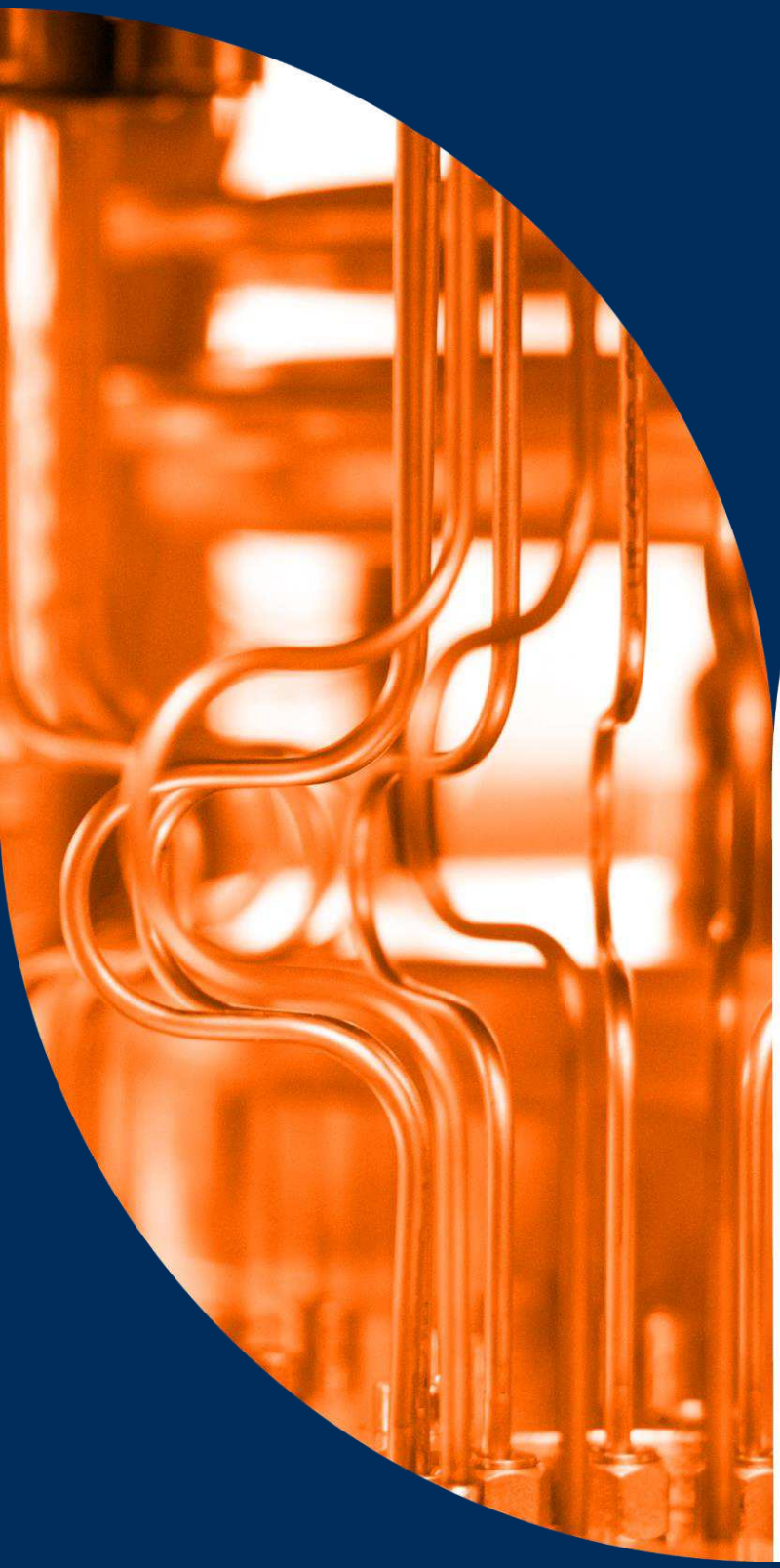

Lessons from AI: Risk Management

White Paper 3

Quantum Delta Nederland

Lessons from AI: Risk Management

Auteurs

MSc. Jord Goudsmit
Dr. Luca Possati
MSc. Lauren Challis
Dr. Ulrich Mans

Inhoudsopgave

Lessons from AI: Risk Management

1 Introduction

1.1 About the series

Technology is not developed in a vacuum. As such, its Ethical, Legal, Societal Aspects (ELSA) must be carefully considered. ELSA posits that technological innovations are not independent of our current society and must be developed to cohesively integrate and enhance ethical, legal and societal values we hold to be paramount. Quantum technology is no different. Even though quantum technologies still are at their nascent stages, research advances are moving fast and the field is gradually moving from science to application. It is the right time to look at how best to consider the ethical, legal and societal aspects of quantum technologies. In recent publications, various groups have argued that quantum stakeholders should ensure that mistakes made in the field of AI should not be repeated, and that there is a need to work out guidelines ahead of fully functional quantum systems.

Quantum Delta NL's inspirational papers series attempts to do just this: look ahead and consider key issues of the development of quantum technology. In doing so, we first took a step back and searched for lessons from the development and regulation of Artificial Intelligence. This is part of Quantum Delta NL's mission to study and facilitate societal impact of quantum technology. Its Centre for Quantum & Society – the first in the world – is the place where this work comes together. This series consists of three papers about AI. One paper focuses on the institutional engagements across policy, science, industry and civil society in the EU context. Who was involved, when, and with what result? The two others zoom in on risk management and communication. Each paper draws from in depth interviews with experts from a varied number organisations, and presents a select number of "inspirations", which we think are worth taking on board on our journey to make quantum technology a positive force for science, business and society.

We thank all the participants for their time and collaboration.

1.2 About Quantum Delta NL

Quantum Delta NL is the keystone of the Netherlands' national ecosystem for excellence in quantum innovation, the foundation that connects the most important knowledge institutions in the field of quantum technology in the Netherlands. Our starting position is excellent: Dutch universities and knowledge institutes are leaders in the

field of quantum technology research, our startup and industrial ecosystem is growing continuously, and our national policy is strongly developed. With the allocation of 615 million euros from the National Growth Fund, we will execute the Netherlands' National Agenda Quantum Technology (NAQT) over the next 7 years. Our mission is to further strengthen our ecosystem to become the most relevant ecosystem for Europe. Quantum Delta NL consists of five major quantum hubs and several universities and research centres. The hubs collaborate on innovation by bringing together top-quality scientists, engineers, students and entrepreneurs, working together on the frontier of quantum technology.

1.3 A short story of AI risk management

As early as 2010, numerous AI initiatives and actors cropped up, initially engaged in research. Soon, stakeholders looked into related regulatory issues. DeepMind is a prominent example, which was established in 2010 in the UK with the goal of developing 'general and capable problem-solving systems, known as artificial general intelligence (AGI)'[2]. At the time, it seemed relatively early, and consideration of regulatory consequences largely remained a longer-term problem. Fast forward to 2014, the European Commission (EC) created the Big Data Value Association (BDVA) as a public-private partnership with a focus on Big Data. This later changed into DAIRO (Data, AI and Robotics), reflecting the move away from Big Data towards AI as the dominant framing used for policy discussions. At the same time, negotiations over data privacy gained traction over the years – and culminated in the adoption of the General Data Protection Regulation (GDPR) in 2018.

With the GDPR, the EU presented a new blueprint for regulatory activity for emerging technologies, and shifted the attention of non-governmental and industry stakeholders from stimulating the use of AI (and growing the workforce) towards regulatory issues. In 2015 OpenAI was established as a private, non-profit initiative to develop AI and to ensure that 'AGI benefited humanity'[3]. Its purpose to develop 'friendly AI' was a tacit recognition that AI needed to be regulated in some fashion to truly harness its benefits. The announcement and establishment of the Partnership on AI in 2016[4] was yet another expression from private industry that AI should be regulated. The scope and extent of such regulation, however, remained contested. Private industry's drive to establish AI standards to regulate its use culminated in the Asilomar AI Principles in 2017, which included 23

ethical guidelines for AI research and development.[5]

From this point onwards, the focus on AI grew steadily across EU institutions and the Brussels landscape of stakeholders. On 10 April 2018, 25 European countries signed a Declaration of Cooperation[6] to collaboratively work on three main pillars for AI: industry capacity, socio-economic issues, and legal and ethical concerns. With the goal of establishing a Digital Single Market, work on these three pillars commenced immediately after the announcement of the Declaration of Cooperation. Also in 2018, the European Group on Ethics in Science and New Technologies called for a common and internationally recognised ethical and legal framework on AI.[7] In June 2018, both the AI High Level Expert Group (HLEG) and EU AI Alliance were set up to pool insights from leading AI experts and prepare documents that would later provide the groundwork for new AI regulation. National governments also got involved. A notable example is the German Data Ethics Commission established in 2018 with the aim to develop ethical and legal frameworks for AI. The UN and the OECD also started to develop AI regulation, for instance when convening the 2017 ITU AI Summit for Good.
The EC released a White Paper on AI in 2020, formally documenting its regulatory vision and strategy for AI. Its publication included a public consultation which solicited public feedback on its regulatory intentions. A year later, in 2021, the proposed Artificial Intelligence Act was released. The proposed Act, which is expected to enter into force after 2025, regulates the use of AI across sectors. Its regulatory scope, application and other details are still being debated and amended by the European Parliament as of 2022.

In this white paper, we explore how the debate on the risks of AI was shaped; why it is important to have a balanced and realistic conversation, and that is is crucial to have an inclusive debate that does not solely focus on the risks of AI. In order to do so, different actors should be equipped with a sufficient level of knowledge. This knowledge distribution is also essential to be able to adequately respond to incidents or scandals and enables society to understand the role of AI in these incidents. We elaborate on how AI experts have experienced the assessment of risks of AI and what lessons can be learned. These lessons will be presented in eight inspirations and that can help Quantum Delta NL to study and facilitate the societal impact of quantum technology.

2 Lessons from AI risk management

#1 Invest in human-machine interaction: support early research on the role of human control

Over the last 10 years, the discourse about technology gradually shifted from a techno-centric perspective into a more human-centric approach. The techno-centric perspective starts with the idea that a machine holds all intelligence. Thus, in the case of a technological malfunction, the only solution is to improve the machine.[1] In this perspective, the human is perceived as an object that can be removed through automatisation. This idea stems from 20th century developments in factories, where machines were used for repetitive tasks, such as tightening bolts. In contrast, the human-centric approach argues that machines should not have the power to dictate what jobs are left for humans, instead humans should decide which tasks they wish to delegate.[2]

Recent literature discusses the idea of the collective perspective, reasoning from an interaction between humans and technology.[3] Instead of solely relying on the outcomes of technology or humans, we ought to consider the maximisation of good for both parties resulting from their interaction.[4] For example, in the case of crime prediction models deployed by law enforcement, we need to rethink to what extent police officers should act according to the outcomes of the algorithm that predicts crime hot spots. And in the case of a high-school teacher; how much should they take into account the outcome of an algorithm that advises on a student's knowledge-level?[5] There is room for meeting mid-way, where the techno-centric and the human-centric perspective merge: the collective perspective benefits from unique traits of both technology, such as identifying complex data patterns, and from humans, such as the capacity of critical reflection.[6]

This collective approach also facilitates attention to feedback loops.[7] As a result of the interaction, the human is enabled to provide feedback to update the technology according to environmental changes. During the development phase of technology, it is notoriously difficult to encompass all variables that will be influential in its implementation environment. Therefore, when in usage, there is a need for constant feedback loops to adjust and update the 'machine'. An example of a successful human-technology interaction based on feedback loops is the development of navigational systems.[8] These systems have exponentially increased in quality due to the interaction of humans in the form of regular improvements from (human)

users. For instance, an algorithm cannot effectively detect when a road is blocked due to construction work. Only through the feedback of humans the changes in the real-world circled back in the algorithm.

Inspiration #1: Invest in studying the interaction between humans and technology in order to avoid techno-centric engineering.

#2 Address unrealistic beliefs about a technology: demystify both overly positive and negative beliefs

As with many new technologies, unrealistic beliefs are circulating around AI. On the one hand it is portrayed as a technology that will finally solve all the major world problems, such as hunger and climate issues. On the other hand, there are stories about AI taking over the world and oppress people. During the Covid pandemic, it was suggested that people would be administered a chip through vaccination and thus connected to an AI system that would continuously receive information about people.[9]

Various events contributed to the emergence of unrealistic beliefs of AI and its associated risks. Images of dancing robots, the story that robot Sofia would want a child, and two AI systems that would communicate with each other in a language incomprehensible to humans, have helped create fear among the general public. "Successes" such as AI systems winning against grandmasters in chess and in AlphaGo have contributed to creating unrealistic expectations of AI. In addition, the term "artificial intelligence" suggests that AI systems will approach or surpass human intelligence and contributed to the issue of unrealistic beliefs.

The risk with these unrealistic beliefs is that it makes us both overly optimistic and pessimistic in our expectations. An overly optimistic expectation of AI can result in a lack of scrutiny. It can lead to incorrectly implemented systems and, ultimately, to disappointment.[10] An overly pessimistic expectation of AI makes people anxious. If gone unchecked, these perceptions may contribute to lower investments: applications with a broad societal benefit might remain undiscovered or unused. There may be a fear of 'becoming the next child benefit scandal', with a general disapprovement of AI as a whole. Moreover, unrealistic expectations can obscure our view of current issues. We can think about a world where AI drives all vehicles, but that is not currently on the cards. In contrast, we should think about semi-autonomous vehicles and its consequences at this time.

Demystifying unrealistic images of AI is generally done by scientists, media and policy. National policies can aim to create a realistic picture of applications, make sufficient information available, support campaigns, encourage scientists and build algorithm registries. A good example of demystifying by sharing knowledge is the 'national AI course'[11]. This course is aimed at providing a basic level of knowledge to the public.

Inspiration #2: Tackle the risks of unrealistic beliefs about a new technology – by investing in outreach that generates a balanced picture of today's and near-term implications. Stay away from science fiction visions.

#3 Prepare for incidents and scandals

All over the world, perceptions about the risks of AI emerged as a result from incidents and scandals. For example, the first fatality caused by a self-driving car sparked a major debate.[12] Cambridge Analytica created a lot of attention for the risks of AI, driven by its impact on the US elections in 2016.[13] The rise of Google Glass and deep fakes have also resulted in negative perceptions; and we have seen campaigns around killer robots or the 'Open Letter on AI' signed by influential individuals like Stephen Hawking and Elon Musk contribute to a fear of the risks of AI. On a European level, the child benefit scandal in the Netherlands is one of the most well-known incidents that has defined the debate on the risks of deploying AI.[14] In addition, various risk classification models, such as SyRI in the Netherlands, have had a negative impact on the debate.[15]

The impact of such scandals and incidents is significant, noting that it is not exclusively negative. For instance, citizen unrest over facial recognition in US cities led to a ban on it. In China, the victory of AI in AlphaGo resulted in a draft regulation on AI within a year. In other words, scandals and incidents can highlight the need for policy and regulation. At the same time, it is important that we adequately respond to and learn from incidents. Therefore, a feedback mechanism is essential. It should be made as easy as possible to let the feedback flow back into the system. Not only on the level of user experience as described earlier with navigation systems, but it should also be incorporated on a systemic level, i.e. into regulation. It is important that we invest time and money on how to implement the regulation instead of focusing too much on its enforcement. Providing best practices can help AI providers with integrating better solutions into their prototypes.

Inspiration #3 Incidents will happen. Make sure to prepare adequate responses to possible incidents, and create feedback mechanisms that can feed into (future) regulation.

#4 Be clear about the exact role of AI

When looking at the (long) list of incidents and scandals with AI, the role of AI in these scandals is oftentimes rather limited. It is important that to make a distinction between the overall system and the specific function of the AI in such a system. The child benefit scandal is one of the most well-known incidents in Europe in recent years. This scandal has a great impact on the affected individuals. Many journals, papers and articles have been dedicated to this topic. Many of these point to AI as the main cause of this scandal.

A closer look reveals a more nuanced picture. An algorithm was indeed used to do a risk assessment on potential fraud with childcare benefits. This algorithm turned out to be discriminating, as information about nationalities was used in an unrightful manner. Based on an AI-fuelled recommendation, a desk officer then reviewed this risk score and made a judgement on whether an application was fraudulent or more information should be requested. These civil servants had no (or at least inadequate) information on why an application was classified by the AI as high-risk.

Clearly, AI played a role in setting the child benefit scandal in motion. However, there were several occasions in which the impact of this discriminatory algorithm could have been reduced drastically. The civil servants handling these applications were always in the loop. Furthermore, perverse incentives pushed the tax authority into overdrive due to financial constraints (the department had to generate funding). Judges ruled unlawfully, and politicians only realised the impact at a later stage. It is important to clarify how specific risks can be avoided. For example, this algorithm used nationality as a variable to make a classification while this is discriminatory and unlawful.[16] This could have easily been avoided by excluding this variable in the algorithm.

Inspiration #4: Be nuanced when explaining on the role of AI in incidents and scandals. Make sure to highlight what can be (and should have been) done to avoid a specific problem to persist.

#5 Ensure (technical) knowledge for public policy-makers

Politicians and lawmakers play a key role in the development of new emerging technologies. It is therefore important that they are in a position to take decisions and are equipped with a sufficient level of (technical) knowledge.

This is particularly true when we think about major incidents and scandals, which have a great impact on the perception, policy, and regulation of AI. More than once, scandals set the political agenda and trigger regulations. Debates and regulations arising as a reaction to scandals tend to focus too much on the negative aspects, and can cause regulation that hinders technological innovation. It is important that politicians and lawmakers bring realism into the debate, and distinguish between unrealistic fears, and (future) possibilities. It is a fine balance to strike. More in-depth technical knowledge also reduces the potential influence of large corporate organisations on regulation. We have seen with large corporate organisations have come to dominate the AI industry, while the technical knowledge gap between the public and the private sector keeps growing. This created a lot of room for market players to voice their concerns and to stage campaigns geared towards the interests of large corporate organisations.[17] This is not to say that corporate viewpoints should be excluded, however in order to have a balanced debate and law-making process, it is crucial that politicians and lawmakers can sufficiently challenge the views of private, large (and resourceful) businesses. Recent years of AI development have shown that victims of problematic AI solutions are, for several reasons, not capable to champion their own rights.[18]

Inspiration #5: Support politicians and lawmakers to gain a sufficient level of technical in order to challenge inappropriate lobby efforts.

#6 Address technical gatekeeping at the start

The debate on (the risks of) AI is subject to gatekeeping.[19] For a long time, the debate on AI was controlled and viewed strictly from the technical perspective by engineers, data scientists and other technicians. This did not do justice to the fact that AI is a societal issue. The recent opening of several ELSA labs in the Netherlands is exemplary for this rising awareness of AI being a societal issue. Instead of having a solely technical debate, care should be taken for digital inclusion. People from different disciplines and corners of society should be involved at an early stage. Until recent years, the research concerning AI was mostly done from a technical point of view, which resulted in a strict technical

perspective on AI, neglecting the societal impact of the technology. For example, a hospital developed an algorithm to feed medical advice with technical teams, without adequate input from the (future) patients. Not only does gatekeeping lead to a limited perspective on AI, it also results in a concentration of knowledge among a relatively small group of engineers and data scientist. This had an excluding effect on many non-technical stakeholders.[20]

It is therefore important to include the perspective of the (end-)user and society. Organisations can actively seek for end-user opinions through citizen participation. You involve people in the process and get a better idea of what the end user actually needs. Then, ideally, products are developed in line with user demands, instead of the users having to be convinced that they need it after development.

In order to enable effective citizen participation, it is essential to have a sufficient level of digital literacy. At the same time, it is also important to include different perspectives in education. For example, technical education should integrate the societal aspects from day one. And currently, only higher education offers courses on digital inclusion. Luckily, this is slowly changing, as there is more room for ethical issues.[21]

Inspiration #6: Improve educational offerings so that technicians learn about ethics and technology awareness grows among those with a non-technical background

#7 Be a Competence Centre – double down on a specific strength

In order to make good use of innovations, a trustworthy data infrastructure is essential. Next to this infrastructure, it is evenly important to have 'local' knowledge. For example, the Netherlands has successfully worked on an environment that enables high quality hardware, software and guidelines, creating a well-founded data infrastructure.[22] Yet, the level of 'locally produced' knowledge in the field of Artificial Intelligence (AI) still lags behind.[23] The rate of knowledge production has disadvantaged the Dutch ranking in the global AI landscape, which has left room for big tech companies to become dominant in this market. To avoid the risk of big tech having too much authority, it is worth becoming a center of excellence for specific themes around an emerging technology. For small and medium-sized countries it will be worthwhile to focus on specific themes within an emerging technology as it is impossible to develop and maintain a

knowledge advantage on a wide range of themes.
Changing market positions is difficult, and especially when the market deals with dominant players. For the Netherlands to reduce their AI-knowledge backlog is a problematic proposition because their well-resourced competitors are continuously developing.[24] This task becomes even harder when considering the amount of university scientists that big tech companies can lure away.[25] Therefore, it is crucial to start early on with knowledge development, in order to maintain an advantage. This can be done by investing in 'local knowledge production' and stimulate the cooperation and knowledge-sharing between universities. In this process, an emphasis ought to be placed on an interdisciplinary approach, where both technical and social scientists collaborate on responsible innovations.[26]

Inspiration #7: Invest in local knowledge production in a specific field of technology regulation, which can become a widely known centre of excellence – to attract and keep talent.

#8 Make sure regulation fits the ecosystem of the technology

The April 2022 proposal for AI regulation (AI Act) by the European Commission aims to implement an ecosystem of trust by proposing a legal framework for "trustworthy AI". While technology such as AI is widely tested and undergoes many iterations, sweeping legislation such as this proposal must reflect the way technology is being brought to the market place. In the case of the AI Act, this is not the case. These discrepancies could have been avoided if the AI Act had been tested in practice.

The AI Act follows a product regulation approach. That means that an AI system will be placed on the market by a provider (or importer or distributer) after which it is used by a user. Consequently, provider and user are the two main actors in the AI Act and these roles come with different obligations. Research on the definitions of these actors shows that definitions are clear on paper but are not mutually exclusive in practice.[28] For example, a company that develops AI systems and place them on the market is a provider. It could be very well possible that this company also uses these in-house built AI systems, which makes the company a user as well.

This simplified taxonomy of actors in the AI Act makes it hard to distribute responsibility across the AI value chain. For example, a hospital uses an AI system for improving the diagnostic process. This AI system is provided by

company ‘A’, which used general purpose AI made available by company ‘B’. The training data are provided by company ‘C’. It turns out that the AI system used by the hospital is biased. As the AI Act does not describe the companies A, B and C as an actor it is difficult to distribute responsibilities in a rightful manner.

Inspiration #8: Spend enough resources on testing legislation in practice before it comes into force, so that implementation works for technology providers and -users.

3 Conclusions

AI communication has received much attention in recent years, with many stakeholders becoming increasingly involved in reinforcing public understanding of AI technologies. This effort resulted in a shifting public understanding of AI – from fears over gloom-and-doom scenarios to a more realistic perception about the social and economic impacts of AI applications. Still, much remains to be done and AI communication is in constant change, with recent efforts focusing on talking with the public instead of talking to the public.

Looking at the current challenges in AI communication, it becomes clear that efforts on future technologies such as quantum should start as soon as possible, adopt the best practices from AI, and should start with addressing potential pitfalls in public understanding today. At Quantum Delta NL we are committed to investing in best practice sharing. We believe that we should learn from the past, and take inspiration from those who have taken similar pathways before. The nine inspirations presented here have been drawn from a select number of expert interviews and are merely a beginning, a conversation starter. We hope they can serve as a starting point for a fruitful discussion within the quantum communication community.

At the same time, we are investing in concrete follow-ups. Where inspirations lead to concrete suggestions on how to do things in quantum, we will take these up. For example, we have started working on impact assessments back in 2021, along the lines of AI impact assessments. We are also working on public awareness campaigns and a quantum course, for free and without any jargon. Both are available from 2023 onwards. Acknowledging that there is much work to be done in communicating quantum, we welcome any support along the way from our partners in the growing quantum ecosystem.

4 Endnotes

[1] Interview AI research scientist
[2] Interview with responsible AI researcher
[3] Peeters, M. M. M., van Diggelen, J., van den Bosch, K., Bronkhorst, A., Neerincx, M. A., Schraagen, J. M., & Raaijmakers, S. (2021). Hybrid collective intelligence in a human–AI society. *AI & society*, *36*(1), 217-238. https://doi.org/10.1007/s00146-020-01005-y. See also: Malone TW (2018) How human-computer 'Superminds' are redefining the future of work. MIT Sloan Manag Rev 59(4):33–42.
[4] Interview with a professor in technologies
[5] Interview with a professor in technologies
[6] Interview AI research scientist
[7] Interview with tech entrepreneur
[8] Interview with CEO of responsible AI company
[9] Interview with a professor in technologies
[10] Idem
[11] https://www.ai-cursus.nl
[12] https://www.forbes.com/sites/davidkiley5/2018/03/19/the-first-pedestrian-fatality-with-an-autonomous-vehicle-could-set-tone-for-lawyers-and-liability/
[13] https://www.theguardian.com/news/2018/may/06/cambridge-analytica-how-turn-clicks-into-votes-christopher-wylie
[14] https://www.politico.eu/article/dutch-scandal-serves-as-a-warning-for-europe-over-risks-of-using-algorithms/
[15] van Bekkum, M., & Borgesius, F. Z. (2021). Digital welfare fraud detection and the Dutch SyRI judgment. European Journal of Social Security, 23(4), 323–340. https://doi.org/10.1177/13882627211031257
[16] https://autoriteitpersoonsgegevens.nl/nl/nieuws/werkwijze-belastingdienst-strijd-met-de-wet-en-discriminerend
[17] Interview with responsible AI researcher
[18] Idem
[19] Idem
[20] Interview with a professor in technologies
[21] Interview with responsible AI researcher
[22] Interview with tech entrepreneur
[23] Interview with CEO of responsible AI company
[24] Interview with a professor in technologies
[25] Interview with CEO of privacy company
[26] Interview with responsible AI researcher
[27] https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex%3A52021PC0206
[28] https://openloop.org/programs/open-loop-eu-ai-act-program/

Quantum Delta

Lorentzweg 1

2628 CJ Delft

E info@quantumdelta.nl

Quantum Delta
Nederland