that we want to go.

### Retrospect part I

The overarching nature of AI makes it a concept that is hard to define and there is no unequivocal and internationally accepted definition. In 2019, the Dutch government launched the Strategic Action Plan for Artificial Intelligence (SAPAI), which describes the intention of the government to speed up the development of AI in the Netherlands and profile it internationally. The document uses the definition of the European Commission: ‘AI refers to systems that display intelligent behaviour by analysing their environment and – with a certain degree of independence – take action to achieve specific goals’. This sentence is filled with broad terms that can be interpreted in different ways: Systems? Intelligence? A certain degree of independence? Specific goals? And yet, based on this holistic definition, a complete action plan is developed.



To get a grip on the development of AI, it is important to get a grip on the operation and application of AI.

That is why in part 1 of this futures research, ‘Submarines don’t swim’, we took a close look at what AI is, how it works, how it relates to human intelligence, to what extent human decision-making can be automated, what the most dominant expert opinions are with regard to the development of AI and which economic and political factors affect the direction in which AI is developing.

### Retrospect part II

When people talk about the future of AI, they often think in extremes: will it be utopia or dystopia? In addition to the fact that such extremes often work better in movies and newspaper headlines, this dichotomy also has to do with the idea that, although it is very unlikely that either scenario will ever occur, the potential impact can be so great that it deserves a certain measure of reflection. That applies both to the utopian vision (we never have to work again) and the dystopian vision (we will become slaves to technology). However, there are multiple flavours.

That is why, in part 2 of this futures research, ‘Computer says no’, we examined several alternative realities and translated them into five future scenarios, each with an increasing level of intensity. In the scenario of *Game* Over, the limited availability of resources means that the promise that AI once was never materialized. In the scenario of *The winner takes all*, the need for control and regulation means that AI is mainly used as a tool to increase human intelligence. In the scenario of *Privacy for sale*, the quest for automation means that people are replaced by AI in different areas. In the scenario of *Robot Rights*, man and machine work and live together as equals, while AI transcends human intelligence in all domains in the scenario of *The Singularity is here*.

These future scenarios help us to come to grips with the changing relationship between people and technology, in addition to allowing us to identify the desirable elements in the future, giving us something to work towards.

*Why this publication?*

The way in which AI will be deployed in the future depends to a large extent on the social context. So it is not just about the performance of the technologies and the availability of the resources, but also about strategic interests and social acceptance. What is often overlooked is that we create that context ourselves. The choices we make today will have a major impact on the possible futures, which is why it is important to examine ethical questions and look for answers. Who is responsible when an AI application messes up? Can we grant rights to technologies? And how do we make sure that AI applications are free of prejudices? Fortunately, there are more and more ethical guidelines that have to guarantee the development of reliable AI systems. However, the question is how you can translate those abstract values into concrete practical applications. At the moment, people mostly talk about ethics, but as yet, there are no practical tools for integrating ethics into the development process. If we want to use ethically responsible applications in the future, now is the time to put those ethics into practice.





*» Many researchers will tell you that the heaven-or-hell scenarios are extremely unlikely. We're not going to get the AI we dream of or the one that we fear, but the one we plan for. Design will matter. «*

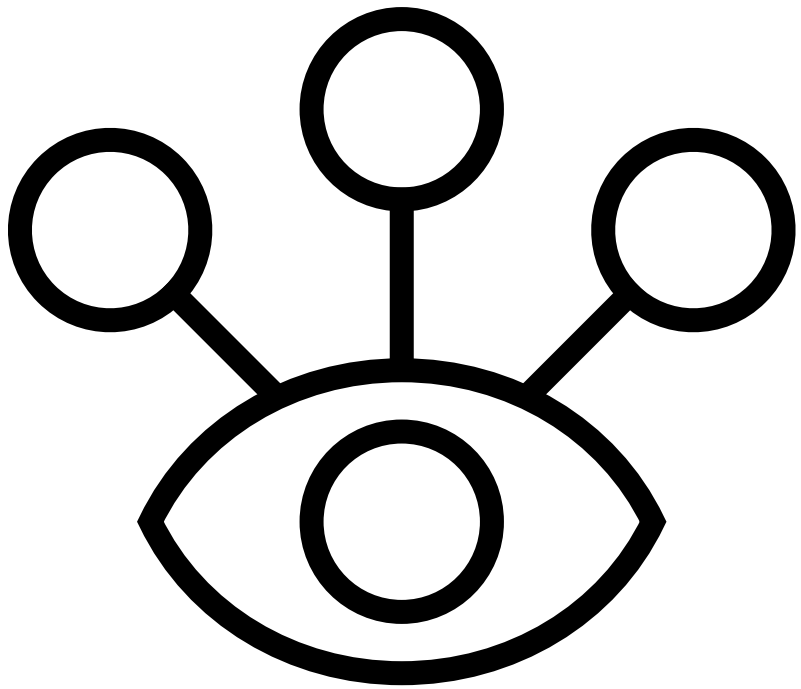
-- Stephan Talty

*Who is this publication for?*

This publication is meant for everyone who is interested in the role of ethics in the development process of AI. The aim is not just to map ethical issues, but also to examine how we develop ethically responsible AI applications in the future. Do you want to know how we can put ethics into practice? And are you open to a new perspective and approach to ethics? Then this publication is for you.

*How is this publication organized?*

In the first chapter, we zoom in on AI & Ethics, trying to explain the emergence of ethics in AI and looking at the ethical issues that play a role in today's discussion. Next, in chapter 2, we look at the ethical guidelines that have already been developed and what their limitations are. In the third and final chapter, we look at a new vision and approach to the integration of ethics in the development of AI, focusing on the design as well as the process.



# 1 ----------
# AI & Ethics

---



---

Pages: 44
Words: 10.130
Reading time: approx. 1,5 hour

## 1. AI & Ethics

When you use the term artificial intelligence ↴ (AI), that will make people pay attention in most cases. We all want ‘something’ with AI. But there is no consensus about the question as to what exactly AI is and what we can do with it. People often speak of ‘technology’ in generic terms, rather than about specific technologies. It is the result of a series of different technologies that together produce a form of intelligence. This artificial form of intelligence ↴ can be realised in different ways. Think, for instance, of whole brain emulation (WBE), which refers to the attempt to transfer a complete brain into a computer. However, when we talk about AI, in most cases, we refer to applications in the area of machine learning, a revolution in which people no longer do the programming (if this, then that), but in which the machines themselves deduce rules from data. With a large quantity of data ↴, computing power and algorithms ↴, there is no AI.

A similar principle appears to apply to AI; it is not easy to achieve an unambiguous view. Ethics is an area of philosophy that concerns itself with the systematic reflection of which actions can be called good or right. With regard to ethical issues, we all have ‘some’ opinions about it, often without being able to tell from which perspective we are reasoning. There is no such thing as ‘ethics’. There are various ethical movements, for instance consequentialism ↴ (which focuses on the results of our actions) and deontology ↴ (which focuses on the starting point, despite the consequences of our actions). Think, for instance, of Robin Hood: he steals from the rich to give to the poor. The question whether his actions are ethically responsible depends on the perspective. In the case of consequentialism, the actions of mister Hood are defensible, because they promote equality. But not in the case of deontology: stealing is wrong, even if it is for a good cause.



The fact that both AI and ethics are overarching concepts that include various movements and approaches makes the discussion that much more complicated. As a consequence, various AI applications and ethnical movements tend to intertwine, which is why it is important to take a closer look at the emergence of ethics in AI and to distinguish different ethical issues and movements.

## Ethics in the spotlight

Ethics ↴ are a hot topic right now, in particular with regard to the development of AI ↴. Sometimes it would appear as though the subject is reserved exclusively for AI. But of course that is not the case. Various ethical issues play an important role in different domains. Think, for instance, about the discussion surrounding cloning in medical biology or the behaviour of bankers in the financial sector. And yet, we seem to be making little headway when it comes to 'the right actions' in relation to AI, which, when you think about it, is actually not that strange.

> As AI is able to operate more and more independently, we will have to relinquish more and more control. For humans, this is new territory and it is bound to create feelings of anxiety, which Hollywood does not hesitate to feed. For almost a century, scenarios about robots rising up against people have been a popular storyline (assuming that Fritz Lang's *Metropolis* is the first real science fiction movie in which robots have bad intensions). In addition, fiction is often overtaken by reality. In 2016, for example, a Tesla running on autopilot crashed, killing the driver. Physical injury as a result of the use of technology is not new, but physical injury without a *human in the loop (HITL)* very much is. We can no longer afford to philosophise about what is right and wrong from a distance. More than ever, ethics is a matter of practical philosophy.

### Ethics on the rise

Who is responsible for an accident with a self-driving car? How do we protect our privacy with data-hungry applications? And how can we prevent unjust actions by AI systems? Nowadays, it is hard to attend an AI-related conference without part of the programme being dedicated to ethics. And while, until a few years ago, ethical specialists were the support act, nowadays they increasingly make up the main event. That shift is not only visible in the programming of the conferences, but also in their organization, for instance that of the largest AI conference in the world, the *Neural Information Processing Systems conference*. A large-scale questionnaire in 2017 showed that the conference did not provide a hospitable environment to female participants. Respondents reported sexual intimidation and sexist or sexually offensive remarks and jokes. The organization therefore decided to introduce a new code of conduct in 2018, to avoid discrimination. In addition, they tried to make the event more inclusive by supporting childcare, among other things, and they also changed the acronym NIPS to NeurIPS, to avoid the association with nipples. A seemingly minor, yet important change. In a previous edition, some male visitors of a workshop about women in machine learning wore a t-shirt with a 'joke' about nipples.

> And it doesn't stop at conferences. Dozens of organizations, from businesses and scientists to governments, have set up ethical guidelines ↴ to be able to guarantee the development of reliable AI applications. Think, for example, of the of the 'Perspectives on Issues in AI Governance' of Google, the 'Asilomar AI Principles' of the Future of Life Institute and the Ethics guidelines for trustworthy AI' of the High-Level Expert Group on Artificial Intelligence (AI HLEG) of the European Commission. The American Ministry of Defence has its own 'AI Principles', with recommendations to safeguard the ethical use of AI within the department. Even the Vatican published guidelines in 2020 for the development and use of AI: 'Rome Call for AI Ethics'. Tech giants like IBM and Microsoft were among the first signatories. Values like *privacy*, *transparency* ↴ and *fairness* ↴ are given ample attention in different guidelines. In addition to different ethical principles, more and more '*ethical boards*' are created to supervise the ethical actions of tech companies. According to Gartner, '*Digital Ethics & Privacy*' was one of the 10 '*Strategic Technology Trends*' of 2019.

## Unethical ethics

However, it would appear that this 'ethical hype' is limited to big tech organizations. Despite the fact that companies around the world expect they will apply AI within their organization, they lag behind in the discussions surrounding ethics according to research by Genesys (2019). More than half of the interviewed employers state that their business does not have a written policy regarding the ethical use of AI at the moment. An interesting fact in this regard is that only 1 in 5 of the respondents express any concern about the possibility of AI being used in an unethical way in their organization. That percentage is even lower for older respondents. While 21% of the millennials are worried about such unethical use in their organization, a mere 12% of generation X and 6% of the baby-boomers share those concerns. The researchers wonder whether that is really the right attitude.

In 2018, the *AI Now Institute* produced an overview of important news about the social implications of AI and the tech industry. The overview confirmed what many people suspected: it was a turbulent year. Just some facts from 2018: an increase in the abuse of data (culminating in the revelation of the Cambridge Analytica scandal, in which millions of personal data were used to influence elections), an acceleration of facial recognition software (like the collaboration of IBM and the New York Police Department, making it possible to search faces bases on race, using camera images of thousands of police officers who unwittingly participated) and an increase in the detrimental consequences by testing AI systems on live populations in high-risk domains (like the deadly collision of a pedestrian by a self-driving Uber). And that is just a small selection from a much larger database.







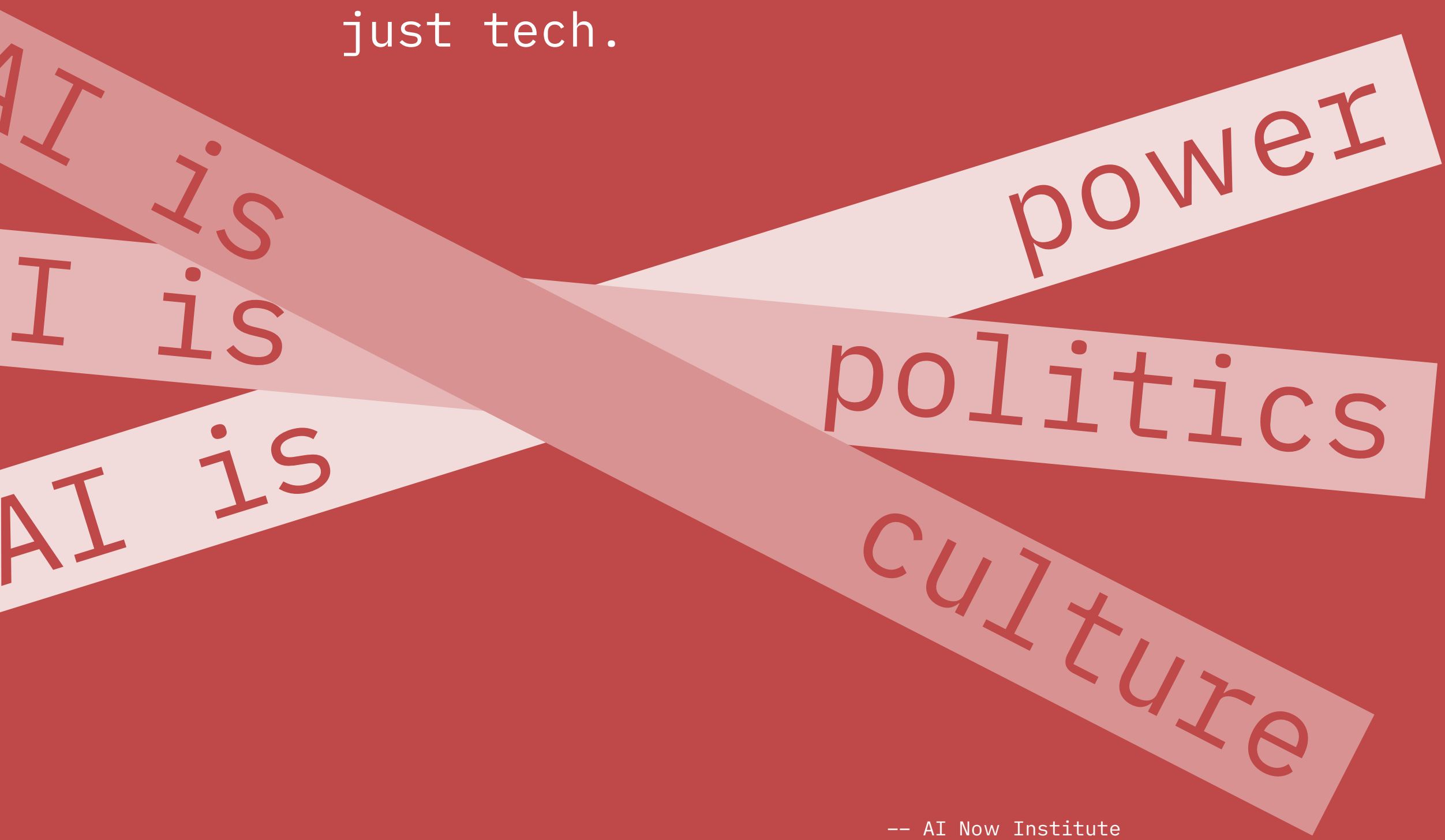
So is it all just bad news? No, fortunately, there are also beneficial side-effects. But it did draw attention to the need for regulation. Large tech companies, like Microsoft and Amazon, have explicitly called on the US government to regulate technologies like facial recognition.

> *» Because AI isn't just tech. AI is power, and politics, and culture. «*
>
> -- AI Now Institute

## Growing resistance

In 2019, the *AI Now Institute* published another overview of important news about the social implications of AI and the tech industry, and it was different from the year before. There appeared to be a growing resistance to hazardous AI applications. A selection from the 2019 database: an acceleration of regulatory measures (including a ban on the use of facial recognition in San Francisco), increasing pressure from shareholders (like the pressure from the shareholders of Amazon against the sale of facial recognition software to governments) and an increase in the number of protests and strikes of employees of large tech companies (like the climate strike of thousands of employees of, among others, Amazon, Google and Microsoft against the damaging effects of AI applications on the environment). The resistance in 2019 reminds us of the fact that there is still room to determine which AI applications are acceptable and how we want to control them. But again, what is desirable appears to be overtaken by reality.

AI isn’t
just tech.




-- AI Now Institute

Because of the rapid spread of the corona virus in 2020, governments are using data-driven apps to monitor the spread of the virus and apply *lockdowns* in a more targeted fashion. People have a track and trace app on their smartphone; when they have been in the vicinity of an infected person, they are notified. Needless to say, that is at odds with our ideas about privacy and, perhaps more importantly, with our views on self-determination and equality. What if people are forced to be quarantined on the basis of a false positive? In the Netherlands, the introduction of these apps met with considerable resistance. At the same time, it appears that many people fail to realize that, when it comes to privacy, these apps may be a lot more friendly than the apps by Facebook or Google that we have been using for years without expressing any privacy concerns. At any rate, it is important to keep asking questions about who benefits from AI, who is disadvantaged by AI and who can and should be allowed to make decisions about that.









1.2 ----------

## Urgent ethical issues

Within the development of AI ↴, the spotlight is firmly on ethics ↴. Questions about explainability ↴ and biases are everywhere. But what exactly do these concepts mean? And how do they express themselves in practice? To be able to answer those questions, we first need to understand which ethical questions are involved and what the potential considerations are. When mapping ethical issues, three central concepts are often used, namely responsibility, freedoms and rights, and justice (Van Dalen, 2013).

### Responsibility

When talking about the development of AI, people often mention the term ‘responsibility’. Most people, for example, have at some point heard the question who is responsible in case of an accident involving an autonomous car. Is it the passenger, the developer or the system itself? We previously only met machines as independent actors in science fiction movies. This leads to a variety of new issues, both legally and socially. The time for armchair philosophers has passed. These issues are now part of reality. But to determine who is responsible, we first need to determine what it is they are responsible for and what behaviour that does and does not include. The question then becomes how we can deduce the level of responsibility.

#### *What are we responsible for?*

A well-known ethical thought experiment is the so-called trolley problem, where the main question is if it is ethically right to sacrifice the life of one person to save the lives of many. To visualize that question, the experiment uses a tram or trolley.

*'A runaway trolley is moving headlong towards a group of five railway workers. You can still intervene by pulling a switch and moving the trolley onto a different track, where there is only one railway worker. Do you save the lives of five people by pulling the switch or do you save the life of one by not pulling the switch?'*

Of course this raises a number of questions. Are five lives worth more than one? Are you responsible when you intervene? And what about if you don't intervene? It is interesting to see whether the decision includes the identity of the five people that may be saved and the one person being sacrificed. Do people make the same decision if that one person is their loved one and the five others are strangers? Or when that one person is a young doctor and the other five are senior citizens? The question then is no longer about quantity, but about quality. And that is a question that is very difficult to automate in AI.

With the arrival of the autonomous car, the trolley problem isn't just a mere thought experiment; we now live in a world where those kinds of situations can actually occur. After all, the semi-autonomous cars that are currently allowed on our streets can break and switch lanes on their own, giving rise to the question what a car should do, for example, when a group of people is crossing the road and the car cannot break in time. Should it keep driving, to protect the passengers, or swerve and avoid hitting the people crossing the road but killing the passengers? And, of course, the question is if it makes a difference who crosses the road and who is inside the car. To answer that question, MIT developed the Moral Machine to see what people would do in such a situation. The experiment started in 2016 and by now has been filled in by over 40 million people. The results were published in 2018 in *Nature*. An analysis of the results showed that there are general preferences in some areas, for instance saving young people over older people and saving people



over animals, while other judgments are culturally determined. Participants from Central and South America tend to save women over men and people with an athletic body over obese people, while participants from countries with a high level of income inequality tended to save people with a higher social status over people with a lower social status.

In short, what we consider to be 'responsible' cannot be objectified completely. However, the question is whether that warrants stopping the development of the self-driving car. After all, how often do these types of dilemmas occur in everyday life? And how does that relate to the number of accidents that can be prevented with autonomous vehicles? Especially when everything is connected (so not just cars to each other, but all people in traffic, including pedestrians and cyclists), autonomous systems are able to look ahead more quickly than people can and anticipate potentially dangerous situations. The trolley problem is a good way to philosophise and to shed light on the complexity of ethical dilemmas, but it is still a thought experiment. In practice, developers are more focused on making the self-driving car as safe as possible. You can also look at the route the car has taken before the dilemma occurs. In the case of the trolley problem, the focus is too much on the existing context: why would we create an environment for autonomous cars where vehicles and pedestrians cross paths in the first place? Elon Musk, for instance, is working on an underground form of mobility with 'The Boring Company' where people can be transported by autonomous vehicles in an underground tunnel, completely bypassing the trolley problem altogether.

» *The discussion about the trolley problem shouldn't be about the forced choice, but about optimising safety.* «

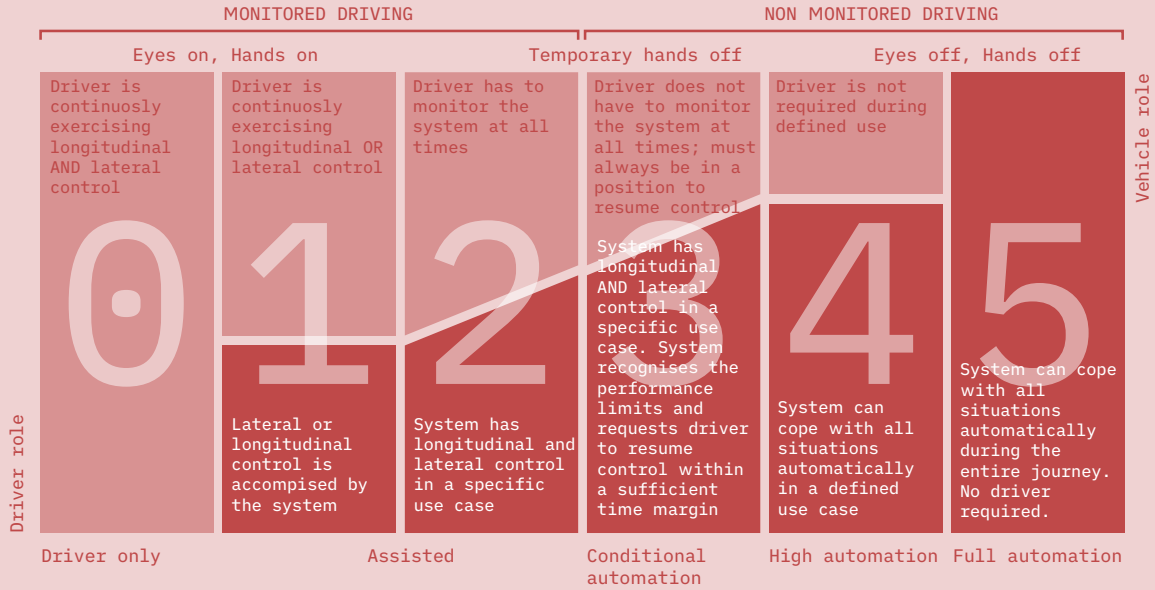
-- Arjen Goedegebure, OGD ict-services

*Who is responsible?*

But even if the trolley system does not occur, accidents can still happen as a result of system errors of autonomous vehicles and the question remains who is legally responsible, for example in case of a collision. Legally prescribed responsibilities are called liabilities. Incidentally, the liability issue is not completely new for AI. Dutch legislation uses the principle of 'product liability'. In Article 6:185 of the Dutch Civil Code, product liability is described as follows: 'The manufacturer is responsible for the damage caused by a defect of his product'. When the driver of a regular car can demonstrate that he suffered injuries and that those injuries are the result of a defective product, the car manufacturer is liable.

According to information from the Dutch government, the rules that apply to normal cars also apply to tests involving self-driving cars. 'The driver is responsible if he is driving the vehicle himself. But if the system does not function properly, the manufacturer is responsible'. According to the government, this legislation is sufficient for the test phase. After all, in the case of the semi-autonomous cars that are currently on the market, the driver is expected to pay attention and intervene when something goes wrong. Various car manufacturers have announced, however, that they intend to market completely autonomous cars in a few years. In that case, product liability will play a bigger role in the future, because fully autonomous cars take over the drivers' tasks, which means that safeguarding their safety is also increasingly the system's responsibility. Despite the reasoning that autonomous vehicles can significantly increase traffic safety, the systems will have their own challenges, for instance in situations where the hardware fails, there is a bug in the software, the system is hacked, the interaction between man and machine falters or when the system fails to properly anticipate other vehicles and people in traffic or unexpected traffic situations (Van Wees, 2018).

However, at the moment, the guidelines surrounding product liability appear to be insufficient for the introduction of autonomous vehicles. For example, right now, it is unclear whether software also falls under different guidelines. The principle of product liability is focused on movable property (like an autonomous car), but the question is whether, legally speaking, software is included in that definition. When autonomous vehicles are introduced from level 3, a paradoxical situation occurs. According to the guidelines of the *Society of Automotive Engineers* (SAE), vehicles below level 3 can, under specific circumstances, drive autonomously and the driver isn't needed to keep an eye on the circumstances. However, the driver has to be able at any moment to intervene, when the vehicle indicates that it is necessary. But a driver who does not have to pay attention hardly seems able to intervene when called upon.




'On the road to automated mobility: An EU strategy for mobility of the future 2018'
– Source: SAE (2019).

In addition, when people no longer have to drive themselves, the driving skills will decrease over time. The question is whether people are still able to intervene when the system demands it. In most cases this will involve complex situations. In that respect, a self-driving car requires a more competent, rather than a less competent driver. This principle also applies in other situations. For example, if a human doctor has to intervene with a robotic surgeon, he must also have his surgical skills up to date and must have knowledge of the complex system. This makes it difficult to guarantee the principle of *human in the loop* (HITL).

*» The curse of automation: the need of higher skilled operators. «*

-- Edgar Reehuis, Hudl

Nevertheless, the development of autonomous systems continues to boom. Tech giants like Google and Alibaba claim to be developing level 4 and even level 5 vehicles, which no longer require a human driver. Waymo (Google's erstwhile *self-driving project*) has been experimenting with fully autonomous vehicles in some suburbs of Phoenix since 2017, so far always with a human behind the wheel, who could intervene when





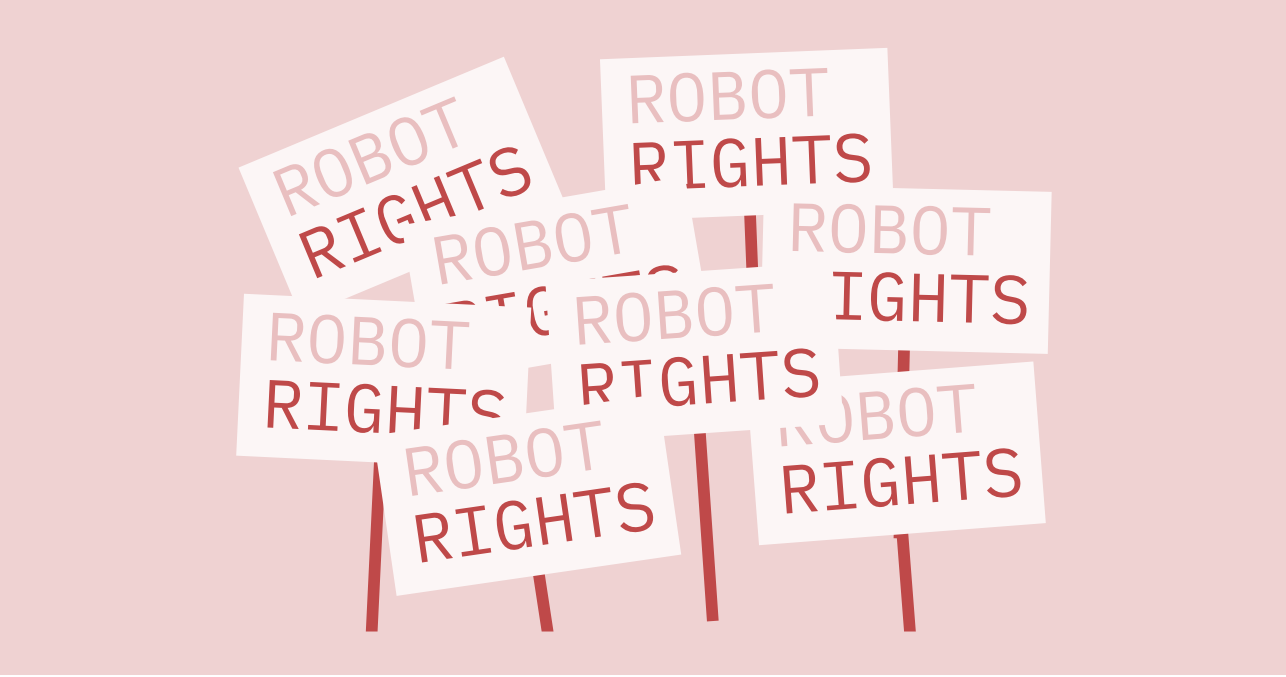
things went wrong, but in 2019, the company announced it will start offering paid taxi rides without a human driver. Of course, traffic is slightly more predictable in the suburbs of Phoenix than it is in the centre of Amsterdam, but it is still a breakthrough.

*» On the road soon: self-driving robot cars without a spare human behind the wheel «*

-- Bard van de Weijer, Volkskrant

Because self-learning systems are involved (where choices are not programmed entirely, but are based on data and experiences of the system itself), the manufacturer cannot be held completely responsible. And without a human driver, the passenger can also not carry all the responsibility. In that case, it should be possible for the system itself to be held to account. That may sound like science fiction, but it isn't. In 2016, Google's self-driving car system was officially recognized in the US as '*driver*' by the *National Highway Traffic Safety Administration* (NHTSA), essentially classifying the AI-system as the car's driver.

In the same year, the European Commission for the first time introduced the term '*e-personhood*'. The term was used in a report to describe the legal status of the most advanced autonomous robots. According to the report, these systems would be able to obtain certain rights and obligations to be able to be held accountable for damage being caused (Delvaux, 2016). And yes, in addition to obligations, the systems would also be given rights. Does that mean that, in the future, systems could also hold humans accountable?




*How can we deduce responsibility?*

As AI will make more and more autonomous decisions (and as a result, will be held accountable more and more as well), it is important to gain insight into the way the decisions of such systems came about. Algorithms ↴ increasingly make the clusters ↴ themselves, without pre-programmed labels, making it increasingly difficult for people to determine on what the decisions were based. As such, people often compare AI to a black box. An important question in this regard is whether the decision-making process of the future is sufficiently transparent and whether the results can be sufficiently explained, which is why a lot of attention is paid to *Explainable AI* (XAI).

A distinction has to be made between transparency, the explanation and explainability. Transparency is mostly about the process and the prior criteria, while the explanation and explainability refer to the explanation and deducibility of the decision afterwards. Explainability is very subjective and context-dependent, while transparency and explanation are more objective. As such, putting explainability in practice is difficult. It involves a move from a black box towards a 'glass box'. In practice, it means coming up with an explanation in retrospect. However, systems that are currently being developed, for instance with the help of deep learning ↴, are so complex that people can hardly understand them. You can't just ask a system for an explanation and it will print out a chronological explanation of its main considerations and decisions. It is a multi-layered web of connections, much like the human brain ↴, come to think of it. We can't look under 'the hood' of the human decision-making process either. When someone causes a traffic accident, we also have to make do with an explanation that is constructed in retrospect. We cannot look into a person's brain and determine exactly what led to the decision that was made.

*» It is easy to put transparency and explainability on the agenda, but putting them into practice is enormously complex. «*

-- Marijn Janssen, Delft University of Technology

Transparency is a misleading concept anyway; it is always a limited representation of reality. Think of a window, for instance. You can look outside through the glass and it seems transparent. But you can see what is visible within the borders of the framework. That leads to a *governance*-related problem. Everything has to be checked and audited; not only the output, but also the decision-making rules and input. However, it is almost impossible to ascertain what the exact origin of the data is, because data is already pruned and reduced before it is used. This so-called *data cleansing* makes it hard to reproduce the process even before it starts.

In today's discussions about the applicability of AI in decision-making processes, people fail to draw enough of a distinction between the different types of decisions and the impact that the decisions have on the people involved. There's a significant difference if it involves a recommendation for a movie, a medical diagnosis on the basis of lung X-rays or an evasive action of an autonomous vehicle. The severity of the impact depends, among other things, on the potential risks. Which is why we need 'levels of explainability'. For instance, there's a less need for explainability in the case of chatbots than in the case of self-driving cars or war drones. As such, the need for explainability and the consequences of making the wrong decisions depend on the context and on the type of decision.

This is closely related to the degree of autonomy that we will grant to the system and thus the relationship we have with technology. Is it 'just' a tool or does it make completely autonomous decisions? In that regard, AI still has a long way to go. Research from the University of Amsterdam into automated decision-making by AI from 2018 shows that many Dutch people are concerned that AI may lead to manipulation, risks or unacceptable results. It is only when more objective decision-making processes are involved, such as a mortgage application, that they feel AI has potential. Human control, human dignity, honesty and accuracy ↴ are considered to be important values when reflecting on decision-making by AI.

*» There may well come a day when we tell AI 'explain yourself', and AI responds 'you wouldn't understand it anyway'. «*

-- Maarten Stol, BrainCreators









## Freedoms and rights

At a fundamental level, freedom is about the freedom people have to determine how they want to organize their lives. It is even a right, the right to self-determination, which is then limited by the duty not to harm others. So you don't have the freedom to break into other people's homes. The formulation of rights is built around the individual's freedom. In some cases, people need to leave you alone or even not do certain things to protect your freedoms (freedom rights), while in other cases, others – usually governments – actually have to make an effort to allow you to realize your freedoms (social rights). Freedom rights and social rights are recognized internationally and have been included in the Universal Declaration of the Rights of Man. The emergence of AI begs the question to what extent we are still completely free and to what extent certain rights are being curtailed, for instance when we are increasingly being monitored with facial recognition software.

*To what extent are we free to make our own choices?*

The discussion surrounding AI often includes the nightmare scenario ↴ in which intelligent robots can take over control and we lose any form of self-determination. That nightmare scenario is based on the assumption that a separate system will have the capacity to exceed human intelligence ↴. What is often overlooked is that, if very powerful specialized systems are connected to each other, that can also create an intelligent system. Instead of '*General AI ↴ gone bad*', it becomes '*Narrow AI ↴ everywhere*'.

Cities are increasingly filled with sensors, which increasingly allows intelligent systems to make independent decisions, even without human intervention. These days, any self-respecting city calls itself a *Smart City*. In other words, a city that uses different kinds of data to optimize processes and tackle problems in areas like traffic, safety and the environment.

Think, for instance, of cameras using facial recognition software that make it possible to ban hooligans from football stadiums or to monitor social media to map and manage tourist flows in the city.

*» In the city of the future, lampposts take part in the conversation, but citizens do not. «*

-- Maurits Martijn & Sanne Blauw, The Correspondent

In that sense, democracy can change in a so-called 'algocracy', where cities (and therefore people) are managed by data. More and more experts warn us about a *black box society*, in which the choices that are made by smart algorithms can no longer be traced, which means that citizens can decide less and less whether or not they want to be a part of this data-driven society. Technology can even be used to create a totalitarian state; Big Brother is watching you. China, for instance, is slowly rolling out a 'social credit system', where Chinese citizens are given a certain score based on their behaviour. On the basis of that score, people can be placed on a blacklist and lose all kinds of rights and privileges. In 2018, 23 million Chinese were banned from buying a train or plane ticket. In addition, restrictions on Internet usage are imposed in more and more places, for example in Pakistan where in 2020, the authorities approved far-reaching new rules that restrict the use of social media, endangering the people's freedom of speech.

The right to make one's own choices can, of course, also be approached from the opposite position. In several areas, AI is (or will be) better than people, in particular when it involves very specialized applications. This goes way beyond extremely powerful chess computers. Even now, algorithms are better at recognizing cancer on lung X-rays than doctors. So the question can then be if certain tasks should even be left up to humans. In that sense, people should be allowed to choose an artificial system over a person.





*Do we have the right to be forgotten?*

AI systems penetrate ever deeper into our lives and sometimes clash with human rights. Think, for instance, of the System Risk Indication (SyRI) that the Dutch government uses to combat fraud in the area of subsidies, taxes and government allowances. On the basis of information involving, among other things, work, income, pensions and debts, the system calculates who might commit fraud. In particular in vulnerable neighbourhoods. An important objection to such a system is that the data of all the people living in a neighbourhood can be analysed, even if they are innocent, making them guilty until proven innocent. Various civil rights and privacy organizations felt that such a system was unacceptable and sued the Dutch State. And with success. In 2020, the courts ruled that the legislation om which the use of SyRI is based violates article 8 of the European Treaty for Human Rights (ETHR), namely the right to respect for our private lives, which requires a fair balance between the social interests served by the legislation and the extent to which it affects our private lives. The courts ruled that the prevention and restriction of fraud was outweighed by our right to privacy.

In 2017, the Rathenau Institute argued in favour of a new European treaty that would update human rights and adjust them to the digital society. The report even mentions new human rights, including the right not to be measured, analysed or influenced (the right to refuse online profiling, tracking and influencing). And not without reason. Applications in the area of, for instance, facial recognition ↴ put more and more pressure on our right to privacy. A good example is the case surrounding the company *ClearView*, which 'scrapes' millions of pictures from Facebook and other sites for the benefit of its facial recognition software and which offers its services to numerous intelligence agencies. Or the singer Taylor Swift who used facial recognition technology to identify potential stalkers

in her audience. Cameras don't even have to be in the neighbourhood any longer. At the moment, facial recognition systems are being developed for the military that can identify people up to a kilometre away.

A version of the right not to be measured became a reality in the European General Data Protection Regulation (GDPR), a piece of legislation that includes the Right to be Forgotten'. The aim is to give people control over their personal information and to see what it is exactly that companies do with that information. That means that all organizations explicitly have to ask for permission to collect and use people's personal data. It also means that users have the right to know what information a given organization processes and how that information is secured. Users also have the right to ask organizations to remove all their personal information. Organizations that fail to comply risk hefty fines. In 2019, the Swedish Authority for Data Protection fined a school for using facial recognition technology to check school attendance. It involved a fine of 200,000 Swedish kroner, which is almost 19,000 euros.

There are exceptions, however. For instance when an organization is legally obliged to use data, for example the data required for a legal ordinance or protecting public health. This makes the discussions surrounding apps that are used to monitor epidemics, for instance during the corona-crisis in 2020, even more intense. In addition, the right to be forgotten does not reach beyond the boundaries of the European Union, according to a ruling by the European Court of Justice in 2019. The balance between the right to privacy on the one hand and the freedom of information of Internet users on the other will, according to a statement by the judge, vary considerably around the world. That means that search engine giants are not obliged to remove information outside the EU countries. As such, in the future, privacy could become a luxury that is only available to a small number of people.






*How can we take the law into our own hands?*

Governments increasingly are in the news for 'spying' on us, which means we are increasingly living in a 'surveillance society'. As early as 2013, former CIA employee Edward Snowden alerted the world to the large-scale surveillance by the American National Security Agency (NSA). Big Brother was indeed watching you. Inspired by those developments and the quote from George Orwell's *1984*, an Australian clothing brand marketed a specially developed clothing line that hides your telephone. The main feature of the 1984 clothes line is the so-called 'UnPocket', a canvass pocket interwoven with special metal materials that blocks Wi-Fi and GPS signals, among other things. If that makes you safe from the NSA remains to be seen, but it does offer people sufficient protection from location tracking.

» *I don't know why people are so keen to put the details of their private life in public; they forget that invisibility is a superpower.* «

-- Banksy

Rather than wait for revised legislation (which is not always sufficient), creative entrepreneurs more and more develop apps and gadgets to help us protect our own privacy. Other examples show that the fashion industry is also very active in this area. For instance the graphic prints on clothes that confuse surveillance technologies or stylized facial masks that make your face unrecognizable for facial recognition software. There is even a clothing item designed to prevent us from being photographed all the time (and often unintentionally posted on social media – without being asked). The designer of Dutch origin Saif Siddiqui developed the ISHU scarf, a special scarf that reflects the flash lights of smartphones thanks to the tiny crystal balls embedded in the scarf, making the pictures unusable.

## Justice

When we talk about justice, essentially we are talking about the equality of people. People should be treated equally *and* be given the same opportunities. That does not mean that there aren't or shouldn't be differences between people. But when people are treated differently, there has to be a demonstrable reason to justify that difference, for instance differences in pay based on experience and education. The question is whether an equal distribution and treatment can be safeguarded with the emergence of AI.



*What is a fair distribution?*

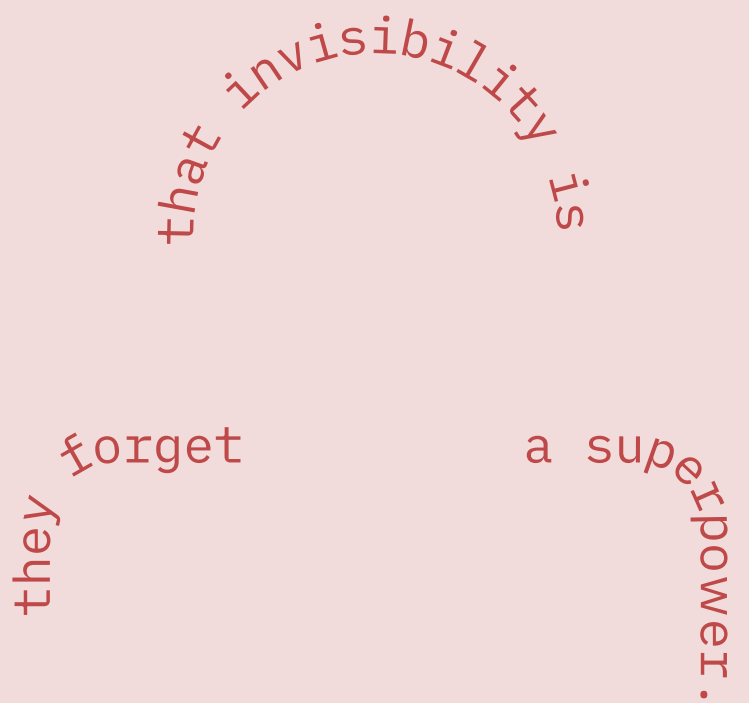
AI is used increasingly to assess credit applications, detect fraud and evaluate job interviews. The starting point is that people are given equal chances and opportunities. However, in practice, that often isn't the case at all. For instance, in 2014, Amazon developed an AI application to evaluate applicants and select the best candidates. It was only a few years later that they discovered that the application was sexist. The problem was that the algorithm was trained with data from the people who had applied at Amazon for the past 10 years. And because, in the tech sector, most of them are males, the algorithm developed a preference for male candidates. The development of the application was discontinued in 2017. To experience the unequal distribution of the algorithms for themselves, four alumni of *NYU Abu Dhabi* developed the game 'Survival of the Best Fit'. The educational game exposes the prejudices of AI ↴ in the application process.

So in those kinds of processes, prejudices – and digital discrimination – are actually reinforced. The problem is that data is not diverse enough and, as a result, the technologies are not neutral. Research from the *Georgia Institute of Technology* shows that even the best image recognition systems are less accurate in detecting pedestrians with a dark skin than pedestrians with a light skin. The researchers indicated that that bias was caused predominantly because of the fact that few examples of pedestrians with darker skins were used in training sets. When the input is incomplete, then so is the output. 'Garbage in, garbage out'.

The question is whether there is enough data from the more vulnerable groups. It appears that prejudices against the handicapped are even more tenacious than discrimination on the basis of gender and race. For instance in the case of self-driving cars. The algorithms are trained, among other things, to recognize pedestrians, to avoid cars from driving over

I DON'T KNOW WHY PEOPLE ARE SO KEEN TO PUT THE DETAILS OF THEIR PRIVATE LIFE IN PUBLIC

-- Banksy



them. So when the training data do not include people in wheelchairs, the technology can place them in hazardous positions. And, although handicaps are relatively common, there are many different kinds of handicap. And they are not always visible. Cars can honk their horn to warn approaching pedestrians, but deaf people won't hear them. In addition, information about handicaps is very sensitive. People are more reluctant to provide information about their handicap than information about gender, age or race. In some situations it is even illegal to ask that information.

In 2018, Virginia Eubanks wrote a book entitled '*Automating Inequality*', describing the way in which automated systems – rather than people – determine which neighbourhoods are checked, which families are given the necessary resources and who will be investigated for fraud. Especially people with fewer resources and opportunities are disadvantaged by those systems.

*» Data protection is not a private but a general interest and is at the heart of the constitutional state. «*

-- Maxim Februari, philosopher and writer

*To what extent are people treated fairly?*

The fact that people need to be treated equally and given equal opportunities is something that is embedded in the Dutch Constitution. Paradoxically enough, it is becoming increasingly clear that the AI systems that are used within our legal system actually promote inequality. Think, for example, of developments within *predictive policing*, whereby criminal behaviour is predicted using large-scale monitoring and data analysis. The risk is that the wrong people are apprehended. What is just in those cases? Do you risk apprehending innocent people or do you risk them committing a crime? The technologies being used, like facial recognition software, are far from perfect. In a test conducted by the *American Civil*





*Liberties Union* (ACLU) in 2018, it turned out that the software of Amazon wrongfully identified 28 members of Congress as people who had been arrested for committing crimes.

In the US, software is often used to predict the likelihood of people becoming repeat offenders. Research by *ProPublica* from 2016 shows that the software being used is prejudiced against people with a darker skin colour. It is extremely difficult to estimate how autonomous systems will behave when they are interconnected, for instance with regard to negative feedback loops. When the data shows, for example, that black men are more likely to end up back in prison after being released, the algorithm will determine that black men should remain in prison longer, which in turn affects follow-up figures: black men will end up serving longer prison sentences, which is once again reinforced by the algorithm, etc. This does not take into account that the figures are biased as a result of human police work, which may well involve profiling. So the biases in algorithms are caused predominantly by biases ↴ in people, which begs the question if a robot judge is indeed more objective.

*» Algorithms are as biased as the people making them. «*

-- Sanne Blauw, The Correspondent

*How can we safeguard fair procedures?*

The question who can and may decide about what is and isn't ethically responsible in relation to the development and application of AI is as yet not asked often enough. So far, it is especially the large tech companies that appear to be in control. They possess by far the most resources, like data, computing power, money and knowledge. Especially the importance of knowledge is often underestimated. Think, for example, of the time when Mark Zuckerberg was questioned by the American Congress in 2018 about the privacy leak

at Facebook. Senators asked him questions that clearly showed they had no idea what exactly Facebook is and does. The question by Utah Senator Orrin Hatch showed it most clearly, when he asked how Facebook was able to make any money, when its users didn't have to pay for it.

*» How do you sustain a business model in which users don't pay for your service? «*

-- Orrin Hatch, US Senator

To which a visibly stunned Zuckerberg responded: 'Senator, we run adds'. This kind of digital illiteracy makes it difficult to regulate complex technologies. However, governments still appear to play an important role in the application of AI, for instance when in 2019, Russian president Vladimir Putin signed a controversial law making it a crime to 'show contempt' for the state and to spread 'fake news' online. In the same year, the Iranian government shut down the Internet during protests against the increasing inflation, making it harder for demonstrators to organise and for journalists to obtain information about the situation. It appears that countries want above all to protect their own digital infrastructure and power.






We see the same inward-looking attitude when it comes to ethics. Various countries have launched their own ethical guidelines ↴. From Germany and Austria to Australia and the United States. The question is not which country has the best ethical guidelines, the important thing is to see where they are in agreement. Ethics isn't a contest, but a team sport. Without a global approach, it is almost impossible to develop reliable AI applications. The different approaches, which are often culturally determined, have to come together. At the moment, these approaches are missing in the global debate about AI and ethics and, as a result, the development of AI looks more like a contest ↴. The financial interests appear to be more important that the moral interests.

In addition to the unethical implications of AI, we also have to think about designing ethical guidelines, or otherwise the ethics will be unethical.

## A matter of ethical perspective

Within the development of AI ↴, there are diverging ethical ↴ issues. Who is responsible in the case of a collision with a self-driving car? How important is the right to privacy ↴ in relation to the right to information of Internet users? And to what extent can a robot judge produce an objective verdict? These issues are far from unambiguous and, as a result, very complex. For instance, we may not even want a robot judge to be objective. What people consider to be fair cannot always be captured in a formula and very much depends on the context and possible extenuating circumstances.

> To be able to assess such issues, it is important to understand that there are different perspectives. There is often no universally accepted truth, which makes it difficult to reach a consensus about what, ethically speaking, the best solution is. Generally speaking, we agree, for example, that privacy is important. But in 2016, a judge had to decide whether or not Apple should give the FBI access to the data on the iPhone of a terrorist. Not only are these issues context-dependent, the outcome also depends on the ethical perspective being used.

### Different kinds of ethics

In the discussions, ethics and morality often intermingle, but there is a clear difference. Morality is the totality of opinions, decisions and actions with which people (individually or collectively) express what they think is good or right, while ethics, on the other hand, is the systematic reflection on what is moral (Van de Poel & Royakkers, 2011). Different ethical guidelines ↴, like the guidelines of the European Commission, are in essence moral guidelines. Morality is about the actions themselves, while ethics is about studying those actions. Within the study of what is morally right, there are various subcategories.





Roughly speaking, we can distinguish two approaches, namely the normative approach and the non-normative approach.

*The normative approach*

Within normative ethics, there are clear moral positions. ‘Good’ and ‘bad’ are translated into general basic principles, designed to show people how to act and to serve as a basis for regulating people’s behaviour. This approach is also known as prescriptive ethics, because it provides people with rules and principles on how to behave.

*The non-normative approach*

With the non-normative approach, no moral positions are taken. There is a distinction between descriptive ethics, meta-ethics and applied ethics. Descriptive ethics is about describing and understanding what people consider to be ‘good’. In the case of meta-ethics, the focus is on studying the central concepts of ethics (responsibility, freedoms & rights and fairness). The aim is also to see whether it is possible to create an ethical framework that can be applied in any situation, regardless of our own opinions. Applied ethics focuses on specific domains, like bio-ethics, business ethics or medical ethics, and looks at ethical questions from a practical perspective. The question is to what extent principles from normative ethics can be applied and provide an answer to the concepts from meta-ethics.

> Because AI applications are used in various domains, it is not easy to create clear ethical frameworks. What is ‘good’ or ‘fair’ always depends on the specific context. Many people are unaware that, in addition to the different contexts, there are also different approaches and starting points that often blend together, which makes ethical discussions surrounding the applications of AI often a little messy.

## Different approaches within ethics

People find different things important, depending on the context. In some cases, the emphasis is on the action itself, while in other cases, the focus is more on the consequences of that action. And sometimes, it is all about the intentions of the person carrying out the action. As such, there are different approaches and opinions within ethics that are sometimes each other's exact opposites. What is 'good' and what are 'good actions' are questions that people have thought about for centuries. And with the arrival of AI, those questions are more relevant than ever before. Technologies are becoming increasingly autonomous and have to be able to act autonomously with regard to those issues.

### *Principle ethics*

In the case of principle ethics, a principle is always used as a starting point, for instance respect for life and human dignity. The solution of an ethical problem has to observe one or more of those principles. The principle has to be applied at all times, regardless of the consequences, so people's actions are considered to be moral as long as they observe said principles. Some actions can be considered to be good, even if their consequences are negative. And vice versa. These behavioural rules, or values, are agreements about how we treat each other. Although there are often no concrete sanctions for violating these values, they are maintained by a society as a whole. Many religions contain such behavioural rules, like the 10 Commandments in the Bible. Some values have been formulated as laws, like laws prohibiting discrimination.

Principle ethics is also known as deontology. The most famous proponent of deontology is the German philosopher Immanuel Kant. In his '*Kritik der praktischen Vernunft*' in 1788, he formulated the 'categorical imperative'. The most well-known statement is 'Act in such a way that you would wish that your principle would be turned into a universal law of nature'. For instance, when you wonder whether you are allowed to throw waste out of your car window, it's easy to realize that, if everyone were to do so (because it is the law), all the world would be a mess, so it is not the standard.

*» Act in such a way that you would wish that your principle would be turned into a universal law of nature. «*

-- Immanuel Kant

### *Consequential ethics*

Consequential ethics states that the consequences of a given action determine whether or not it was 'right'. In other words, the consequences have to be positive, even if it undermines certain principles. So the action itself is not called into question, only its consequences. To assess the consequences of an action, values are used. A value is a goal that we want to achieve as a society through our actions, for example justice or freedom. They are also known as end-values. Qualities that help people realize those end-values, are also known as instrumental values, like helpfulness and responsibility.

Consequential ethics is also known as consequentialism or utilitarianism. The founder of utilitarianism is the British philosopher Jeremy Bentham. With utilitarian theory, the moral value of an action is measured by the contribution that action makes to the common good. So the question is to what extent the action contributes to the maximisation of happiness. The goal of an action is always to provide the maximum amount of happiness to the largest possible group of people. If, in exchange for that, a small group of people has to face negative consequences, that is considered to be acceptable.

*» The end justifies the means «*

*Virtue ethics*

In the case of virtue ethics, the moral focus is not on the rules or certain principles, but on the character of the person performing the action. Again, the action is separated from its explicit consequences. To be able to perform morally sound actions requires certain character traits, or virtues. A virtue is a positive character trait steering a person's behaviour. When virtues are used to assess a person's actions, the focus is not on any individual action, but on the person involved and his or her intentions.

Aristotle is considered to be the founder of virtue ethics. Unlike deontology and consequentialism, the human being is taken into consideration. According to Aristotle, good actions are actions that make you a better human being, which means that people have to keep working on themselves. Virtues can be developed. A virtue is seen as a kind of happy medium between more extreme behaviour characteristics. For instance, bravery is a virtue that lies between hubris and cowardice.

*» Doing a good deed is easy; developing the habit to always do that isn't. «*

-- Aristotle

Not every great thinker can be assigned to one of the three categories described above. The German philosopher Friedrich Nietzsche, for instance, was seen as an 'ethics critic'. He argued that ethical opinions at the time (1887) were based on a 'slave morality' and that people were docile and no longer thought for themselves. Nietzsche wanted people to free themselves from this (at the time above all Christian) morality. Man should design his own ethics and create his own values. According to Nietzsche, ethics is not a matter of duty or virtue, but of personal preferences.



Different opinions about what is ethically responsible

To be able to judge the various ethical discussions within the development of AI systems, it is important to determine the ethical principles being applied. The question where it is acceptable to sacrifice the life of one person to save the lives of several people (trolley dilemma) to a large extent depends on the starting point. A distinction has to be made between the action (deontology), the consequences of the action (consequentialism) and the person – depending on the outcome of the responsibility issue ↴, in this case, the person in the car, the manufacturer or the system itself – performing the action (virtue ethics).

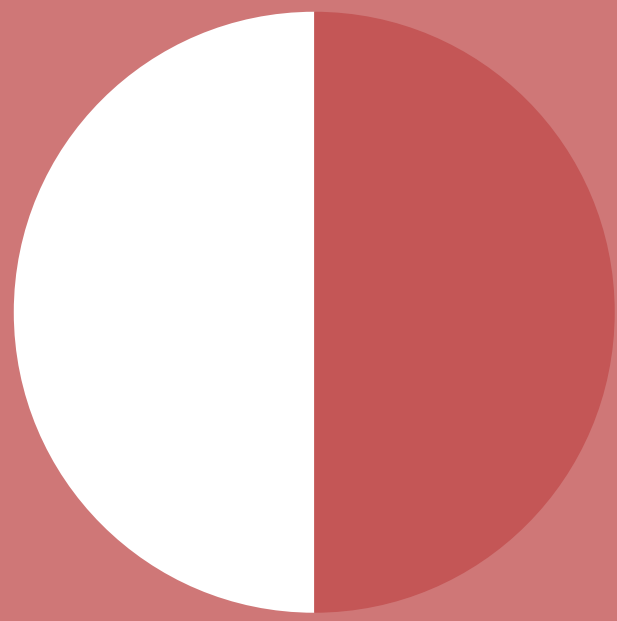
*What should a car do when a group of people is crossing a zebra crossing and the car cannot break in time? Continue to save the life of the person inside the car or swerve to save the people crossing the road, while killing the person inside the car?*

The question is, then, if it can be justified if the car does not intervene and protects the life of the person inside the car rather than the lives of the group of pedestrians. Within the framework of consequentialism, it cannot be justified, because it kills the group of pedestrians instead of one person. However, within the framework of deontology, it can be justified, because intervening would make the car deviate from its natural course and the system would be responsible for killing the person in the car, and killing someone is against the law, even if it saves the lives of others. Within the framework of virtue ethics, the opposite can be argued based on a similar starting point. When the car has the opportunity to intervene but fails to do so, it displays a lack of virtue and its actions are immoral. You always take other people into consideration and act responsibly.

developing
the habit to
always do that
isn’t.

-- Aristotle

Doing
a good deed
is easy;



However, these types of dilemma's are not always clear-cut. Think, for instance, about whether or not Apple should give the FBI access to the data on the iPhone of a terrorist, which was the central question in a court case in 2016. A year earlier, a terrorist shot and killed fourteen people in San Bernardino. At the time, the FBI confiscated the man's iPhone, to gain access to vital data, like information about contacts and possible accomplices. Of course, the iPhone is protected with a PIN code. When the wrong code is entered 10 times in a row, all the information is erased from the phone, which is why the FBI asked Apple to help circumvent this security. At the time, that was not yet possible, and nor did Apple want to help develop it. At face value, it appears to be a relatively simple question, assuming it was just about the one phone. But that was not the case. If Apple were to agree to the FBI's request, it would have to weaken the security of all iPhones. Apple had been working for years to improve security and this could do considerable damage to their reputation.

In addition, they wanted to make an example: when they allowed the FBI to gain access to the personal data of their users, Apple feared that would be the thin end of the wedge and other government agencies would make similar requests, which would potentially violate people's right to privacy. Which is why Apple argued that only the user has access to the data of a secure phone and nobody else.

It is hard to determine whether or not Apple's position was morally responsible. After all, it is far from clear that the right to privacy should always trump the war on terror. As such, it is unclear which action is moral in this case and what the interests are of the wider population. The right to privacy or the war on terror? It also has to do with the short-term versus the long-term impact. Here, it was just the one case, but in the future, the lives of many people could potentially be spared if the FBI was able to fight terrorists more effectively. Ultimately, the judge ruled that Apple had to help the FBI to unlock the terrorist's iPhone. Contrary to what was demanded earlier, namely to create a 'backdoor' for the FBI, Apple only had to guarantee access to that specific iPhone.

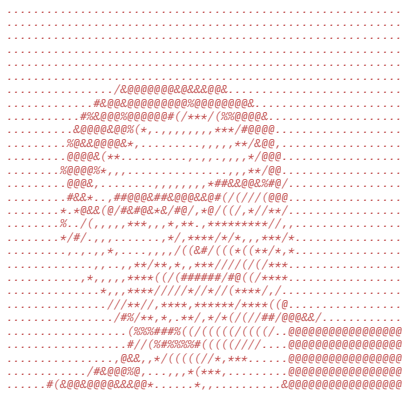




It is unclear whether the verdict provided justice for all involved. In many cases, there is no consensus about what the best solution is. Ultimately, it is about the question what the best society is and it will take us a while to figure that out. For example, what is better? A society in which a large group of people is marginally happy or a society in which a small group is very happy? There is no absolute philosophical theory. In the end, it should be society itself that determines what the best society is. It is a democratic issue. And even when we reach a consensus as a society about what we consider to be important, the question as to how you can integrate that into AI remains unanswered. But first things first. Before we can determine how we can integrate ethical guidelines into the design of AI applications, we need to map which ethical guidelines there are (deontology) and to what extent those guidelines endure in practice (consequentialism and virtue ethics).

*» My ethics are not by definition also your ethics. «*

-- Patrick van der Duin, STT



# AI governance - the good, the bad and the ugly

*By Marijn Janssen, Professor ICT & Governance, Delft University of Technology*

Artificial intelligence (AI) is used in various places to create a better society. It is used, for instance, to detect illnesses at an earlier stage and to tackle social problems (‘the good’). At the same time, AI us used to influence our political preferences and to commit large-scale fraud (‘the bad’), while it can also copy human prejudices and dictate everyday life without people noticing. AI monitors our behaviour to determine our health and insurance premiums and implicitly discriminates population groups (‘the ugly’, in other words the mean one). It is especially in the latter category that AI can be pernicious, because the unintended effects are not immediately visible and we are often unaware of them. Whereas ‘the good’ and ‘the bad’ are more explicitly visible, ‘the ugly’ is less visible and harder to ascertain.

The social call for transparent, responsible and fair AI is therefore a justifiable one. Often, it is argued that AI should be transparent, but what that means in practice is not addressed. Implicitly, it is said that we need to understand the algorithm, but is the situation without AI transparent and do we understand what happens in the human brain when people make decisions? Unknown makes unloved. Is it even reasonable to expect a complex algorithm to be transparent and that everybody will be able to fathom its complexity? Most people don’t even bother to read the conditions of an app or website they use and agree without thinking. Complex algorithms involve advanced maths. For most people, complete transparency is an illusion and the question is how we will deal with AI as a society.

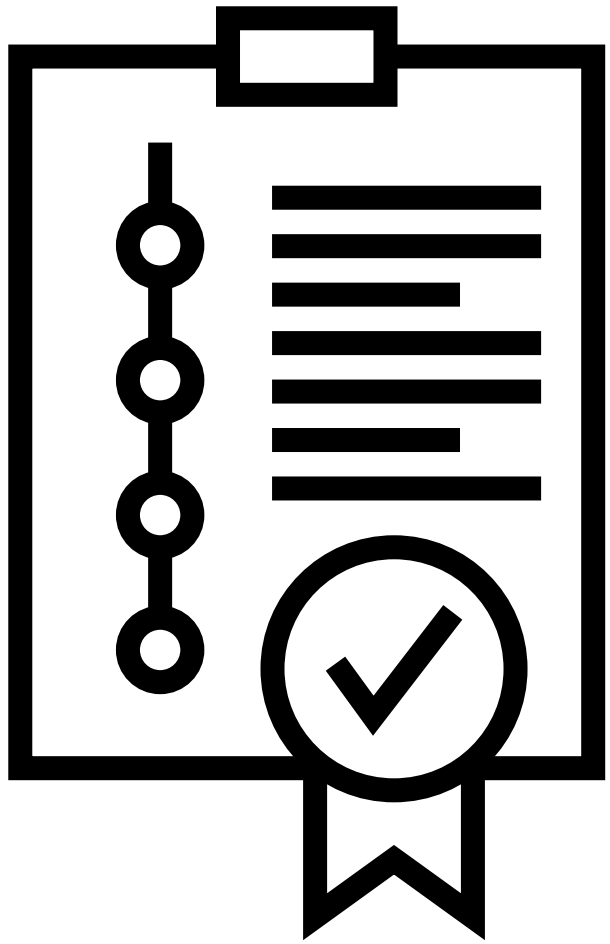




In the science fiction show Star Trek, the Vulcans have a solution. “Logic is the cement of our civilization, with which we ascend from chaos, using reason as our guide” (T’Plana-Hath, Matron of Vulcan Philosophy, Star Trek IV: The Voyage Home, 1986). Logic should lead to a discussion about prejudices, making sure we have all the facts and preventing us from seeing only what we want to see. The logic that the Vulcans have adopted leads to a technocratic perspective and also has its disadvantages, because there are always people playing politics and gaming the system. In addition, information is almost never complete and not all snakes in the grass are visible. We need logic, but we also need a suitable governance model.

The sensitivity of algorithms to change and their capricious nature because of the many variables make it hard to govern AI algorithms. Such a governance model not only needs to steer the technology, but also comprehend it. And that takes more than technological know-how alone. It takes a governance model that contains a well-defined shared model on how to deal with AI. Before people are allowed to use AI, it should first be tested, the way we only market medicines after rigorous testing and insight into possible side-effects. AI systems have to meet requirements to prevent the creation of a surveillance state, where all of our actions are affected by AI without our knowledge or consent.

Furthermore, organizational knowledge is needed to understand the ethical consequences. In the case of AI-driven autonomous cars, we may no longer need a steering wheel, but that doesn’t mean that the car is no longer being driven. The governance model has to make sure the bad and ugly sides of AI don’t occur. At the moment, AI governance is not yet mature, even though we are already using AI on a large scale. Let’s use AI governance to move toward an AI society where people are in the driver’s seat and computers provide support.



# 2 Ethical guidelines for AI



Pages: 30
Words: 6427
Reading time: approx. 50 minutes

## Ethical guidelines for AI

The time when AI systems ↴ predominantly made the news by beating chess grandmasters appears to be far behind us. Of course, at the time, people wondered, if computers can beat us in chess or Go, in which other areas would they be able to beat us? What does that do with the relationship between people and technology? And what does it mean for our humanity? Interesting questions, but still fairly abstract and philosophical in nature. Since then, AI systems have been applied in a wide variety of domains and we are faced with a growing number of issues involving applied ethics. It's no longer a question of 'what if?', but of 'what now?'.

> For instance, autonomous vehicles have made their first lethal victims, our right to privacy ↴ is undermined by the use of location apps and entire sections of the population are disadvantaged by fraud detection systems. Now the hypotheses have been confirmed, it would appear that we have woken up. From the private sector to social organizations and governments, they have all began to formulate ethical guidelines. A good and important first step. Especially for proponents of principle ethics. It is interesting to see what the similarities and differences between these guidelines are.

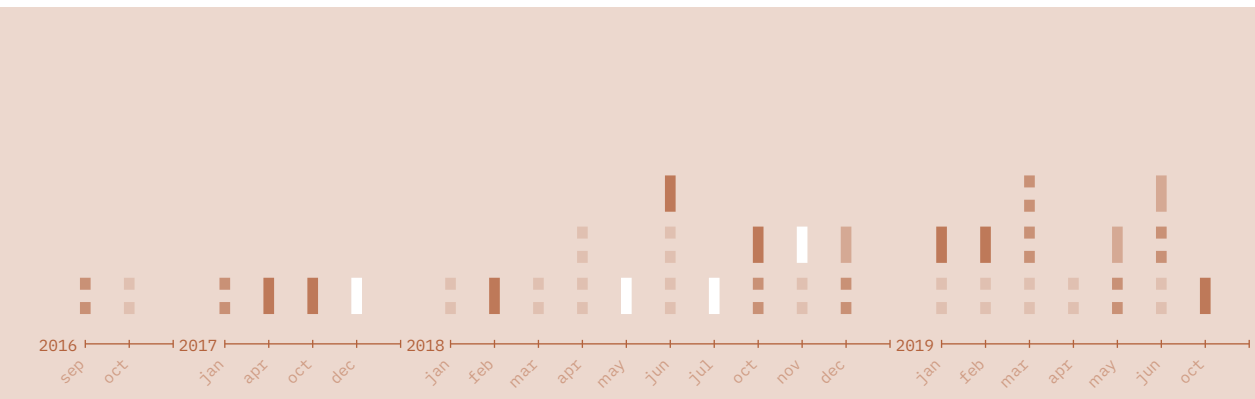
The question is, however, whether the development of guidelines is enough. Ultimately, those guidelines have to be translated into practice. And, as always, practice is less malleable than theory. For instance, how do we deal with conflicting values? And what challenges still await us? It is important to also look at the consequences. Fortunately, there are more and more organizations that have been working on a variety of checklists, assessments and toolkits, which brings us a step closer to putting ethics ↴ into practice, although we need to make sure that these tools really help us along in the design process. Do these tools actually allow us to act ethically?

## From corporate to government

In recent years, various companies, research institutes and government organizations have set up different principles and guidelines for ethical AI, at a national, continental and global level.

### Emergence of ethical guidelines

To create order in the fragmented discussion about the development of ethically responsible AI applications, researchers of the *Berkman Klein Center* in 2020 carried out an analysis of the 36 most prominent AI guidelines, which they also translated into a clear timeline.




'A map of Ethical and Rights-based Approaches to Principles for AI'
– Source: Berkman Klein Center (2020).

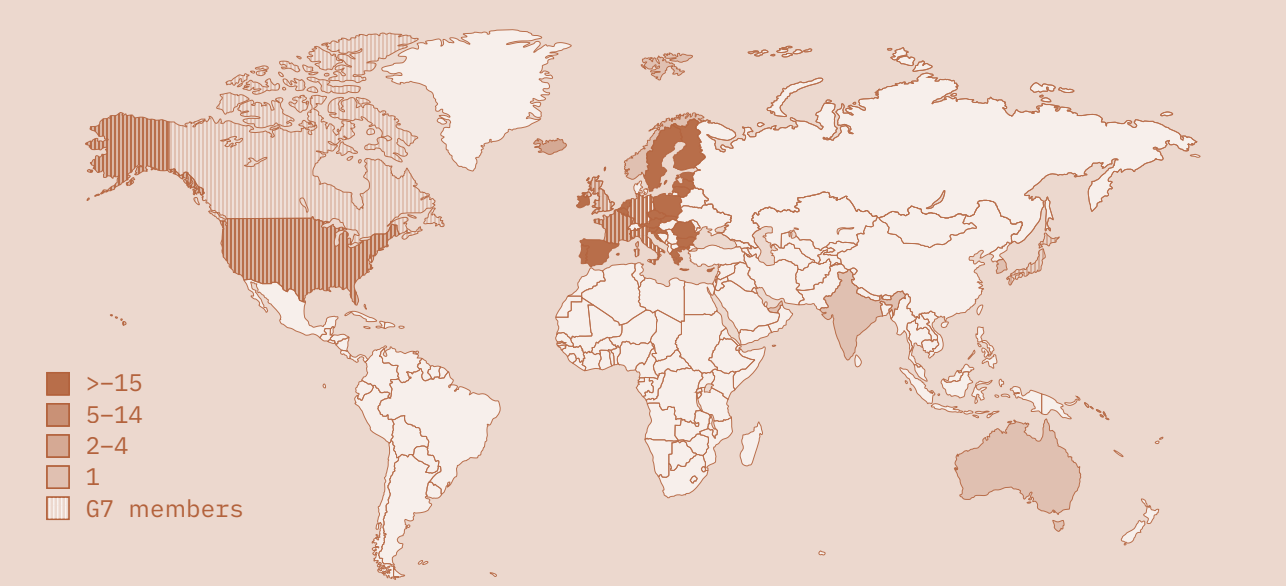
The timeline clearly shows that the frequency of publications in recent years increased enormously. A distinction is made between principles and guidelines of:

> Social organizations, like the Top 10 Principles for ethical AI' of the UNI Global Union (2017);
> Governments, like the 'AI Ethics Principles & Guidelines' of Dubai (2019);
> Intergovernmental organizations, like the 'Principles on AI' of the OECD (2019);
> Multi-stakeholders, like the 'Beijing AI Principles' of the Beijing Academy of AI (2019);
> The private sector, like the 'Everyday Ethics for AI' of IBM (2019).

Ethics is clearly no longer only a European affair. In 2019, 'even' the Chinese government launched the 'Governance Principles for a New Generation of AI'. Obviously, China realizes that, if it wants to continue to do business in AI internationally, it needs to address the subject of ethical AI applications. They do like to stay in control and find fraudulent practices as undesirable as any other government.

### Distribution of ethical guidelines

Ethical principles and guidelines come in a variety of types and sizes. Despite the fact that there are so many principles and guidelines, their distribution is limited. In 2019, researchers of ETH Zurich analysed no fewer than 84 ethical guidelines that were published worldwide in recent years. From the private sector to social organizations and governments. Research shows that most ethical guidelines come from the US (21), Europe (19) and Japan (4).




‘Geographic distribution of issuers of ethical AI guidelines by number of documents released’ – Source: ETH Zürich (2019).

The highest ‘guideline density’ is found in the United Kingdom, where no fewer than 3 ethical guidelines were published. Member states of the G7 produced the highest number of ethical guidelines. The G7 (Group of 7) consists of seven important industrial states, namely Canada, Germany, France, Italy, Japan, the United Kingdom and the United States. In 1997, the European Union also joined the G7, but the name was not adjusted to reflect that. It is no coincidence that these countries publish the most guidelines. In 2018, the ministers of the member states responsible for innovation have signed a G7-declaration involving *Human-centric AI*, in which they presented a joint vision that is designed to strike a balance between encouraging economic growth through AI innovation, increasing the confidence ↴ in and acceptance of AI and promoting inclusivity in the development and implementation of AI.

Although the overview is a snapshot (for instance, two new guidelines were published in China after the publication of the study), it does present a clear division. So far, it is above all the richer countries that dominate the worldwide discussion about AI. Although some developing countries were involved in international organizations setting up guidelines, only a few of them actually published ethical guidelines of their own. However, according to the researchers, that is very important, because different cultures have different opinions about AI. A global collaboration is needed to provide ethical AI in the future that contributes to the welfare of individuals and societies.

A first attempt to realise a global collaboration was made by the *Organisation for Economic Co-operation and Development* (OECD), a coalition of countries aimed at promoting democracy and economic development, which, in 2019, announced a set of five principles for the development and use of AI. However, because China isn’t a member of the OECD, it has not been included in the creation of the guidelines. The principles involved appear to be at odds with the way AI is used. Especially with regard to facial recognition and supervision of ethnic groups that are being associated with political dissidence. But, especially in the case of conflicting opinions, it is important to open a dialogue and try to reach a kind of consensus.

### Similarities and differences

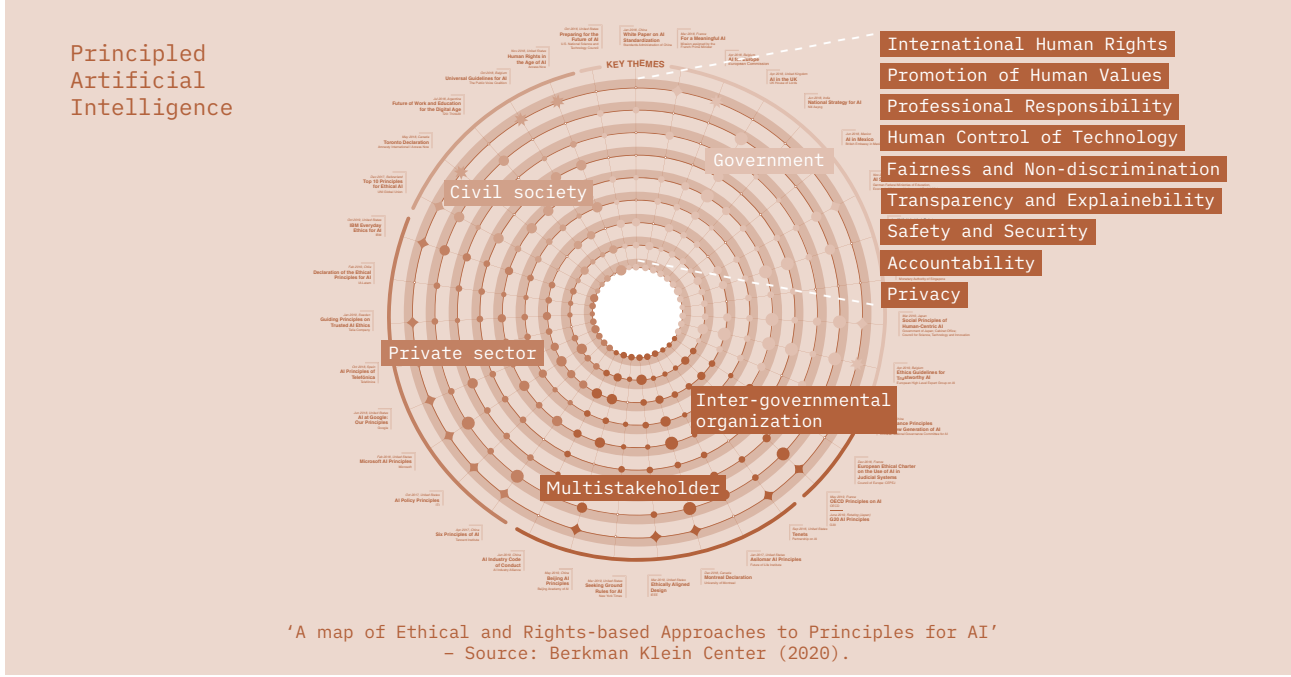
The researchers of ETH Zurich not only looked at the geographical distribution of the ethical principles and guidelines, but also at the similarities and differences between the principles. The study shows that, although no ethical principles are exactly identical, there is a clear convergence surrounding the principles of transparency ↴, justice ↴ and fairness ↴, reliability, responsibility ↴ and privacy. These principles are mentioned in more than half of all sources.

However, there are significant differences in the way ethical principles are interpreted. In particular the specific recommendations and areas of attention that were based on each principle vary enormously. For instance, according to some guidelines, AI is above all meant to make the decision-making process explainable, while other guidelines argue that it is necessary for the decisions of AI to be completely traceable as well. That is why the researchers emphasize the need to integrate the various guidelines to reach a worldwide consensus about adequate implementation strategies.

The researchers of the *Berkman Klein Center* arrived at a similar conclusion. Based on an analysis of the terms being used, they presented a list of eight overarching principles, which in broad lines match the terms of the study by ETH Zurich.

| Ethical principle | Number of documents | Ethical principle | Number of documents |
|---|---|---|---|
| Transparency | 73/84 | Beneficence | 41/84 |
| Justice & fairness | 68/84 | Freedom & autonomy | 34/84 |
| Non-maleficence | 60/84 | Trust | 28/84 |
| Responsibility | 60/84 | Sustainability | 14/84 |
| Privacy | 47/84 | Dignity | 13/84 |
| | | Solidarity | 6/84 |

'Ethical principles identified in existing AI guidelines'
– Source: ETH Zürich (2019)




'A map of Ethical and Rights-based Approaches to Principles for AI'
– Source: Berkman Klein Center (2020).

Again, it is clear that, when you look at the details and interpretations, there are clear differences between the ethical principles and guidelines. Not only in the extent to which certain principles have been worked out, but also in the extent to which they refer to international human rights, both as a general concept and in terms of specific documents, like the '*Universal Declaration of Human Rights*' or the 'United Nations Sustainable Development Goals'. Some ethical guidelines even use an explicit 'Human Rights Framework', which means that human rights are the basis for the formulation of ethical guidelines for the development of AI applications. Against expectations, it is above all the guidelines of the private sector that refer to human rights, and to a lesser extent the guidelines of governments.

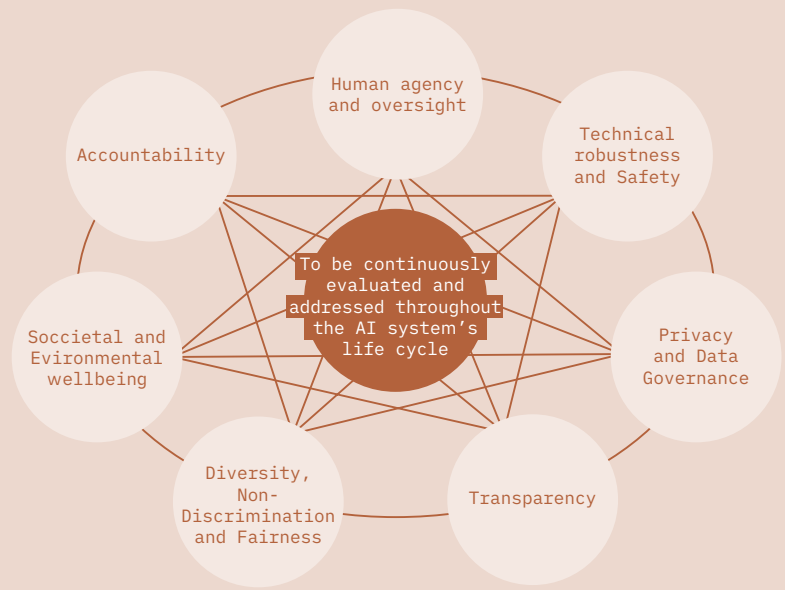
## Conflicting values

Ethical ↴ principles ↴ can help us reach an agreement about what we consider to be important and allow us to develop the AI ↴ applications to match. The question is, however, to what extent these guidelines survive in practice, because the focus is very much on the conditions and to a lesser extent on the consequences. However, in practice, dilemma's occur, which can create value conflicts. For instance, on a fundamental level, we all agree that it is wrong to kill another person. But what do you do if a terrorist threatens to kill 100 people? Should an autonomous weapon drone be allowed to intervene? In practice, there are often extenuating circumstances. Another example is that we all agree that stealing is wrong. But what if a single mother steals to feed her baby? Should a robot judge simply apply the rules and punish the mother to the full extent of the law? Ethical guidelines shouldn't just be about what we think is important, but also how important we consider different values to be in relation to each other. And in which circumstances.

> When we take 'Ethics Guidelines for Trustworthy AI' of the European Union as a starting point, we notice that such trade-offs are hardly mentioned. In fact, it is emphasized that all the requirements are equally important and that they support each other.

The report includes only one small paragraph about trade-offs, stating that trade-offs can occur and that the pros and cons have to be weighed, although what those pros and cons are and how they can be weighed remains unclear. The report only indicates that the pros and cons have to be evaluated and documented.






'Interrelationship of the seven requirements: all are of equal importance, support each other, and should be implemented and evaluated throughout the AI system's lifecycle'
– Source: High Level Expert Group on AI (2019).

### What do we find really important?

According to the guidelines of the European Union, the seven 'key requirements' are equally important. However, when you ask people to rank them, it turns out that there are differences in how much value we attach to different principles. Our own research shows that the values autonomy ('Human agency and Oversight'), privacy ↴ ('Privacy and Data Governance') and Equality ('Diversity, Non-Discrimination and Fairness ↴') have a much higher score than explainability ↴ ('Transparency') and responsibility ('Accountability ↴'). We presented respondents (n = 108) with seven principles and asked them to rank them in order of importance: 1 = a low priority and 7 = a high priority.

| Principle | Average score |
|---|---|
| People staying in control (*autonomy*) | 4,81 |
| Protecting personal data (*privacy*) | 4,80 |
| Fighting inequality (*preventing biases*) | 4,67 |
| Optimising human choices (*efficiency*) | 3,88 |
| Increasing explainability of the system (*solving black box*) | 3,77 |
| Solving accountability problem (*legislation*) | 3,34 |
| Improving international position in AI (*geopolitics*) | 2,73 |

The other questions also show that autonomy is considered to be very important. For instance, a large majority of the respondents indicate that strict legislation and regulation is needed to maintain control of AI, even if that were to slow down the development of the technology.

However, when we zoom in, we can see that there are differences, especially between the different groups of respondents. We presented the same questionnaire to three different groups, AI experts, administrators and students. The analysis shows that AI experts tend to take more risks with AI than the other groups. For instance, they are more open to AI systems making autonomous decisions and they are prepared to accept a higher margin of uncertainty of the systems. When it comes to acquiring an international advantage ↴ in the development of AI, on the other hand, it is the administrators who are willing to take greater risks, indicating that, as a country, we should do everything we can to secure an international lead in the development of AI, even if that leads to international tensions. Students are relatively speaking more concerned about their privacy.



It becomes even more interesting when the different principles are pitted against each other. In general, both privacy and equality are given a very high score. However, when we pit them against each other, most respondents prefer to stay in possession of their data instead of relinquishing control of their data to further equality. The differences are even greater in other areas. For instance, some respondents indicate that, if we do not understand how an AI system obtains its results, we shouldn't use it, while as many of the other respondents state that it doesn't matter how an AI system gets its results. They find it above all important that the system performs as expected. So people have different opinions when it comes to these types of principles, opinions that are context-dependent, making it difficult to translate general guidelines into practice. In many cases, a balance will have to be found.

*» Norms and values are universal. They are the moral judgments that vary among people. «*

-- Wiegert van Dalen, Ethicist

### What trade-offs are there?

When AI systems are applied in practice, there are various possible value conflicts. Both between values and within values. What we consider to be important depends, among other things, on the context in which the system is applied. There's a big difference whether it involves a recommendation for a movie, a diagnosis on the basis of lung X-rays in the hospital or a recommendation concerning a business takeover. Furthermore, there are also technical considerations. For example, when we go all out for transparency, that will affect the level of privacy. That doesn't mean we have to relinquish our privacy altogether.

It is not a zero sum game where we have to trade one thing for another, but choices will have to be made in the design. To be able to maximise values in relation to each other, any potential tensions first have to be identified.

*Technical trade-offs*

One important trade-off in practice is that between accuracy and explainability ↴. Methods that are currently being used in the development of AI, like deep learning ↴, are so complex that the exact decision-making processes are impossible to trace. At the moment, the optimisation of these types of systems takes place on a trial and error basis: the input is tweaked to see what it does to the output. If you want to optimise the accuracy of the system, you will have to give up part of the explainability. On the other end of the spectrum we find linear regression, which, compared to deep learning is a method that is far from flexible, but is easy to explain ↴. Sometimes people choose this method for the sake of explainability, even if they know that the relationship between the underlying variables isn't directly proportional.

*» If you have arbitrary data and you want to be able to learn from it, you pay a price for that. «*

-- Maarten Stol, BrainCreators



A similar trade-off occurs between privacy and accuracy. Generally speaking, the more complete and encompassing the data set is with which an AI system is trained, the more accurate the system will be. For instance, when AI is used to predict future purchases of consumers on the basis of their purchasing history, the model will be more accurate if the data it can use is enriched with, for instance, demographical information. However, collecting personal data can violate the privacy of the customers.

When the data set is incomplete, that can lead to skewed or discriminating results. There can be a trade-off between privacy and fairness. Organizations can take various technical measures to limit the risk of that happening, but most of those techniques will make the system less accurate. For instance, when you want to prevent a credit rating system to assign people in a certain class based on where they live or what their ethnicity is, the model should not include those data. However, although that can help prevent discriminating results, it will also lead to less accurate measurements, because a person's zip-code can also be an indicator for a legitimate factor, like job security, so it will reduce the accuracy of the results.

In turn, these considerations affect the safety of the system. If your model is less accurate, the likelihood of errors is greater, which will affect safety. If you optimize safety at the expense of explainability, that will in turn affect accountability, because that will be harder to deduce when the system's explainability is limited. As such, trade-offs can be placed on a spectrum and people need to decide where in that spectrum they feel most comfortable. There is no one size fits all in that, it has to be tailor-made.

*» An algorithm that performs exactly as intended and with perfect accuracy is not necessarily an ethical use of AI. «*

-- Kalev Leetaru, George Washington University

*Alignment trade-off*

In 540 BC, King Midas wished that everything he touched would turn to gold. That meant, however, that he also turned his food and loved ones into gold, which made him so lonely and hungry that he relinquished his superpower. When he formulated the end, he failed to consider the means. That is also known as the Value Alignment Problem (VAP). Theoretically speaking, an intelligent machine that is programmed in such a way as to produce as many paperclips as possible would do everything in its power to make that happen. In his book *Superintelligence* ↴, Nick Bostrom philosophises that the machine will clear from its path anything that comes in the way of production. Even people, because after all, they don't contribute to the production of paperclips. A machine can be so goal-oriented that the results don't match what we want.

It is important, then, to determine what a successful outcome is. The question is whether we programme on the basis of desirability or on the basis of reality. When you googled 'CEO' a number of years ago, that would yield pictures of predominantly white middle-aged men. You would have to scroll quite a bit to see a picture of a woman. However, if we look at statistics, that is not completely inaccurate. Women continue to be underrepresented in the top management positions. The data of *Pew Research Center* shows that, in 2018, the percentage of female CEO's in the *Fortune 400* was a mere 4.8%, but when you google 'CEO' in 2020, 3 of the first 20 pictures that you see are of women. Which is 15%. Still a low percentage, but a lot higher that it is in reality. And the question is who determines what a successful outcome is.



If, for instance, you want to use an algorithm ↴ that determines what you can cook on the basis of ingredients, it is very important who determines what a successful outcome is. Parents want their children to eat healthy, but the kids themselves would much prefer something that tastes good. At the moment, the decision as to what a successful outcome is still lies in the hands of a very small group of people.

*Contextual trade-offs*

What we find important to a large extent depends on the context. Take, for instance, privacy. If a doctor asks you to take off your pants to examine your private parts, that is usually not a problem. But if your baker asks you the same question, that's a violation of your privacy. The same applies to explainability, the need for which will be lower when talking about a chatbot compared to a self-driving car. As such, the importance depends, among other things, on the risks involved. The acceptable level of subjectivity also depends on the context. For example, a 'faulty' recommendation by Netflix won't do much damage, but when a medical diagnosis is wrong, that may have far greater consequences. Although the subjectivity of Netflix's recommendation system is much higher than the image recognition software of a hospital, the latter has to meet much higher standards.

The question what we find important also depends on the perspective, which is often culturally determined. In many cases, the data for image recognition ↴ software is still labelled by people. In some cultures, an image of a man or woman with a glass of beer is labelled with having a good time, being together, partying, etc., while in other cultures, it is tagged with alcoholism, rowdiness, etc. The perspective also depends very much on age. For instance, many people consider it inhuman to have robots take care of elderly people. The National Future Monitor of 2019 shows that most Dutch people have a negative view about having intimate relations with a robot. But many of the elderly in need of help think it's

an ideal outcome. For instance elderly people who are no longer able to eat independently. They feel ashamed when they are being fed by strangers in a home. Only their kids are allowed to do that and otherwise they prefer not to eat. Robots fill a need and actually provide autonomy. What we consider to be ‘well-being’ is often a subjective affair. Emotionally speaking, we want to prolong life as much as possible, especially when it concerns the people in our own environment. The question is, however, what the value of life is for terminally ill people themselves. What are the optimisation goals for AI in such situations? As efficiently, socially, sustainably or humane as possible? And does humane mean prolonging life or reducing unnecessary suffering?

We are suddenly faced with similar issues by the corona crisis in 2020, where it becomes clear that different interests are interwoven and have a mutual influence. For instance, for purely health-related reasons, a complete lockdown could be sensible. However, for people in developing countries – some of whom are dependent on day labour – that would also mean a complete loss of income and possible starvation. What is worse? Dying because of the virus or starving to death? A crisis like this one underlines the need for posing fundamental questions. It turns out to be difficult to express the value of life in measurable values, although this is done during a pandemic. The entire economy grinds to a halt to protect the vulnerable. How far should you go in that? These are not popular questions, but they cannot be avoided. Especially when AI-systems will play a greater role in decision-making processes, we need to determine where the balance is between rationality and emotion, objectivity and subjectivity, the long term and the short term. Are we willing to accept ‘objective’ decisions by AI systems or should we also build in emotions ↴ and moral intuition?



### What trade-offs are we willing to make?

In the Netherlands, more and more cameras and trackers are used to follow and record movements. When it is about improving safety and liveability, people appear to be willing to accept the use of sensors and the collection of sensor data. But there are specific conditions. Research by the Rathenau Institute from 2019 indicates that the acceptance depends predominantly on the context. People are not *a priori* against the use of bodycams or Wi-Fi trackers, but it depends on when and in what situation they are deployed. These insights are supported by research by the European Commission from 2020, which shows that 59% of all respondents are willing to securely share part of their personal information to improve public services. Especially when it concerns the improvement of medical research and care (42%), the improvement of the response to a crisis (31%) or the improvement of public transport or reducing air pollution (26%).

Analysis by the Rathenau Institute show that there are two crucial factors, namely the level of safety people experience and the type of living environment in which sensor technology is applied. In situations in which citizens feel unsafe, they will accept the use of sensors more easily than in situations where they feel safe. The use of sensors is considered less acceptable in private spaces that in public spaces where there are many people. As such, the use of sensors is desirable to improve safety and liveability in crowded public spaces, but not in private spaces. It is interesting to note that people not only weigh safety and privacy, but several other values as well, like democratic rights, transparency, efficiency and human contact.

Therefore, ethical principles and guidelines cannot simply be copied and pasted into practice. Each specific situation requires specific considerations. It is always necessary to determine in the context which values conflict and what is acceptable in this. A trade-off that is accepted in certain situations may be completely unacceptable in other situations. Together we will have to determine what we can and cannot accept in different situations.

*» Sometimes we think that technology will inevitably erode privacy, but ultimately humans, not technology, make that choice. «*

-- Hu Yong, Peking University









2.3 ----------

## Practical challenges

The formulation of ethical ↴ principles and guidelines is an important first step in the realisation of ethically responsible AI applications ↴. However, it is not easy to translate those guidelines into practice. For instance, when talking about transparency ↴, its meaning depends, among other things, on the domain and environment in which the AI application is used. For example, when Spotify recommends a song I don't like, I don't need an explanation as to why that happened, but if an algorithm ↴ causes me to be rejected during a job application process, I would like to know the criteria on the basis of which I have been rejected. In practice, there are a number of potential value conflicts, for instance between transparency and privacy ↴. How do we safeguard our privacy when we demand transparency? In many cases, it is not exactly clear which ethical questions play a role in the development of AI applications.

Fortunately, there are more and more organisations, at a national as well as international level, that have developed various tools that can help identify ethical dilemmas in practice and that can be used to map the ethical implications of applying AI in practice and the ethical issues that are involved. Think, for instance of:

> The Algorithmic Impact Assessment (AIA) of the *AI Now Institute*
> Data Ethics Decision Aid (DEDA) of the Utrecht Data School
> Artificial Intelligence Impact Assessment (AIIA) of the ECP

Based on different questions, organisations are helped to get a clearer view of which ethical issues play a role in their AI projects and how they want to handle them. Examples are questions like 'Are personal data being used in the project?', 'Are all the various groups of citizens represented in the data set(s)?' and 'Who are we missing or aren't yet visible?'. That can help expose and prevent potential biases ↴ in the application. It also helps organisations to document their considerations, making the process more transparent and allowing them to be accountable to their stakeholders. However, such guidelines are often hard to express unambiguously and difficult to quantify.

> *» There's no such thing as a single set of ethical principles that can be rationally justified in a way that every rational being will agree to. «*
>
> -- Tom Chatfield, Tech philosopher

## The clarity of guidelines

It is obvious that we all want to prevent the use of AI to treat people unfairly and to discriminate them on the basis of their gender or ethnicity, which is why fairness is a commonly used principle in ethical guidelines and assessment tools. However, it is not easy to determine what 'fair' exactly means. It is an issue that has kept philosophers busy for hundreds of years. Is a society in which everyone is treated exactly the same fair? The arrival of AI gives this issue a new dimension, because the concept of fairness has to be expressed in mathematical terms ↴. Think, for example, of the use of AI in the legal system. The use of predictive policing can help predict criminal behaviour through large-scale data monitoring and data analyses. However, there is always a risk that people who do not meet the criteria being used get a positive score (false positives) and that people who do meet those criteria get a negative score (false negatives).



> What does 'fair' mean in these cases? Do you risk locking innocent people up or do you risk having them commit crimes?

In America, software has been used on a large scale to predict the likelihood of recidivism. Research by *ProPublica* from 2016 shows that the software is prejudiced against people with a darker complexion. It is obvious that that is unfair, but it is not easy to determine what is fair and how you can measure it. Is fairness defined by using the same variables or by using the same statistics after using different variables? Does fairness mean that the same percentage of black and white individuals are given high risk assessment scores? Or that the same risk level should result in the same score, regardless of race? So the question is whether fairness is about treating people equally (with the risk of an uneven result) or about getting equal results (with a possible unfair treatment). Research shows that different mathematical definitions of fairness are mutually exclusive (Selbst et al., 2019). So it is impossible to meet both definitions at the same time, which means that, at some point, a choice has to be made.

> However, it is not possible to make a universal choice. How we define fairness depends on the application domain. You cannot simply transpose a system that is used to produce fair legal verdicts onto an application process. However, people often think that a powerful system can be applied in multiple domains. Different cultures and communities have different ideas about fairness. Not only do different standards apply, there are also different laws. In addition, our opinions about what is right or wrong can change over time. That makes it difficult, perhaps even undesirable, to determine in advance how an AI system should act.

» *By fixing the answer, you're solving a problem that looks very different than how society tends to think about these issues.* «

-- Andrew Selbst, Data & Society Research Institute

## The measurability of the guidelines

The various ethical principles and guidelines have different levels of abstraction, which end values and instrumental values getting mixed up. For instance, *Societal well-being* and *Safety* are end values that we, as a society, aim for, while *Accountability* ↴ and *Transparency* are instrumental values that we can use in our pursuit of those end values. Some guidelines are easier to quantify than others. For instance, the level of accuracy can be made measurable, but in the case of transparency, it's more complicated. When is something 'transparent enough'? At 70% transparency? And what exactly does that mean?

The question is whether we should even aim for 100% transparency. Research by Microsoft Research from 2018 shows that too much transparency can lead to information overload. It turns out that it is even harder to detect and correct the errors in transparent models. In addition, there is a risk that people may trust transparent models when they shouldn't. A follow-up study by Microsoft Research from 2020, in collaboration with the University of Michigan, shows that the use of visualisations about the training results of machine learning ↴ tools create a misplaced trust about the possible applications of the models. Even when the data had been manipulated and the explanation didn't match reality.



## The motives behind the guidelines

In 2019, Google introduced the *Advanced Technology External Advisory Council* (AETAC), an external ethical board designed to make sure that the company would adhere to its own guidelines for ethically responsible AI applications. However, the ethical board was disbanded after a week. Immediately after announcing who was on the board, there were intense discussions, in particular about the position of Kay Coles James. The president of the Heritage Foundation is known for her conservative opinions, among other things about the rights of the LHBTI community.

> Even apart from that, one has to wonder why Google decided to create such an advisory board in the first place. According to experts, the ethical guidelines and advisory boards of Google and other commercial organisations are aimed at circumventing government regulations. This is also known as 'ethical washing'. It is said to be a way to deflect criticism, not to act in a genuinely ethical way. Because the advisory boards have no real power, the organisations don't actually have to adjust their behaviour. And it seems hardly surprising that Google created its advisory board after a period when it had been under considerable pressure. At the time, Google worked together with the Chinese government on *Project Dragonfly*, a search engine that blocked results that the Chinese authorities considered undesirable. According to Amnesty International, the modified search engine threatened the freedom of speech and privacy of millions of China's citizens. Later, the employees of Google also started a protest and wrote an open letter to Google's management. In 2018, Google announced it would stop working on Project Dragonfly, but employees doubt that that actually happened.
> The advisory board appears to be above all a way for Google to tell the world: 'Look, we are doing everything we can'.

It doesn't appear to be a coincidence that other tech organisations also launched ethical guidelines and advisory boards in a period in which numerous problems in the tech sector came to light, like the Cambridge Analytica scandal in 2018. For instance, Microsoft created its *AI ethics committee* and conducted broad research into the transparency of AI systems, while Amazon sponsors a research programme aimed at promoting 'fairness in artificial intelligence' and Facebook has invested in an 'AI ethics research center' in Germany.

*» Ethics boards and charters aren't changing how companies operate. «*

-- James Vincent, The Verge

The challenge is to arrive at binding guidelines. According to different experts, legislation is necessary to make sure that ethical guidelines are observed. The first step in this direction is described in the 'Whitepaper on AI', which was presented by the European Commission in 2020 and in which the commission explains proposals to promote the development of AI in Europe, taking European fundamental rights into account. An important part is the proposal to develop a 'prior conformity assessment' for risky AI applications, based on the ethical guidelines of the *High Level Expert Group*. That legal framework is designed to tackle the risks facing fundamental rights and safety, allowing reliable AI systems to get a quality mark, making it clear to users which systems they can trust.

Although it is very important to make sure that ethical guidelines are applied in practice and respect European laws and fundamental rights, it provides insufficient tools for actually integrating the guidelines in the development process. The currently available checklists and assessment tools are insufficiently quantifiable. Now, every aspect of the list can be 'checked' without completely meeting them. Ethical guidelines and assessments are therefore mainly tools for evaluating whether the AI application complies or will comply with the principles, not how the requirements must be integrated into the design itself (and thus withstand the evaluation). While many guidelines and assessments provide different checklists and questionnaires, they do not answer how AI systems can make ethically responsible decisions.

*» Despite an apparent agreement that AI should be 'ethical', there is debate about both what constitutes 'ethical AI' and which ethical requirements, technical standards and best practices are needed for its realization. «*

-- Effy Vayena, ETH Zurich

Ethics boards and charters aren’t changing how companies operate.

-- James Vincent, The Verge

# Compromises surrounding reliability of AI

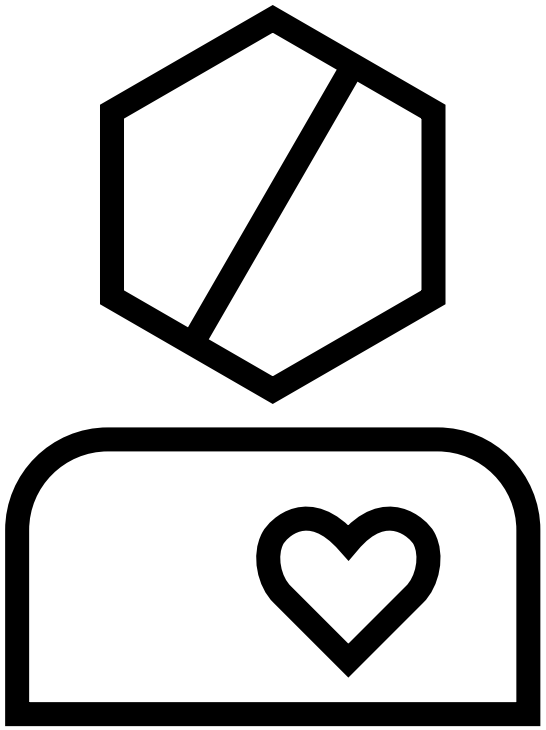
*By Maarten Stol, Principal Scientific Advisor, BrainCreators*

In 2019, the European Commission published the '*Ethics Guidelines for Trustworthy AI*'. It contains a list of seven requirements an AI system has to meet to be allowed to be called 'reliable'. In parentheses, because, even though we all have a sense about what reliability is, in the case of advanced technical systems, the definition is hard to make exact. Such lists have been made before (think for example of the *Asilomar AI Principles*), but the European try to cover the term 'reliable' as completely as possible and at the same time keep the list from becoming confusing:

1. *Human agency and oversight.*
2. *Technical robustness and safety.*
3. *Privacy and data governance.*
4. *Transparency.*
5. *Diversity, non-discrimination and fairness.*
6. *Societal and environmental well-being.*
7. *Accountability.*

However, the guidelines also make it clear that there can be fundamental tensions between some of these requirements. And that's where the complexity of the issue lies. Leaving aside the question how to meet these requirements on an individual level (easier said than done, technically), I would like to address the possible interactions and trade-offs, starting with Safety vs. Accountability. The behaviour of current AI systems is largely determined by data, more than by programming code alone. The goal of machine learning is to try to use that data and automatically teach a programme that can make a decision in the general trend of the input data. However, since machine learning is statistical in nature, it will never be 100% correct, in other words: not 100% safe. To get as close as possible to 100%, increasingly complex systems are being developed. Due to the nature of the technology, the result of that increasing complexity is that such systems take on the character of a black box. That is to say that, even though the decisions may be of a good quality, the origin of each individual decision is less accessible. So explainability is sacrificed in favour of safety. Safety vs. Accountability is a trade-off that can never be avoided completely with existing technologies.

There are other similar trade-offs. Privacy vs. Transparency: how much privacy must citizens give up to unlock data that must make sure that AI systems are transparent to other citizens? Human agency vs. Societal well-being: which interests are more important, those of the individual or those of the collective? Technical robustness vs. Environmental well-being: how much energy is the continuous training and maintenance of AI systems allowed to use? Etcetera. It would therefore appear to be reasonable to assume that, in the future, we will not be able to put all the guidelines into practice. Instead, we shall need to compromise. Both industry and government have a role to play in that. Industry will have to assume a measure of responsibility to try and follow the guidelines. And government will have to think about legislation designed to manage these developments. One thing is certain: when it comes to the ethical use of AI, the future will not be without compromises.



# 3 Ethics in AI design



Pages: 40
Words: 7770
Reading time: approx. 1 hour

## Ethics in AI design

From governments and social organizations to scientific institutes and the private sector, for different reasons, they all think it is important to develop ethically ↴ responsible AI ↴ applications, which is why ethical principles and guidelines ↴ pop up like mushrooms. Different guidelines dictate, among other things, that AI systems have to be secure and robust, safeguard our privacy ↴ and treat us fairly ↴. The question is to what extent such noble goals affect each other in practice. Can we develop AI systems that are both accurate ↴ and explainable ↴? And are explainable systems by definition also fair? Can we translate fairness into mathematical terms ↴ without treating people unfairly? And do we agree in the first place about what a fair society should look like?

> Ethical guidelines – and in particular their interpretation – are to a large extent context-dependent and subjective. In that sense, ethical guidelines are more about morality than about ethics. Ethics should be about aligning the different applications to that morality. From that perspective, ethics are more a design issue than a collection of opinions about what we find important. If we want to use ethically responsible AI applications in the future, we will have to concern ourselves now with the question as to how we can develop applications that are in agreement with morality. It is not enough only to assess whether or not an AI application meets the guidelines. Ethical principles and guidelines actually have to be integrated into the design. And in the design process as well. That requires a different approach to ethics in the development of AI.





## A new perspective on ethics

In the past, ethics ↴ was mostly about human actions. However, with the arrival of AI ↴, there is a new player in the game, namely self-learning technology. As technology becomes more and more autonomous, takes over more decisions and makes it harder to trace the decision-making rules, new ethical questions emerge. The question is to what extent the current ethical terminology is still sufficient. What is the relationship between people and technology? And how will that relationship develop?

### A broadening of the concept of ethics

According to Peter-Paul Verbeek, professor of Philosophy of Man and Technology at Twente University, the terminology of ethics needs to be expanded. To be ethical, according to existing opinions, it is necessary to have intentions and to be able to act freely. According to Verbeek (2011), technologies that help shape moral decisions have no intentionality ↴ and people who are guided by technology in their moral decisions are not free. Opinions are seriously divided on whether or not technology possesses intentionality, but it is clear that we are no longer completely autonomous and that our lives are affected by technology.

As such, the idea of a completely autonomous person that is behind existing ethics is often erroneous. The question is not *whether* we want completely explainable systems (that ship has long sailed), but *how* we can apply systems that are no longer completely explainable. So we should focus more on mapping the impact of these systems and come up with a constructive plan on how we can and want to deal with them. Rather than focusing exclusively on the question what is right and wrong, we need to develop a concrete ethical framework that allows us to deal with the errors of AI systems. We cannot approach that from a purely theoretical perspective, but have to experience

it in practice. We need to experiment in controlled settings. From *human in the loop* (HITL) to *human on the loop* (HOTL). AI actually allows us to have more control over inequality with regard to things like gender or race. AI is far better equipped to manage the conditions that create biases ↴ than people are. That way, we can create enough diversity in the data sets and in the background of the programmers developing the algorithms ↴ and the supervising legislators. We need to try and create systems that divide unfairness more evenly. AI may not be able to solve our mistakes entirely, it can distribute them more fairly.

*» The worst form of inequality is trying to make unequal things equal. «*

-- Aristotle

### Ethics as leitmotiv

Many ethical discussions focus on the question when the use of AI is and is not acceptable, in an attempt to create hard boundaries and rules to 'control' the technology. However, that would suggest that it is possible to separate society and technology from one another completely. But everyday practice is not that black and white. People and technology affect each other mutually; we shape technology and technology shapes us. In addition to broadening the concept, it also requires a different approach to ethics.

### *Guidance ethics*

In 2019, ECP, the Platform for the Information Society, in collaboration with Peter-Paul Verbeek, published a report about 'guidance ethics'. Instead of asking how we can assess AI, according to their approach, we should focus on how best to guide the implementation of AI in our society and how we can deal with it in a responsible way. In this approach, the focus is on the development of technology with action perspective.



*» Instead of seeing ethics as 'judging', it could also be seen as a normative 'guidance' of technology in society. «*

-- Peter-Paul Verbeek and Daniël Tijink, ECP

A clear distinction is made between three kinds of activities. With regard to the first kind (ethics by design), the focus is primarily on the design of the technology itself. Values like privacy ↴ have to be included in the design. For instance, it has to be possible to use AI to make medical diagnoses without putting the privacy of the patients at risk. AI systems could be trained, for example, with data from different hospitals, without the data in question ever leaving a hospital building or touching the servers of a technology company. With regard to the section kind of activity (ethics in context), the focus is on context-specific agreements. The introduction of a new technology is often accompanied by changes in the environment, which are not always visible, however, for instance the use of facial recognition software. That's why it has to be clear to everybody that AI is being used, what information is collected, who has access to the data, what possible ways there are to contest the decisions of an AI system, etc. In the case of the last activity (ethics by user), the focus is on the use of the technology. It is important that everybody has enough knowledge to deal with the technology in a critical and responsible way, both the developers and the users. There could still be driver's licences for people in self-driving cars, for example, to reflect the new skills that have to be learned, for instance for the communication with the autonomous systems or when there is a need to intervene when the system requests it.

This leads to a more holistic approach to ethics. Both the use and the users, and the technology are included in this approach, which does justice to the dynamic practice of AI: the adjustment between people and technology is an ongoing process. It is no longer either people or the technology making the decisions. They mutually shape one another. The limitation of this approach is, however, that it is suggested that technology will keep developing and that we might as well resign ourselves to that. But the question where we should use AI in certain contexts continues to be relevant. We mustn't see AI as a goal in itself, but as a means to a certain other goal. We cannot skip the question whether or not AI is the best way to achieve that goal. So we must keep asking ourselves what kind of society we want to be, given all the technological developments.

*» AI is an Ideology, Not a Technology. «*

-- Jaron Lanier & Glen Weyl, Wired

*Beckoning perspective*

When we talk about the future of AI, it is often suggested that ethics slow down innovation and that, for instance, Europe has to make a choice: either impose strict regulation and slow down innovation or innovate and try to join the US and China. According to a report by PwC Netherlands from 2020, there is a trade-off between strict regulation and quick innovation. The 'White Paper on AI', in which the European Commission voices its ambition to speed up the development of AI and at the same time announces stricter regulation, could be a case of putting one foot on the brake and the other on the accelerator.





*» Those who introduce strict regulations, accept that the development of AI will initially be slower. «*

-- Mona de Boer, PwC Netherlands

There is a middle way, however. Think, for instance, of agriculture, which has become intensely innovative because of government regulations. They have become so innovative that our greenhouses are considered to be the most sustainable in the world, which is why Dutch greenhouse builders are increasingly active abroad as well and contribute to the realisation of sustainable horticulture projects around the world. So ethics doesn't have to be an anchor that you have to drag around, it can provide a beckoning perspective for innovation.

Uniting values

Trade-offs are still often presented as forced binary choices. Do you want the government to know everything about you, or live in unsafe conditions? The same thing occurred with the use of tracking apps during the corona crisis in 2020. Newspaper headlines and articles were focused on the trade-off between privacy and public health. The emphasis appears to be on the individual versus the collective and people are expected to surrender their privacy for the greater good. However, the notion that these apps only work if people give up their privacy is mistaken. There's a reason that the Dutch government decided against using any of the seven apps proposed during the 'appathon' in 2020, because they violated privacy guidelines. Apple and Google also announced that apps that use the location data of users would not be given access to their operating systems.

-- Jaron Lanier & Glen Weyl, Wired

# AI is an Ideology

not a
Technology

» *Those who would give up essential Liberty, to purchase a little temporary Safety, deserve neither Liberty nor Safety.* «

-- Benjamin Franklin

*Value-Sensitive Design*

It is possible, then, to unite values in the design, which is the starting point for Value-Sensitive Design (VSD). ‘Designing for values’ provides an alternative approach to innovation. According to Jeroen van den Hoven, Professor of Ethics and Technology at Delft University of Technology, we need to use innovation to serve values and remove value conflicts. That way, we can innovate with AI in a responsible manner. You only talk about trade-offs if you actually experience them in practice, so the challenge is to avoid those situations through design, for which we have to create environments in which we do not have to choose between different values, but in which we can maximise values in relation to each other. Yes, there are choices, but by choosing the right design, you can ensure that the choices don’t do any damage.

» *Ethics to a large extent is a design discipline and has to do with shaping our society and living environment in a responsible way.* «

-- Jeroen van den Hoven, Delft University of Technology







## Ethics by Design

When we want to use ethically ↴ responsible AI systems ↴ in the future, we need more than just guidelines ↴ and assessments. We need tools that will allow us to actually integrate such principles into the design, which means we have to move from evaluating to integrating. In short, we need Ethics by Design. Despite the fact that more and more ethicists share that view, they rarely go beyond stating *that* we need to integrate values into AI applications, which leaves the question as to *how* that is supposed to happen in practice. Examples are mentioned of AI applications that have come about in an ethically responsible way, but that says little about the AI system itself. Think, for example, of the Fairphone: a smartphone that is both fair to the environment and to people in the production process. Although this phone shows that it is possible to unite values, it does not answer the question whether or not its operating system is able to make ethically responsible decisions.

The same goes for the existing impact and assessment tools that are used to assess AI. They focus more on the development of the AI system than on its actions. Most assessment tools are checklists that focus mostly on the use of the datasets. Has the data been anonymised? And is the process transparent? They often don’t answer the question how an AI system can arrive at ethically responsible decisions. To develop genuinely ethical AI, we need to look at the different ways an AI system can learn what is and isn’t ethically responsible. So the question is how we can build systems that are able to act in an ethically responsible way in different situations. Can you programme ethical rules into the system? Do we have to equip the system with ethical target functions? Or is the system itself able to make moral judgments? To answer these questions, we distinguish three different system approaches, namely static learning, adaptive learning and intuitive learning.

Static learning

In the case of static learning, ethical principles and rules are programmed into the intelligent system, which implicitly makes the goal of the AI system part of the algorithm ↴, to be filled in by a programmer. If we want an autonomous vehicle to bring us from A to B as quickly as possible, the exceptions also have to be embedded in the algorithm. We don't want the vehicle to violate traffic rules and just drive in a straight line at 200 miles an hour. Objectively speaking, as quickly as possible literally means as quickly as possible. The algorithm also has to consider values like safety. This approach also appears to be the one that is used by ethicists and developers within the Value Sensitive Design (VSD) community. The starting point is that values like safety have to be made explicit as early as possible in the design process. The values can then be formalised and embedded in the AI system.

*» Ethics has to be part of the design of technology. «*

-- Jeroen van den Hoven, Delft University of Technology

*Advantages*

The major advantage of this approach is that ethical principles are fairly transparent and relatively easy to interpret by people, which allows us to think together about what we, as a society, consider to be important and embed that in AI systems. That provides a certain level of human control. We can monitor the development of AI even before it being marketed and give quality marks to the applications that meet the relevant ethical guidelines. In this context, it is relatively clear when certain guidelines are being violated, making it possible to hold organisations that violate the rules responsible, set up supervisory bodies and monitor the development of AI.



*» It is not ethicists, but engineers who are at the frontline of ethics. «*

-- Peter-Paul Verbeek, Twente University

*Disadvantages*

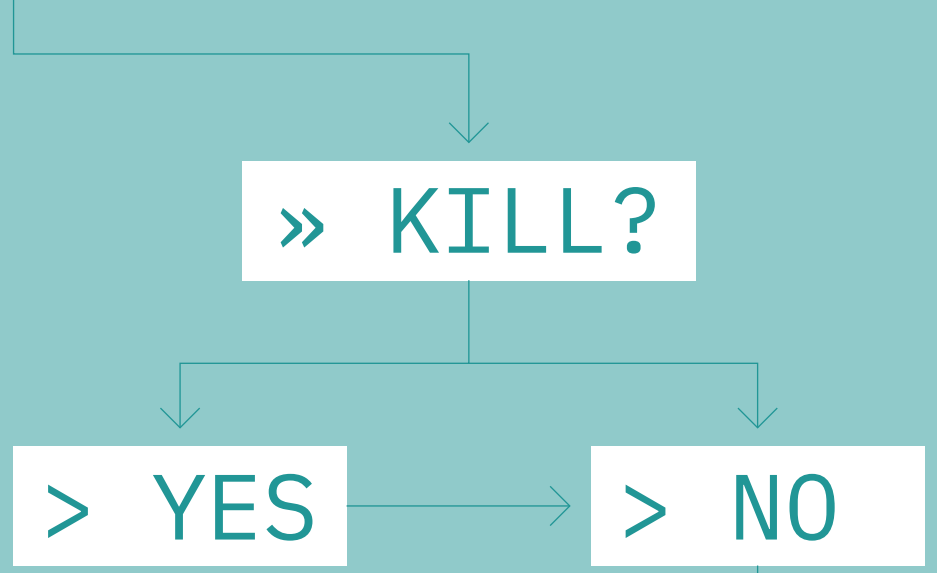
However, this approach fails to take exceptions into account that can occur in practice. It requires that there be rules for *every* possible situation, which in practice is virtually impossible to realise. In addition, there are situations in which contradictory rules apply. It is, for instance, not allowed to run a red light, but when an autonomous vehicle has to avoid hitting a group of people, it is allowed to run a red light. It is almost impossible to record all the exceptions. In addition, it is impossible to predict all the possible consequences. When GPS functionalities were developed for the aerospace sector, nobody could predict that that functionality would ultimately end up as an app on our smartphones. With positive and negative consequences. In addition, AI systems are updated: is it necessary to apply for a new quality mark with each update? It's virtually impossible to capture all that in ethical guidelines in advance.

*» You can tell a security robot not to hurt people. But that will be a limitation when that robot has to prevent a terrorist attack. «*

-- Leon Kester, TNO

*You can tell a security robot not to hurt people.*

-- Leon Kester, TNO




*But that will be a limitation when that robot has to prevent a terrorist attack.*

This approach also places too much responsibility on the programmer, because the guidelines don't specify how the values should be formalised in mathematical terms ↴. The question as to what exactly is *fair* ↴ depends on the context and the specific user application, and there is always a risk that, for instance in the case of a medical test, the result of patients is erroneously classified as positive or negative, which could lead to healthy people being administered medication and sick people not being treated. It is unfair as well as irresponsible to leave such configurations only up to the programmer.

### Adaptive learning

In the case of adaptive learning, the rules are not pre-programmed, but the system instead learns what is 'right and wrong' based on human behaviour. To that end, the algorithm is equipped with a goal function, with which it can be specified on what the algorithm should be optimised. A clear distinction is made between the problem-solving ability of the intelligent system and the goal function. That way, specific application goals can be combined with ethical goals, allowing, for instance, an autonomous vehicle to bring us from A to B as quickly as possible (application goal) as well as take our safety into account (ethical goal). When there are contradictory rules in practice, the AI system has to be able to make a judgment. This approach is popular with, among other things, Open AI and the Future of Life Institute. According to AI pioneer Stuart Russell, AI systems can only make those kinds of judgment when the systems learn in practice what human values mean. In his TED-Talk in 2017, he discussed three pillars to be able to develop safer AI applications.



1. The only goal of the AI system is maximise the realisation of human values;
2. It is initially unclear to the AI system exactly what those values are;
3. Human behaviour provides the AI system with information about human values.

In other words, learning on the job. It is important that machines learn everything about human values, to ascertain what is really important to us.

#### *Advantages*

The advantage of this approach is that the AI system learns what the right action is within the context, allowing it to handle conflicting values, which is virtually impossible when the rules are pre-programmed. It makes AI systems much more flexible and easier to use. In addition, this approach makes it possible for the system to learn from human behaviour, without copying undesirable qualities, because the system doesn't just learn from individual people (who all do something 'bad' on occasion), but from society as a whole (placing 'bad' behaviour in a broader context). The system can learn, for instance, that people sometimes steal things when they don't have enough money to send their kids to school. Rather than learning that stealing is allowed in such a situation, it will try to help find a way to send the kids to school. Systems are not 'burdened' with human urges and emotions, like status and power, which are the result of biological evolution.

*» The robot does not have any objective of its own. It's purely altruistic. «*

-- Stuart Russell, University of Berkeley

*Disadvantages*

The challenge of this approach is that an AI system has to act according to the human values of society as a whole, not just those of the user. If a machine puts your interests first, that can be at the expense of others. In one way or another, the system has to weigh the preferences of many different people. In his Talk, Russell lists a number of examples where this approach can go wrong. Imagine you have forgotten your wife's birthday and you have a meeting you cannot cancel. An AI system can help you by delaying the flight of the person you are meeting, allowing you to take your wife out to dinner. But that would upset the lives of other people. It can also happen the other way around. Imagine you are hungry and you ask your 'robot chef' to make you a ham sandwich, it can refuse your request, because there are people elsewhere on the planet who are more hungry. According to Russell, it's also possible that different human values have to be weighed against each other. Imagine that your robot chef decides to make you that ham sandwich, but there's no meat in the fridge. There is, however, a cat in the house. Which value is more important: your need for food or the sentimental value of a pet? According to Maslow's pyramid, the cat loses. Because we don't indicate in advance what is good or bad, we relinquish a large part of the control over the system. When an AI application causes unintended consequences, it is very hard to intervene.

*» We had better be quite sure that the purpose we put into the machine is the purpose which we really desire. «*

-- Norbert Wiener, 1960



Another challenge is that many of our moral opinions are implicit ↴. We don't express them literally, which makes it hard for an AI system to learn. Also, it is hard for a computer to assess emotions ↴ correctly. When people laugh, it is hard to determine if they are sincere, or whether there is another underlying emotion or motivation.

*» Computers can't tell if you're happy when you smile. «*

-- Angela Chen, MIT Technology Review

Intuitive learning

In the case of intuitive learning, elements of static and adaptive learning are combined. In this approach, the algorithm is also given a goal function, but it does not learn from the behaviour of people. Instead, people determine how much value they attach to a goal by assigning weight factors. The weight is determined based on the usefulness the goal has for society, which is why it is also called a 'utility function'. This allows the system to make reasoned assessments on the basis of the pre-weighted factors. When an autonomous vehicle has to move us from A to B, there are various goals that are relevant, like travel time, comfort, safety and sustainability. Different weights are assigned to these different goals. The autonomous vehicle will use the goal function to decide which route to take and which driving behaviour best matches both the wishes of the passengers (comfort and time of arrival) and of society (safety and the environment). Depending on the weight and the current state of the surroundings (like the amount of traffic on the road), different outcomes are possible.

Computers can't tell if you're happy when you smile.

OK :)

-- Angela Chen, MIT Technology Review

This form of learning looks a lot like the human decision-making ↴ process: it is intuitive. It is possible for humans to drive the car on the basis of laws and rules, because they translate them to the specific context. That last step cannot be programmed. For instance, the speed limit was lowered to 100 km in many places in the Netherlands, the idea being that it is better for the environment and traffic safety. However, although people are allowed to drive 100 km per hour, that's not what they do all the time. The speed is constantly adjusted to the surroundings.

*Advantages*

The advantage of this approach is that the added value of static and adaptive learning is combined: it provides the control of the static approach and the flexibility of the adaptive approach. That way, an AI system is provided with the values that society considers important, which it can integrate in a recommendation or decision, utilizing its computing power to calculate the best possible outcome in every situation, safeguarding ethically responsible outcomes without completely relinquishing control. After all, it's still people who assign weights to the factors that are included in the calculation. This also bypasses the sharp trade-off between good *or* bad. In reality, situations occur all the time where we have to choose between two 'evils', and which evil prevails depends on the context. With this approach, the system can determine which outcome is best for the individual and for society.

*» That will allow an autonomous vehicle to choose between two alternatives that are undesirable in principle. «*

-- Leon Kester TNO





*Disadvantages*

The disadvantage of this approach is that it assumes that algorithms are capable of making nuanced considerations. According to Peter Eckersley, research director of The Partnership on AI, algorithms are designed to pursue a single mathematical objective, like minimising costs or maximising the number of apprehended fraudulent people. When an attempt is made to pursue more than one goal at the same time – with some of which competing with each other – the development of AI is faced with practical and conceptual problems, in what is also known as the 'impossibility theorem'. In particular when immaterial values like freedom and wellbeing have to be maximised, Eckersley argues that, in some cases, there simply is no mathematical solution. It turns out that ethics is about more than calculating costs and benefits. It also involves less tangible things, like empathy ↴, compassion and respect. In a by now 'infamous' article, Eckersley describes that it is impossible to formally specify what a good result is for a society without violating human ethical intuitions.

*» Such systems should not use objective functions in the strict mathematical sense. «*

-- Peter Eckersley, The Partnership on AI

Despite the criticism, intuitive learning seems to be the only way to deal with the complexity of ethical issues, as TNO researcher Leon Kester and PhD candidate Nadisha-Marie Aliman also argue in their research. We need a system that is able to reason with uncertainty and that has a notion of 'self' that will allow it to deal with 'trolley-like' problems. We need to include the value ↴ that society places on the action, the consequences ↴ of the action and the person ↴ (or object) performing the action, as is also evident from research by ethicist Bart Wernaart. This is why 'ethics' shouldn't be literally part of the design, but the system itself ought to be able to make ethical considerations. And, instead of speaking of 'Ethics by Design', we should call it 'Designing for Ethics'. At the moment, technology isn't sufficiently advanced yet, but people are working hard on new perspectives, for instance in the area of hybrid AI, which involves systems that can see (using neural networks ↴, among other things) *and* that are able to reason (using formal logic ↴, among other things). This approach is also known as *Deep Reasoning* ↴, a combination of *deep learning* and *symbolic reasoning*. It increases AI system's ability to learn intuitively.

> *» Deep Reasoning is the field of enabling machines to understand implicit relationships between different things. «*
>
> -- Adar Kahiri, Towards Data Science

## The Ethical Scrum

There are different approaches to allow AI systems ↴ to learn about what is and isn't ethically ↴ responsible. Ethical principles and guidelines ↴ can be programmed into the system (static learning), the system can learn from human behaviour and optimise human values (adaptive learning) and the system can weigh multiple goals on the basis of predefined weight factors (intuitive learning). In many ethical discussions, these approaches are not included sufficiently, because it is the system approach that determines how we need to formulate what is important and what we do and don't include in the system. In the case of static learning, that is a complete set of rules and exceptions, while, in the case of adaptive and intuitive learning, we have to determine which goal functions to include. When we opt in favour of intuitive learning, we not only have to determine which goals we consider to be important, but also what their relative weight is in different situations.

No matter which approach we select, it is virtually impossible to map in advance what the possible implications of the different design choices are, because the optimisation of AI systems is a process of trial and error, which means that new challenges emerge during the development process. Think, for instance, of an area like safety. The original safety principles are affected by choices in the process. To avoid vulnerabilities in the system, choices have to be made during the process and design criteria have to be adjusted. That is why it is important to look not only at the design, but also at the design process.



*» The problems with ethics are not located in the perspectives, but in the processes. «*

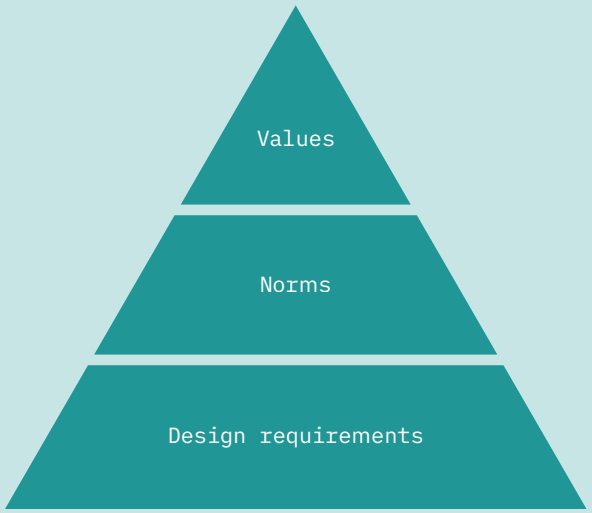
-- Robert de Snoo, Human & Tech Institute

### Ethicists versus Technicians

In discussions about AI and ethics, it is often the ethicists who are talking about technology, or technicians who are talking about ethics. In both cases, it is not their area of expertise, and there is a huge difference between the various approaches. Technicians approach the question mostly from the point of view of optimisation. How can I formalise ↴ values like privacy ↴ and transparency ↴? From that point of view, ethics is a problem that has to be solved; it requires concrete answers. Ethicists, on the other hand, are much more focused on examining the question itself. Existing questions may lead to new questions. However, such abstract insights are hard to translate into the concrete user practice.

To create ethically responsible AI systems, we need both approaches, which is why it is important to adopt a more holistic approach. At the moment, scientists from different disciplines are still competing with each other, when they can complement one another. The development of AI goes beyond technology and philosophy. The use of AI affects society as a whole; the way we work together and live together. Ethicists and technicians should start working together with sociologists and economists. But also with biologists and psychologists. The decision-making process of systems more and more resembles human decision-making processes ↴, so we need to understand how such processes take place in the human brain and express themselves in human behaviour. And that requires a more transdisciplinary approach.

In 2019, researchers at various American universities and companies published an article in *Nature* in which they call for a transdisciplinary scientific research agenda. The aim is to gain more insight into the behaviour of AI systems. Developments in the area of AI bring machines that have a form of 'agency' ever closer. In other words, machines that act independently and make autonomous decisions. That turns machines into a new class of actors in our society, with their own behaviour and ecosystems. Experts argue that this calls for the development of a new research area, namely that of *Machine Behaviour*. The starting point is that we need to study AI systems in the same way we do with animals and humans, namely through empirical observation and experiments.




'Values hierarchy'
– Source: Van de Poel (2013)

### Operationalisation in practice

The different approaches by ethicists, technicians and others can be explained via the so-called 'values hierarchy'.

At the top of the pyramid, we find the most abstract values that many ethicists are concerned with. At the bottom, we find the concrete design requirements that many technicians work with. Ultimately, values have to be operationalised. Design requirements have to be made measurable to allow them to be used for and by AI systems. Normalisation can help close the gap, which roughly speaking involves three steps in the design process:

*1. Conceptualisation:* first of all, values have to be defined. Their meaning has to be clear and universally applicable. That is what many ethical principles and guidelines do. Despite the limitations that many guidelines have, it is an important step that is necessary.

*2. Specification:* the defined values have to be translated into the specific context, because values have different meanings in different situations. This produces more concrete norms that can guide the design process.

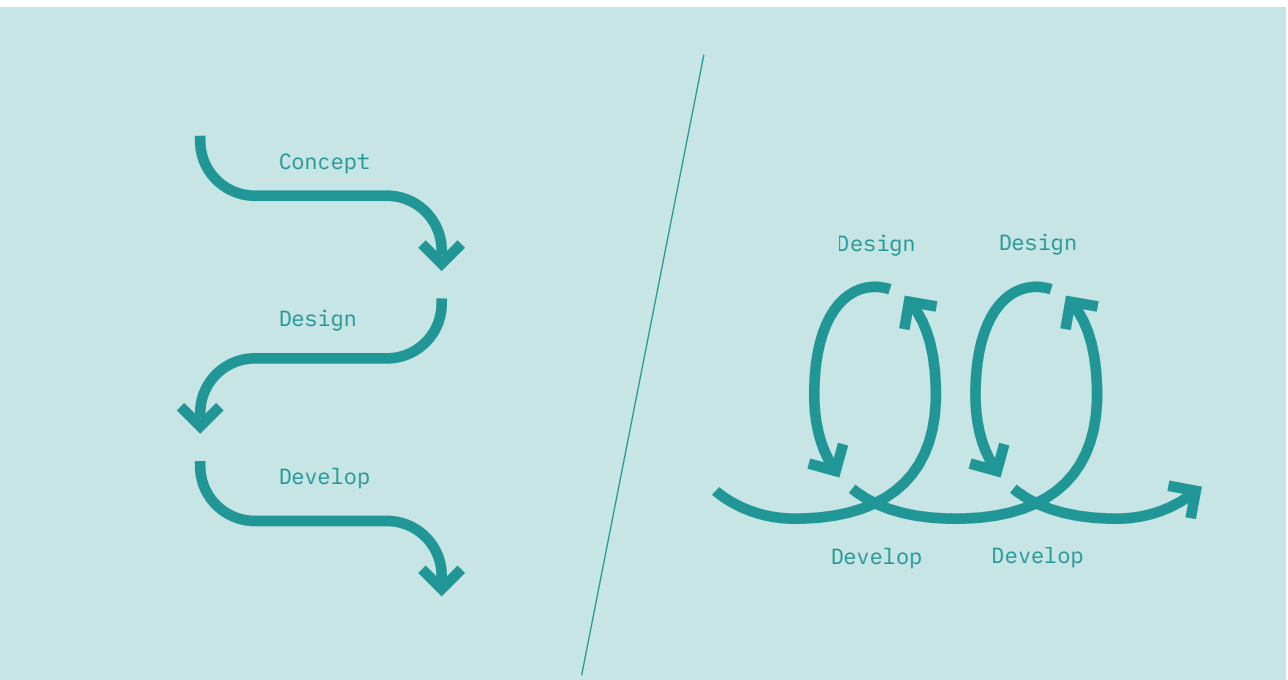
*3. Operationalisation:* the specified norms then have to be translated into measurable design requirements. That way, different design choices can be weighed against each other.

These measurable requirements can then be translated into technical standards, on which quality marks and certificates can be based. The challenge is to be able to harmonise such standards on an international level. That way, technicians have more concrete tools at their disposal to develop ethically responsible AI applications.

*» Standards are broadly supported agreements about the ethics, governance and technology of AI, allowing AI to meet the same requirements everywhere. «*

-- Yvette Mulder, NEN

However, there's an important step that's missing from this process, namely the quantification of what is 'good' and 'not good'. Without such considerations, a system cannot determine what the right action is within a given context. Machines need a kind of moral intuition that has to develop alongside society. Universal values may not change very quickly, but the weighing of different specifications of those values in different contexts does, which requires a different approach to the design process.




### The ethical design process

There are different approaches to the development of software, from so-called 'waterfall' models to 'agile' approaches (Leijnen et al., 2020). In the case of waterfall models, the design is determined at the start of the process, while agile models adopt a more iterative approach, which doesn't go from A to B in a straight line, but accepts the possibility of encountering new challenges in the course of the development process. Sometimes, that means having to take a step back to get ahead or understanding that you shouldn't aim for B, but for C instead.

To a large extent, the waterfall model matches the current value-driven ethical design disciplines, like Value-Sensitive Design (VSD), where the starting point is to define values as early on in the design process as possible, making it possible to maximise as many values as possible against each other, while innovation serves the optimisation of those values. For that to happen, it is important to choose the design at an early stage and make the values explicit. However, a limitation of this waterfall approach is that the focus is primarily on value conflicts that exist at an abstract level, without sufficiently taking into account the tensions that can occur during the design process. In particular when we are talking about AI applications, where factors like safety are crucially important, it is hard to determine in advance which design requirements have to be embedded, because those requirements are subject to change. When you approach this in a static way, that may lead to vulnerabilities in the system. The assumption that requirements that are defined in advance will flow over into the next stage of the design process is wrong. However, in practice, there is a risk that parts of this will be lost and fall outside of the process, making it possible that new requirements are needed to safeguard an ethical application of AI.

An agile approach makes it easier to deal with unforeseen circumstances and adjust the design in the course of the process, making sure that ethical considerations play a role throughout the design process.

However, existing agile approaches focus too much on the functional system requirements. The focus appears to be on the system, not the human element. Developing AI systems that can make ethically responsible decisions requires an approach that also looks at less tangible system requirements, like values. At the moment, the development of AI systems still often focuses on the question how we can improve the reliability of AI systems, instead of on the question how we can develop AI systems that can assign the right value to people. It is therefore important to determine the relative weight of different values in relation to each other in a given context. To be able to do that, we need to understand that the weighing process depends on different factors. There are at least four factors that play a role:

> The stakeholders involved
> Social goals (values)
> Specific interests
> Context

To put ethics in practice, all these factors have to be translated to the design process. The context is different for every application and there are other interests. In this process, the more specific the context is, the stronger the design will be, because it is more customised. And that is exactly what is missing in the current ethical discussions. We need a process that translates ethics to the context.





*Scrum*

Within the agile method, it is especially the scrum process that appears to be able to provide inspiration to safeguard ethics in the design process, for instance because scrum applies so-called ‘user stories’, the advantage of which is that they focus on people and on the different factors that play a role in determining the relative weight of values. A user story is constructed as follows:

As……(*stakeholder*) I want……(*values*)
in order to……(*interests*) given……(*context*).

When we translate this to a specific application domain – like the self-driving car – and place it in a specific context – like a collision between two autonomous vehicles, for the value ‘transparency’, that leads to the following *user stories*:

> As manufacturer, I want to increase traceability, in order to be able to track the system error and avoid collisions
> As user, I want to increase communication, in order to be informed about actions and further steps to be taken in the case of a collision
> As legislator, I want to increase explainability, in order to impose even stricter requirements on the system in case of a collision
> As insurer, I want to increase explainability, in order to be able to determine the guilty party in case of a collision.

By placing the universal value of ‘transparency’ in context, it is specified, drawing a distinction between traceability (the data set and the processes that generate the decision of the AI system), explainability (a suitable explanation of the decision-making process of the AI system) and communication (the communication about the level of accuracy ↴ and limitations of the system). This makes it clear that different stakeholders have different interests within the same value, which also applies to other values, like privacy:

> As manufacturer, I want to increase the value and integrity of the data, in order to help avoid in-accuracies, errors and mistakes in case of a collision
> As driver, I want to increase privacy and data protection, in order to guarantee that my personal information is protected in case of a collision
> As legislator, I want to control access to data, in order to be able to create protocols and manage access to data in case of a collision
> As insurer, I want to control access to data, in order to get clarity about who can access data under what circumstances in case of a collision.

This creates requirements at user level and makes it possible to achieve customisation. The advantage of this approach is that programmers are used to working with these types of processes, which will make their implementation in the design process easier.

*Weighing*

For a manufacturer to be able to develop a self-driving car, it is not only important to know which different interests there are, but also to know how these different interests relate to one another. That can also serve as inspiration during the scrum process, by looking at ‘planning poker’. Normally speaking, this method is used to determine which activities in the design process have priority and need to be carried out first. Measurable values, like 0, 1, 40 and 100, are assigned.



In the case of weighing ethical interests, however, it is better to apply the so-called ‘T-shirt sizing’ method, where the different interests are weighed by assigning S, M, L and XL to express the relative weight of the interests, without creating the illusion that one interest is 100 times more important than another interest. Also, it is not possible to exchange a bad deed for a good deed. Ultimately, all interests are included in the design process, while making it possible to make choices in the design process.

These methods can also be used to detect potential value conflicts and weigh the values involved against each other. What is more important in case of a collision between two autonomous cars? The traceability of the decision-making process for the manufacturer (transparency)? Or protecting the personal information of the user (privacy)? These interests can also be weighed against each other, showing that most trade-offs take place between different users groups, and to a lesser extent between people and system. And again, it is not so much an exchange as it is a ranking of priorities. The goal is ultimately to maximise these values in relation to each other via the design.

*Ethical design game*

A book about ethics in the design process of AI which argues that a lot is written about ethics and that an action perspective is often missing, has to go beyond the written word, which is why we have developed an ethical design game, inspired by the scrum process, that can be used to better streamline the discussion about ethics. The game was developed in collaboration with the standards commission AI of the NEN (The Royal Netherlands Standardization Institute) and the Artificial Intelligence lectorate of the HU University of Applied Sciences Utrecht. In offers policy-makers, developers, philosophers and essentially everybody who is interested in ethics an opportunity to take part in the discussion about what we, as a society, find important when it comes to the development of AI. It helps provide more insight into the various stakeholder perspectives and is designed to contribute to a more constructive discussion about AI and ethics. Because different interests can be weighed against each other, it is possible to make more concrete choices in the design process.

*A final piece of advice*

To safeguard ethically responsible AI applications in the future, all movements and approaches ultimately have to be united. We need ethical guidelines, assessments and standards (principle ethics) to be able to map value conflicts at an early stage and bring values together (consequential ethics), and to then develop AI applications that can make ethically responsible decisions (virtue ethics). That means that the best elements of static and adaptive learning have to be combined to develop AI systems that can aspire to our human intuition and make the right decisions within specific contexts. Complete control over these technologies may be an illusion, but we can design AI systems that will act in our interests.
Now and in the future.



For that to happen, technicians, ethicists, sociologists, economists and others have to join forces to work on AI systems that can make ethically responsible decisions. It is important to specify what we consider to be ‘good’ and ‘bad’ behaviour and how much value we assign to different elements. So there’s an important task for legislators. Programmers should only be responsible for making the system as intelligent as possible and for optimising the goal functions. It is up to the legislator to define those goal functions after listening to society. Without clear legislation, a developer or user doesn’t know on the basis of what they will be assessed. We’re not there yet. But I do hope that this publication and tool have brought the development of ethically responsible AI applications a little bit further.

*» Real ethics starts where the use of words stops. «*

-- Albert Schweitzer, 1923

Real ethics starts
where the use of words

stops.

-- Albert Schweitzer, 1923

# The ethics of AI in practice

*By Bernard ter Haar, Special advisor, Ministry of the Interior and Kingdom Relations*

A lot is thought, written and spoken about the importance of securing ethical values for artificial intelligence (AI), or even for the entire digital ecosystem. There's a lot of theory, and little practice. And all that theory has a bit of a paralysing effect. As though ethics is really too lofty or difficult a subject to put into practice. That is, at any rate, an enormous misunderstanding.

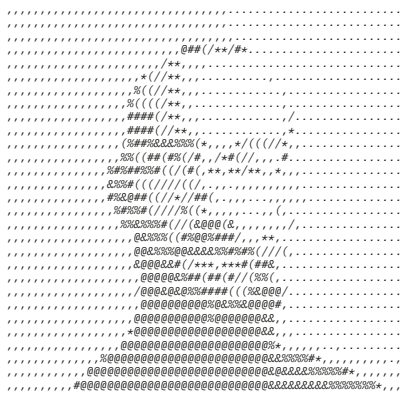
Simply put, I see ethics as thinking about good and evil. And we do that every day. We constantly judge developments in this world on whether they are good or bad. For instance, many people think it's ethically irresponsible to let refugees suffer on some Greek island, while others think it's ethically important to defend authentic Dutch culture. And not only do we judge every day, our judgment also shifts over time. Until the 1960s, female teachers had to quit their jobs once they got married, based on our ethical stance in relation to family values. Nowadays, we see that as oppression of women, at least most of us do. There's never a complete consensus about ethical values. In China, they thought long and hard about good and bad and the possibilities of AI. They set up a kind of social score system, in which people are scored on the basis of whether they behave well or badly. In the eyes of the Chinese a good way to increase the ethical content of social behaviour. In the Netherlands, we abhor such an approach, because it is at odds with our views on individual freedom and expression. The fact that an ethically positive goal like social cohesion is used to sugar-coat purely commercial interests, like Facebook does, is also viewed with increasing scepticism.





So securing ethical values is an everyday activity, and an essential and very dynamic element of our democratic constitutional state. We secure our ethics in laws and rules that are created democratically. And yet, as I wrote earlier, when we are talking about ethics and AI, we see a lot of theory and little practice. Where do things go wrong? Both on a departmental and parliamentary level, legislators appear reluctant to act. There is a lack of knowledge and insight. There is the question at which level the responsibility lies, national, European or global. And there is a widespread fiction about the value of total freedom of the Internet. Protection of property is the basis of almost every ethical system, but do digital data have an owner? Of course it does, but that needs to be translated into laws and ethics.

To create ethically responsible AI, there are a few important things. Legislators need to know what is happening in the digital world. It goes further than that, obviously. We all need to know. After all, democracy can only function when there's a broad social discussion. The transparency of every digital action has to be increased considerably. That is complicated, which is why we need to give it serious thought. Once we know more about what happens in the digital realm, which includes the developing world of AI, we can decide on a day to day basis what is acceptable and what is not. And we will also start limiting the latter category in one way or another. Experience has to make us wiser, there's no way we can regulate everything in advance. That's no different in the physical world. It does require legislators who are able to keep up with the pace of digital development. A challenge, but one we cannot walk away from!



# Don’t put all the responsibility on the programmer’s shoulders

*By Rudy van Belkom*

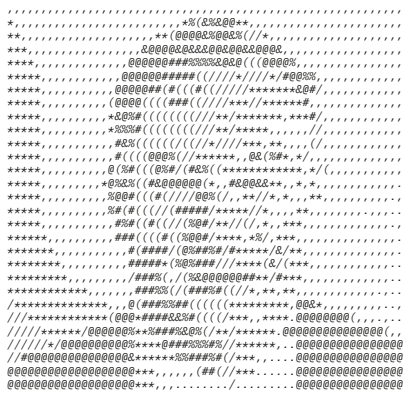
There was a lot of criticism regarding the ‘appathon’ organised by the Dutch Ministry of Health to try and use smart technologies to stop the corona virus from spreading. The process was too hasty and chaotic. Developers had to try and make some last-minute improvements to the tracking apps in a pressure cooker. According to the experts involved, the results were disappointing. None of the apps met the relevant privacy guidelines. Ethically irresponsible, was the final verdict.

> And yet, the way I see it, the real criticism doesn’t involve the appathon itself, but the way we all conduct ethics. At the moment, the final responsibility lies with the programmers, which is not as it should be (and may even be unintentional). We think that we have covered everything with different ethical principles and guidelines, but nothing could be farther from the truth, because those principles and guidelines say nothing about the way values like privacy need to be expressed in mathematical terms. And the development of AI is all about statistics. Smart systems have to be able to extract patterns from large amounts of data and learn from that independently. In addition to the fact that that means that personal information is exposed, the system can unintentionally disadvantage certain groups of people by misinterpreting the data.

In that sense, it’s not so much about privacy, but about fairness. And the question in that case is what is fair from a statistical point of view. Do we not want to overlook any corona cases, or don’t we want to unjustly quarantine people? And what if it turns out that people in some areas have a higher risk of contamination? Do those variables have to be factored in, or do we want everyone to be treated equally? The question then is what is ‘fair enough’. What percentage of erroneous quarantine cases can we and do we want to accept? In practice, the various mathematical definitions of fairness turn out to be mutually exclusive. So if we don’t quantify these guidelines, programmers will, according to the current approach, have to make a choice themselves.

> Furthermore, what we consider to be fair is context-dependent. Several months ago we couldn’t even imagine having to place people in quarantine in the first place, and we will probably respond differently when we are talking about prison sentences, rather than a relatively luxurious quarantine in our own homes. Universal values may not be subject to change, our norms certainly are. That makes it not only difficult, but also irresponsible to programme principles and guidelines into AI systems. Ultimately, the system itself has to be able to make moral considerations. That sounds scary, but without a morally intuitive system, it is almost impossible to apply AI in practice in an ethically responsible way. The only reason that people are able to deal with rules is because we can translate them into behaviour within a given context. At the moment, social distancing is the norm, but if someone were to stumble and fall down a flight of stairs, I believe I should catch that person if I am able to.

As such, we should spend less energy in setting up ethical guidelines and spend more time building systems that can make ethically responsible decisions within a given context. Ethics is above all a design issue. In addition to programmers and ethicists, sociologists, psychologists, biologists and economists should also be involved. If we are unable to make that happen, then perhaps we shouldn’t want to use AI systems at all. Or accept that our ethics are unethical.



# Ethics in action

*By Patrick van der Duin,*
*Director Netherlands Study Centre*
*for Technology Trends*

Credit where credit is due. It was former Prime Minister Jan Peter Balkenende who, in 2002, said we should talk more about norms and values. I was among those who found that rather amusing and felt that it was an ancient discussion. And that his norms and values aren't the same as mine (and that it's actually values and norms). But now, in 2020, norms and values are more important than ever before. Balkenende proved to be a veritable prophet, who was rewarded when a norm was named after him: the Balkenende norm, which states that managers in the public and semi-public sector are not allowed to earn more than a government Minister.

> The current 'ethical turn' (perhaps similar to the 'linguistic turn') also fills me with a sense of nostalgia. In 2002, I was working at the Technology, Governance and Management faculty at Delft University of Technology. There was a section Philosophy of Technology, but that was hardly a grand affair going by the small number of staff members: a mere handful. But in 2020, it is the largest section of the entire faculty. Under the header of 'responsible innovation', the ladies and gentlemen ethicists and philosophers have stood up from their respective armchairs to examine how they can put ethics into action. No more idle philosophies, but research into how design processes can be managed ethically and to that end develop practical methods that really have to make the world a little more humane.

The STT study into AI in the future, carried out by project leader Rudy van Belkom, has to be seen in this context of the increasing importance of ethics in our society, economy and technology. This final part is a logical conclusion of the trilogy, which began by examining what AI is, how diverse it is and also what it is not.



Not to limit the research but above all to create clarity about what AI means, to allow us to put the hype surrounding AI into perspective. The second study deliberately created some confusion by using a number of scenarios to show the various possible futures of AI, with the aim of breaking through the dominant discourse that AI is something that is really bad for humanity and to show that there are alternatives. Alternatives that didn't just materialise out of thin air but that are the result of what we, as a society, want and desire. Many argue that AI may well be the most far-reaching technology that mankind ever has developed and will develop, which is exactly why 'human agency' is important, so that we shape AI the way we want to.

> In this third and final part, Rudy van Belkom has shown how you can turn theory surrounding AI into practice. Not just by explaining that ethics (like AI) is a many-headed, well-intended monster, but also by arguing that it can only be put into practice by including anyone and everyone who is interested in and cares for AI. Ethics is something to talk about, but it's also something to do. This STT study is one of the first to establish a direct connection between thinking about the future and acting accordingly, which, as far as the development of AI is concerned, is not a luxury, but a necessity. Not only to prevent AI from going into the 'wrong' direction, but above all to use the possibilities of AI to further the norms and values of our society.

# Glossary

Accountability

Responsibility that is legally prescribed is called accountability.
III.1.2 Urgent ethical issues ↴

Accuracy

Generally speaking, the more complete and elaborate the dataset is with which an AI system is trained, the more accurate the system will be. III.2.2 Conflicting values ↴

Affective computing

Affective computing refers to systems that can detect and recognize emotions.

Artificial Intelligence (AI)

The most dominant association with AI is machine learning. Research by the World Intellectual Property Organisation (WIPO) from 2019 also shows that machine learning is the most dominant AI technology included in patent applications, which is why the focus in this study is mainly on machine learning and related methods.

Algorithms

An algorithm is a mathematical formula. It is a finite sequence of instructions that works from a given starting point towards a predetermined goal.

Biases

Because we are programmed by evolution to save as much energy as we can, the way we process information can lead to fallacies, also known as cognitive biases.

Brain

We are not yet able to model complex concepts that we want to link to the brain, like awareness and free will, and we cannot connect them individually to certain areas of the brain.



Clusters

When creating clusters, the algorithm independently searches for similarities in the data and tries to recognize patterns.

Common sense

Common sense consists of all knowledge about the world; from physical and visible aspects, to cultural and therefore more implicit rules, like how we should treat each other.

Consequential ethics

Consequential ethics states that it is the consequences of a certain action that determine whether the action is ‘right’. In other words, the behaviour has to have positive consequences, even when it undermines certain principles. So it is not about the action itself, but about the consequences. III.1.3 A matter of ethical perspective ↴

Contest

At the moment, the interest in AI is so great that world powers have entered into a kind of AI contest. Research by PwC from 2017 shows that, in 2030, the worldwide Gross Domestic Product (GDP) will be 14% higher thanks to developments in the area of AI. According to Russian President Putin, the country with the best AI will rule the world (2017).

Deep learning

Deep learning is a machine learning method that uses various layered artificial neural networks.

Deep reasoning

This new approach tackles problems in the old approaches by combining them. Deep reasoning solves the scalability problem of symbolism (it is impossible to programme all options efficiently), while at the same time tackling the data problem of neural networks (large data sets are often not available or incomplete).

### Decision-making

We like to believe that human beings are rational creatures. But the decision-making process is capricious. In addition to factual information, perception and ambition also play an important role.

### Dependence on data

One of the main limitations of AI is that it depends on huge amounts of data, which is why, in the case of deep learning, people sometimes talk about data-hungry neural networks. As a result, the technology does not perform well in peripheral cases where there is little data available.

### Empathy

Empathy is the ability to imagine yourself in the situation and feelings of other people. Empathy also allows us to read and understand the non-verbal communication of others.

### Ethical guidelines

In recent years, various companies, research institutes and government organisations have set up different principles and guidelines for *ethical AI*, at a national, continental and global level. III.2.1 From corporate to government ↴

### Ethics

Ethics is a branch of philosophy that engages in the systematic reflection of what should be considered good or right actions. III.1 AI and Ethics ↴

### Explainability

It is increasingly difficult for people to determine on the basis of which data the results of AI are based, which is why AI is still often compared to a black box.

### Explanation

Within the decision-making process of AI, the explanation and explainability involve the explanation and traceability of the decision in retrospect. III.1.2 Urgent ethical issues ↴



### Fairness

Fairness is a much used principle in ethical guidelines and assessment tools. Philosophers have thought for hundreds of years about the concept of fairness, and with the arrival of AI, a whole new dimension has been added, because now, the concept of fairness has to be expressed in mathematical terms. III.2.3 Practical challenges ↴

### Formal logic

In the first phase of AI, from 1957 to the late 1990s, formal logic, in other words if-then rules, were the main tool being used. This form of AI was focused predominantly on high level cognition, like reasoning and problem-solving.

### Freedoms and rights

In essence, freedom refers to the freedom people have to determine how to organise their lives. This is even a right, the right of self-determination. However, that right is then limited by prohibition to harm others. III.1.2 Urgent ethical issues ↴

### General AI

General Artificial Intelligence should be able to carry out all the intellectual tasks that a human being can also perform.

### Image recognition systems

Image recognition systems often use a so-called Convolutional Neural Network (CNN), which acts as a filter moving across the image, looking for the presence of certain characteristics.

### Intelligence

Intelligence can be described as a sequence of mental abilities, processes and skills, like the ability to reason and adapt to new situations.

Intentionality

Even if you were to programme all the knowledge in the world into a computer, the question remains whether that computer genuinely understands its actions. That understanding is also referred to as intentionality.

Justice

When we are talking about justice, in essence, we are talking about the equality of people. People should be treated equally and be given equal opportunities. III.1.2 Urgent ethical issues ↴

Machine biases

Not only human intelligence, but also AI can be biased. The output of algorithms can be biased in terms of gender and race. The explanation for that is simple. When your input isn't pure, then neither will the output be. So the biases of algorithms are caused by the cognitive biases of people.

Machine learning

Machine learning involves a revolution in which it is no longer people who programme (if this, then that), but in which machines themselves deduce rules from data.

Mathematics

In essence, AI is 'ordinary' mathematics. Albeit a very advanced form of mathematics, but mathematics nonetheless. It is above all a tool to realize an optimisation goal.

Morality

In discussions about AI, people often confuse ethics and morality, even though there is a clear difference. Morality is the entirety of opinions, decisions and actions with which people (individually or collectively) express what they think is good or right. Ethics, on the other hand, is the systematic reflection on what is moral. III.1.3 A matter of ethical perspective ↴





Narrow AI

Narrow Artificial Intelligence is a form of AI that is very good in carrying out specific tasks, for instance playing chess, making recommendations and making quantifiable predictions.

Neural networks

Artificial neural networks can be used within deep learning and are originally based on the human brain, whereby neurons are connected to each other in a layered fashion.

Nightmare scenarios

Scenarios about robots rising up have been a popular storyline for almost 100 years (assuming that director Fritz Lang's 1927 movie 'Metropolis' is the first real science fiction movie in which a robot has bad intentions). III.1.1 Ethics in the spotlight ↴

Principle ethics

In the case of principle ethics, a principle is used as the starting point, for instance respect for life and human dignity. When solving an ethical problem, one or more of these principles need to be taken into account. The principle has to be applied at all times, regardless of the consequences. III.1.3 A matter of ethical perspective ↴

Privacy

In the European Treaty for Human Rights, privacy is included as the right to respect of people's private lives, which requires a fair balance between the social interest that a technology serves and the extent to which it violates people's private lives. III.1.2 Urgent ethical issues ↴

Responsibility

In discussions about the development of AI, the term ‘responsibility’ is often mentioned. For instance, who is responsible in an accident involving a self-driving car? However, to determine who is responsible, we first need to determine what it is they are responsible for and what behaviour can and cannot be defended. In addition, the question is how we can deduce the level of responsibility. III.1.2 Urgent ethical issues ↴

Superintelligence

Artificial Super Intelligence (ASI) can be realised when AI transcends the abilities of the human brain in every possible domain.

Transparency

Within the decision-making process of AI, transparency is primarily about the process and the predetermined criteria. III.1.2 Urgent ethical issues ↴

Trust

Research from, among others, the University of Pennsylvania from 2014 shows that, when people see an algorithm make a small and insignificant mistake, chances are they well have lost all trust. Among researchers, this is also referred to as algorithm aversion.

Virtue ethics

In the case of virtue ethics, it is not the rules of certain principles that are central to moral judgments, but the character of the actor, whose actions are separated from their explicit consequences. Right actions require certain characteristics, or virtues. III.1.3 A matter of ethical perspective ↴

## Sources


AI Now Institute (2018). AI in 2018: a year in review. Consulted on https://medium.com/@AINowInstitute/ai-in-2018-a-year-in-review-8b161ead2b4e

AI Now Institute (2019). AI in 2019: a year in review. Consulted on https://medium.com/@AINowInstitute/ai-in-2019-a-year-in-review-c1eba5107127

Aliman, N. & Kester, L. (2019) Transformative AI Governance and AI-Empowered Ethical Enhancement Through Preemptive Simulations. Consulted on https://delphi.lexxion.eu/article/DELPHI/2019/1/6

Araujo, T. et al. (2018). Automated Decision-Making Fairness in an AI-driven World: Public Perceptions, Hopes and Concerns. Consulted on http://www.digicomlab.eu/wp-content/uploads/2018/09/20180925_ADMbyAI.pdf

Beijing Academy of Artificial Intelligence (2019). Beijing AI Principles. Consulted on https://www.baai.ac.cn/blog/beijing-ai-principles

Blauw, S. (2018). Algoritmes zijn even bevooroordeeld als de mensen die ze maken. Consulted on https://decorrespondent.nl/8802/algoritmes-zijn-even-bevooroordeeld-als-de-mensen-die-ze-maken/4676003192124-d89e66a9

Boer, M. de (2020). The many futures of Artificial Intelligence: Scenarios of what AI could look like in the EU by 2025. Consulted on https://www.pwc.nl/nl/actueel-publicaties/assets/pdfs/the-many-futures-of-artificial-intelligence.pdf

Bostrom, N. (2016). Superintelligence: Paths, Dangers, Strategies. Oxford: Oxford University Press

Dalen, W. van (2012). Ethiek de basis; Morele competenties voor professionals. Groningen/Houten: Noordhoff Uitgevers





Delvaux, M. (2016). DRAFT REPORT with recommendations to the Commission on Civil Law Rules on Robotics. Consulted on https://www.europarl.europa.eu/doceo/document/JURI-PR-582443_EN.pdf?redirect

Duin, P. van der, Snijders, D. & Lodder, P. (2019). The National Future Monitor: How do Dutch people think about technology and the future? Den Haag: STT

Eckersley, P. (2018). Impossibility and Uncertainty Theorems in AI Value Alignment (or why your AGI should not have a utility function) Consulted on https://arxiv.org/abs/1901.00064

ECP (2018). Artificial Intelligence Impact Assessment (AIIA). Consulted on https://ecp.nl/wp-content/uploads/2019/01/Artificial-Intelligence-Impact-Assessment-English.pdf

Elsevier (2018). ArtificiaI Intelligence: How knowledge is created, transferred, and used. Consulted on https://www.elsevier.com/research-intelligence/resource-library/ai-report

Est, R. van & Gerritsen, J. (2017). Human rights in the robot age Challenges arising from the use of robotics, artificial intelligence, and virtual and augmented reality. Consulted on https://www.rathenau.nl/nl/digitale-samenleving/mensenrechten-het-robottijdperk

Eubanks, V. (2018). Automating Inequality; How High-Tech Tools Profile, Police, and Punish the Poor. New York: St. Martin's Press

European Commission (2020). Attitudes towards the impact of digitalisation on daily lives. Consulted on https://ec.europa.eu/commfrontoffice/publicopinion/index.cfm/survey/getsurveydetail/instruments/special/surveyky/2228

European Commission (2020). White Paper on Artificial Intelligence: a European approach to excellence and trust. Consulted on https://ec.europa.eu/info/publications/white-paper-artificial-intelligence-european-approach-excellence-and-trust_en

Fjeld, J., Achten, N., Hilligoss, H., Nagy, A. & Srikumar, M. (2020). Principled Artificial Intelligence: Mapping Consensus in Ethical and Rights-Based Approaches to Principles for AI. Berkman Klein Center Research Publication No. 2020-1. http://dx.doi.org/10.2139/ssrn.3518482

Future of Life Institute (2017). Asilomar AI Principles. Consulted on https://futureoflife.org/ai-principles/

G7 Innovation Ministers (2018). G7 Innovation Ministers' Statement on Artificial Intelligence. Consulted on http://www.g8.utoronto.ca/employment/2018-labour-annex-b-en.html

Gartner (2019). Top 10 Strategic Technology Trends for 2019: Digital Ethics and Privacy. Consulted on https://www.gartner.com/en/documents/3904420/top-10-strategic-technology-trends-for-2019-digital-ethi

Genesys (2019). New Workplace Survey Finds Nearly 80% of Employers Aren't Worried About Unethical Use of AI — But Maybe They Should Be. Consulted on https://www.genesys.com/en-gb/company/newsroom/announcements/new-workplace-survey-finds-nearly-80-of-employers-arent-worried-about-unethical-use-of-ai-but-maybe-they-should-be

Google (2019). Perspectives on Issues in AI Governance. Consulted on https://ai.google/static/documents/perspectives-on-issues-in-ai-governance.pdf

High-Level Expert Group on Artificial Intelligence (2019). Ethics guidelines for trustworthy AI. Consulted on https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai

Hill, K. (2020). The Secretive Company That Might End Privacy as We Know It. Consulted on https://www.nytimes.com/2020/01/18/technology/clearview-privacy-facial-recognition.html

Hoven, J. van den, Miller, S. & Pogge, T. (2017). Designing in Ethics. Cambridge: Cambridge University Press



IBM (2019). Everyday Ethics on AI. Consulted on https://www.ibm.com/design/ai/ethics/everyday-ethics

Jobin, A., Ienca, M. & Vayena, E. (2019). The global landscape of AI ethics guidelines. *Nat Mach Intell 1, 389–399.* https://doi.org/10.1038/s42256-019-0088-2

Kant, I. (1788). Kritik der praktischen Vernunft. Keulen: Anaconda Verlag

Kaur, H. et al. (2020). Interpreting Interpretability: Understanding Data Scientists' Use of Interpretability Tools for Machine Learning. Consulted on http://www.jennwv.com/papers/interp-ds.pdf

Leijnen, S. et al. (2020). An Agile Framework for Trustworthy AI. ECAI 2020 position paper.

Maurits, M. & Blauw, S. (2019). In de stad van de toekomst praten lantaarnpalen mee en burgers niet. Consulted on https://decorrespondent.nl/9148/in-de-stad-van-de-toekomst-praten-lantaarnpalen-mee-en-burgers-niet/4859813360776-3fcc1087

Microsoft Research (2018). Manipulating and Measuring Model Interpretability. Consulted on https://arxiv.org/pdf/1802.07810.pdf

Ministry of Economic Affairs and Climate Policy (2019). Strategic Action Plan for Artificial Intelligence. Consulted on https://www.government.nl/documents/reports/2019/10/09/strategic-action-plan-for-artificial-intelligence

NeurIPS. (z.d.). Code of Conduct. Consulted on https://nips.cc/public/CodeOfConduct

OECD (2019). Principles on AI. Consulted on http://www.oecd.org/going-digital/ai/principles/

Pew Research Center (2018). The Data on Women Leaders. Consulted on https://www.pewsocialtrends.org/fact-sheet/the-data-on-women-leaders/#ceos

Poel, I. van de & Royakkers, L. (2011). Ethics, Technology and Engineering; An Introduction. Hoboken: Wiley-Blackwell

Poel, I. van de. (2013). Translating Values into Design Requirements. In: MichelFelder D., McCarthy N., Goldberg D. (eds), Philosophy and Engineering: Reflections on Practice, Principles and Process. Vol. 15 of Philosophy of Engineering and Technology, Springer, Dordrecht (pp.253-266)

ProPublica (2016). Machine Bias. Consulted on https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

Rahwan, I. (2019). Machine behavior. Consulted on https://www.nature.com/articles/s41586-019-1138-y

Russel, S. (2017). 3 principles for creating safer AI. Consulted on https://www.ted.com/talks/stuart_russell_3_principles_for_creating_safer_ai

Selbst, A.D. et al. (2019). Fairness and Abstraction in Sociotechnical Systems. Consulted on https://dl.acm.org/doi/pdf/10.1145/3287560.3287598

Snijders, D., Biesiot, M., Munnichs, G. & Est, R. van (2019). Citizens and sensors: Eight rules for using sensors to promote security and quality of life. Den Haag: Rathenau Instituut

Smart Dubai (2019). AI Ethics Principles & Guidelines. Consulted on https://www.smartdubai.ae/pdfviewer/web/viewer.html?file=https://www.smartdubai.ae/docs/default-source/ai-principles-resources/ai-ethics.pdf?sfvrsn=d4184f8d_6

The Pontifical Academy for Life (2020). Rome Call for AI Ethics. Consulted on http://www.academyforlife.va/content/dam/pav/documenti%20pdf/2020/CALL%2028%20febbraio/AI%20Rome%20Call%20x%20firma_DEF_DEF_.pdf

UNI Global Union (2017). Top 10 Principles for ethical AI. Consulted on http://www.thefutureworldofwork.org/media/35420/uni_ethical_ai.pdf

U.S. Department of Defense (2020). AI Principles: Recommendations on the Ethical Use of Artificial Intelligence. Consulted on https://media.defense.gov/2019/Oct/31/2002204458/-1/-1/0/DIB_AI_PRINCIPLES_PRIMARY_DOCUMENT.PDF

Utrecht Data School (2017). Data Ethics Decision Aid (DEDA). Consulted on: https://dataschool.nl/deda/deda-worksheet/?lang=en

Verbeek, P.-P (2011). De grens van de mens: Over ethiek, techniek en de menselijke natuur. Rotterdam: Lemniscaat

Weijer, B. van de (2019). Binnenkort op de weg: zelfrijdende robottaxi's zonder reservemens achter het stuur. Consulted on https://www.volkskrant.nl/economie/binnenkort-op-de-weg-zelfrijdende-robottaxi-s-zonder-reservemens-achter-het-stuur~b671f376/

Wernaart, B. (2021). Developing a roadmap for the moral programming of smart technology. Consulted on https://www.sciencedirect.com/science/article/pii/S0160791X20312690

Wilson, B., Hoffman, J. & Morgenstern, J. (2019). Predictive Inequity in Object Detection. Consulted on https://arxiv.org/abs/1902.11097

# Appendices

## About the author

Rudy van Belkom MA studied Brand, Design & Reputation Management at the European Institute for Brand Management (EURIB). For his Master's thesis, he conducted research into the factors that are decisive in the acceptance of radical technological innovations. Trust in the technology turns out to play a crucial role. Until joining STT, he conducted research into the role of science fiction in futures research at the Fontys lectorate of Futures Research & Trend watching and for years taught concept development and trend research in higher education.

## Participants

For this part of the futures research, an online questionnaire was distributed among three groups, namely experts, administrators and students, which was filled in by over 100 respondents. On behalf of STT and myself, many thanks for filling this questionnaire. In addition, a word of thanks to the experts that contributed their thoughts, reading and writing to this part of the futures research.

| | | |
|---|---|---|
| Roland Bijvank | HU | Lecturer Informatics |
| Marc Burger | Capgemini | CEO |
| Patrick van der Duin | STT | Director |
| Arjen Goedegebure | OGD IT-services | Data Scientist |
| Bernard ter Haar | Ministry of the Interior and Kingdom Relations | Special advisor |
| Fred Herrebout | T-Mobile | Senior Strategy Manager |



| | | |
|---|---|---|
| Jeroen van den Hoven | Delft University of Technology | Professor of Ethics and Technology |
| Marijn Janssen | Delft University of Technology | Professor ICT & Governance |
| Leon Kester | TNO | Senior Research Scientist |
| Maria de Kleijn-Lloyd | Kearney | Senior Principal |
| Stefan Leijnen | HU | Lector Artificial Intelligence |
| Leendert van Maanen | Utrecht University | Assistant professor in Human-centered AI |
| Michel Meulpolder | BigData Republic | Lead Data Architect |
| Evelien Mols | STT | Graduate student |
| Gerard Nijboer | Municipality of Rotterdam | Process Manager Innovation |
| Yvette Mulder | NEN | Committee manager AI |
| Eefje Op den Buysch | STT | Futures Explorer |
| Marieke van Putten | Ministry of the Interior and Kingdom Relations | Senior Innovation Manager |
| Edgar Reehuis | Hudl | Engineer |
| Jelmer de Ronde | SURF | Project Manager SURFnet |
| Martijn Scheltema | EUR | Professor in Civil Law |
| Klamer Schutte | TNO | Lead Scientist Intelligent Imaging |
| Robert de Snoo | Human & Tech Institute | Founder |
| Marc Steen | TNO | Senior Research Scientist |
| Maarten Stol | Brain-Creators | Principal Scientific Advisor |

## Netherlands Study Centre for Technology Trends

The Netherlands Study Centre for Technology Trends (STT) was founded in 1968 by the Royal Institute of Engineers (KIVI). STT is an independent foundation that is funded by contributions from the government and the business community. STT carries out broad futures explorations at the crossroads of technology and society that transcend domains and are multidisciplinary. The General Board (GB) of STT consists of top people from government, business community, the research community and social organisations. The GB contributes to the STT programme, is involved in explorations and forms an important think tank within which the board members discuss future technological developments and innovation.

In addition, the STT Academy is involved in various activities, like co-funding special chairs, method development and the management of Network Futures Research and Young STT. The latter consists of young high potentials from the participating organisations.

Information about STT, its activities and its products can be found on www.stt.nl.



## Previous publications STT

STT90 Bioprinters, resource-roundabouts and brainternet? How we produce, consume and redistribute in 2050. Silke den Hartog-de Wilde, 2019

STT89 Safety in the future. Carlijn Naber, 2019

STT88 Long Live Learning. About Learning, Technology and the Future. Dhoya Snijders, 2018

STT87 And then there was light …; Transitions towards an emission-free energy system. Soledad van Eijk, 2017

STT86 Data is power. About Big Data and the future. Dhoya Snijders, 2017

# Colophon

*STT92 part 3: AI no longer has a plug*

Research and project leader: Rudy van Belkom STT
Text and language editing: Japke Schreuders, STT
Graphic design: Yurr Studio

NUR no 950

ISBN 978-94-91397-23-3

Keywords
AI, artificial intelligence, ethics, future, society, design, backcasting

In this series about the future of AI, we previously published part 1: 'Submarines don't swim', about the technology of AI, and part 2: 'Computer says no', about the social impact of AI.



Netherlands Study Centre for Technology Trends
Koninginnegracht 19
2514 AB The Hague
+31(0)70-302 98 30
info@stt.nl