WHITEPAPER

AI: IN SEARCH OF THE HUMAN DIMENSION

INVOLVE CITIZENS AND EXPERIMENT RESPONSIBLY

TNO innovation for life

INDEX

ABSTRACT

Artificial intelligence, or AI, has resulted in many innovations in recent years, prompting government organisations to consider how they too can use AI to address social issues or to improve their performance. AI can, for example, be used in algorithmic decision-making. There are also drawbacks to using AI in this way, however, one being the risk of discrimination. To guard against such adverse effects, EU and national policymakers aim to establish ethical frameworks for AI development and to regulate the use of AI. In addition to that TNO is developing a methodology for testing AI systems for algorithmic decision-making, known as the Dynamic Impact Assessment. This system will make it possible to analyse the shorter and longer-term effects of an AI application and to consider the interests of the various stakeholders, including the general public.

1. INTRODUCTION

Expectations regarding the innovativeness of artificial intelligence are running high. In its *Strategic Action Plan for AI*, the Dutch government states that AI will 'make a substantial contribution to economic growth, prosperity and well-being of the Netherlands.'[1] Although AI has been around as a technology since the 1950s, the rise of big data and advances in computing power have created an unprecedented boom. AI is expected to deliver breakthroughs in medical research and lead to greater road safety thanks to (partially) self-driving vehicles. Inspired by these examples, governments are also increasingly using AI to explore social issues and take decisions, for example to expedite the processing of visa applications, to gain a better understanding of poverty or debt issues, or to maintain bridges and locks. This is referred to as *algorithmic decision-making*, in which the AI algorithm determines some or all of the output by automated means.[2]

In these examples, technology has a direct or indirect impact on people's lives, for example whether they qualify for debt assistance or are given a fine. This can sometimes have unintended adverse consequences (see box).[3]

1 Strategic Action Plan for Artificial Intelligence (2019),
2 This position paper focuses on the use of AI in algorithmic decision-making in the public sector and therefore does not address its use in self-driving vehicles, robots and other autonomous systems, for example. TNO and others are studying the safety of such AI systems and how to ensure that they comply with ethical standards. See, for example, Aliman, N.M., Kester, L., Werkhoven, P., & Ziesche, S. (2019). Sustainable AI safety? Delphi, 2, 226.
3 Sources: Guide to AS and A level results for England, 2020 - GOV.UK (www.gov.uk); Why did the A-level algorithm say no? - BBC News.

Using an algorithm to score final examinations
In the United Kingdom, pupils' A-level scores determine the likelihood of admission to the university of their choice. When the Covid-19 pandemic prevented the authorities from administering A-levels, an algorithm was used to predict pupils' results based on a combination of their individual ability and how well their school performed in exams in recent years. This meant that bright children in low-achieving schools were disadvantaged, while pupils enrolled in public schools – which are often smaller and where individual pupils receive more attention – had a leg up. The use of the algorithm to predict A-level scores was heavily criticised as a result and led to protests.

It is not only in the United Kingdom that the use of algorithms by government has led to a public outcry. The question, then, is how to prevent unwelcome consequences. Government's role in this is twofold. First of all, it must act responsibly itself in its use of algorithms. Second, it must consider which existing laws can ensure the responsible use of algorithms and develop new policy where needed. One important piece of legislation is the European Union's General Data Protection Regulation (GDPR), designed to guarantee data privacy. In addition, ethical frameworks for AI development are being drafted that define principles for the responsible use of AI, based on human rights.[4] A good example is the set of seven requirements for trustworthy AI produced by the European Union's High-Level Expert Group on AI (AI HLEG).[5] The European Commission recently published a *Proposal for a Regulation laying down harmonised rules on artificial intelligence*, based in part on these requirements.[6] The proposal suggests regulating high-risk applications by requiring that they demonstrate in advance that they comply with certain requirements, for example human oversight. It does not make a distinction between the private and the public sector in the use of AI technology. It also allows for AI innovation by noting the importance of experimentation.

Based on a survey of current regulatory systems for AI and an analysis of existing ethical frameworks and methods, this paper introduces a methodology for testing the effects of AI, i.e. a *Dynamic Impact Assessment* for AI systems focused on social issues.

The paper begins by describing the use of AI in algorithmic decision-making and the associated risks. It then surveys current policies governing AI and goes on to analyse ethical frameworks and methodologies for developing value-driven technology. Finally, it introduces TNO's methodology for testing responsible AI systems for algorithmic decision-making.

4 Kamerbrief over artificiële intelligentie, publieke waarden en mensenrechten | Kamerstuk | Rijksoverheid.nl
5 European Commission High-level Expert Group on AI, Shaping Europe's digital future (europa.eu).
6 European Commission (2021), Proposal for a Regulation laying down harmonised rules on artificial intelligence | Shaping Europe's digital future (europa.eu).

2. ALGORITHMIC DECISION-MAKING

Algorithms have played a role in decision-making – including government decisions – since the advent of the computer in the 1950s. Algorithms can identify who is liable to pay tax and who is not, for example. Such algorithmic decisions are underpinned by a set of well-defined statutory rules: someone who earns income from employment, for instance, is liable for income tax. This rule can then be used to determine, without human intervention, whether someone will be sent an automated letter summoning them to file a tax return. The algorithms used for this purpose have been refined over time as the technology has improved.

A recent TNO study shows that the use of AI in public services has increased in the past two years.[7] The Netherlands Court of Audit (NCA) recently investigated government use of *predictive* and *prescriptive* (AI) algorithms.[8] An example of the former is an AI-enabled model that uses sensor data to predict when a bridge or lock is due for maintenance or replacement. An example of a prescriptive algorithm is the one given above, in which an automated system determines who is liable to pay tax and thus receives a letter summoning them to file a tax return.

A mistake in either example could be inconvenient or annoying, for example if a bridge is closed unnecessarily or someone is mistakenly classified as being liable for tax. In its audit, the NCA found that most of the algorithms currently in use by the Dutch government are relatively simple. Even so, even straightforward decisions or decisions that are only partly automated can have a major impact on individuals, for example when AI is used to score final exams, and can also lead to unwanted effects, such as inequality. The question then is whether it is appropriate to use AI in algorithmic decision-making and whether its advantages outweigh its adverse consequences.

Research by the Rathenau Instituut,[9] the NCA and TNO highlights the risks and adverse consequences of algorithmic decision-making. There are several reasons why this should be a concern. One is that bias in data or in an algorithm may result in the unequal treatment of individuals or groups. An algorithm can also reflect the bias of its programmer, who – unlike the customary decision-makers – may not have much actual knowledge and experience of how the algorithm will be used. When it comes to social issues, however, it is especially critical to consider the interests of those affected by a decision and to weigh them up against those of the decision-makers.

Table 1 lists the adverse consequences of algorithmic decision-making based on AI systems. Some of these consequences are not exclusive to AI but are also associated with the large-scale use of personal data or with the broader application of algorithms. AI can be a contributing factor in both cases. In addition, AI-driven algorithms have specific adverse consequences, especially when applied to machine learning in decision-making processes.

7 TNO (2021), Quickscan AI in publieke dienstverlening II | Rapport | Rijksoverheid.nl.
8 Netherlands Court of Audit (2021), Understanding algorithms | Report | Netherlands Court of Audit.
9 Rathenau Instituut (2020), Artificial Intelligence, what's new? | Rathenau Instituut.

Table 1: Adverse consequences of algorithmic decision-making and potential solutions

Adverse consequences	Cause; origin	Solutions (actual or potential)
Unwanted or unintended impact of predictions; discrimination or exclusion Systemic inequality	Incomplete, incorrect or biased data Misclassification of cases based on self-learning or other algorithms	Representative and high-quality datasets Redress mechanisms such as complaints procedures, democratic oversight, litigation, protests
Confusion about accountability of decisions or of systems that take decisions	The algorithm or system lacks transparency, and the processes and outputs of algorithmic decision-making are difficult to explain	Transparency, e.g. through algorithm registers; explaining and accounting for the workings of algorithms and systems
Dehumanisation	Bureaucratisation of decision-making ('computer says no')	'Meaningful human control', monitoring and oversight

One of the main adverse consequences of automated decision-making is discrimination and exclusion owing to incomplete, incorrect or biased data. When developing an AI system, it is important, first and foremost, to have enough useful data from which the system can 'learn'. This is especially the case for supervised/unsupervised machine learning. A meagre dataset makes it difficult to use an algorithm for problem-solving. In addition to data quantity, data accuracy and bias play a major role. In the latter case, the dataset used to train the algorithm may not be representative of the population in which the AI system will be used, or it may be representative but in so being deepen social inequality. One example is the COMPAS tool in the United States (see box).[10] The same adverse effect may also result from an algorithm that has learned faulty patterns or assessments. That is because autonomous algorithms can detect random anomalies and label them, erroneously, as patterns, leading to the misclassification or exclusion of individuals.

10 Jeff Larson, Surya Mattu, Lauren Kirchner and Julia Angwin (2016). How We Analyzed the COMPAS Recidivism Algorithm. ProPublica.

COMPAS in the United States
The Correctional Offender Management Profiling for Alternative Sanctions or COMPAS is an algorithm-driven tool used by the US courts to assess potential recidivism risk, a factor in sentencing. The dataset used to train this algorithm consisted mainly of Black men, a group that historically has received higher sentences than other groups for the same offences in the US judicial system. As a result, the algorithm greatly accelerated this effect. This happened without any human intervention.

The second adverse consequence concerns the explicability and limited accountability of decisions or of the systems that produce them, something that concerns our understanding of both the algorithm itself and its outputs. Both the NCA and TNO have concluded that in many instances in which government takes decisions about citizens, it outsources data processing and the algorithm itself to commercial parties. This often means that the precise design of the algorithm is regarded as the property of that external party and that insight into its precise operation is thus not always forthcoming, making it more difficult for a government organisation to exercise oversight over and be accountable for an algorithmic system.[11]

Dehumanisation is the third adverse consequence of algorithmic decision-making. This phenomenon occurs when a system produces an output without taking the human situation into account. Even when there is human agency in decision-making, the decision-makers are likely to adopt the system's outputs almost automatically, making human intervention in the system difficult and also making it hard for people to get an explanation for the decision. As a result, not only do they sometimes face a disadvantageous (and incorrect) outcome, but they are also denied an intelligible answer regarding the reason for this decision.

The adverse consequences described above illustrate the impact that automated decision-making can have on individuals. Such consequences could well lead to a crisis of trust in algorithmic decision-making or even in government as a whole. The protests against using an algorithm to score A-levels in the UK are just one example. Individual officials and policymakers may also be biased, leading to systemic discrimination and exclusion. There are democratic oversight mechanisms meant to minimise bias, but the use of algorithmic decision-making puts pressure on such checks and balances. To prevent algorithms from leading to systemic discrimination, inequality and other adverse consequences, the EU and national governments, including the Netherlands, are developing sets of rules and policies to govern the use of AI.

11 And this should prompt a review of the public procurement process for AI systems. See Van Noordt, C., Misuraca, G., Mortati, M., Rizzo, F. and Timan, T., (2020). *AI Watch - Artificial Intelligence for the public sector*, Publications Office of the European Union, Luxembourg.

3. EU POLICY ON AI

The main objective of the EU's policy on AI is to encourage the development and use of AI along two axes: trust and excellence.[12] After all, the technology is regarded as a key driver of innovation and economic growth. The EU does not wish to lag behind the huge investments being made elsewhere in the world, such as in the United States, where most of the big software and data companies are located, and in China, where the state is investing heavily in AI development. By introducing rules that aim to limit the influence of foreign platforms[13] and policies encouraging European AI start-ups and SMEs, the EU is attempting to gain better control of the AI innovation landscape.

But there is another reason to stimulate AI development in the Union: ascertaining whether American or Chinese technology adheres to 'European values' – human dignity, equal treatment, the right to privacy, information security – is even more difficult. The EU's second objective, then, is to ensure the responsible use of AI. This means regulating the development and deployment of European AI systems to ensure that it is done responsibly. Table 2 lists the main European AI policies, all of which focus on both encouraging and regulating AI.

12 European Commission (2021), Proposal for a Regulation laying down harmonised rules on artificial intelligence | Shaping Europe's digital future (europa.eu).
13 European Commission (2020), Digital Markets Act: Ensuring fair and open digital markets (europa.eu).

Table 2: The EU's AI strategy

EU AI strategy	Encouraging development of the technology	Regulating responsible use
COMMUNICATION FROM THE COMMISSION Artificial intelligence for Europe {COM(2018) 237 final}	– Boost the EU's technological and industrial capacity and AI uptake by the private and public sectors – Prepare for socio-economic changes brought about by AI	– Ensure an appropriate ethical and legal framework, based on the Union's values and in line with the Charter of Fundamental Rights of the EU
WHITE PAPER on Artificial Intelligence - A European approach to excellence and trust {COM(2020) 65 final}	– Organise innovation by establishing a network of Digital Innovation Hubs – Secure access to data for AI – Ensure data compliance with FAIR principles	– Regulate high-risk AI applications – Set up voluntary labelling and certification for no-high risk AI applications – Establish networks of national authorities
Report on the safety and liability implications of Artificial Intelligence, the Internet of Things and robotics {COM(2020) 64 final}	– Extra risk assessment procedure for products subject to important changes during their lifetime (e.g. due to new software, algorithms or data)	– Explicit obligations for producers to consider the immaterial harm their products could cause to vulnerable users – Requirements regarding transparency, robustness, accountability and human oversight of algorithms – Obligations on developers of algorithms to disclose design parameters and metadata in the event of accidents
Proposal for a REGULATION LAYING DOWN HARMONISED RULES ON ARTIFICIAL INTELLIGENCE (ARTIFICIAL INTELLIGENCE ACT) (COM(2021) 206 final)	– Community of excellence – National 'regulatory sandboxes' – Give small-scale providers and start-ups priority access to AI regulatory sandboxes	– Community of trust – Regulation of high-risk AI systems – EU and national oversight of high-risk AI systems

One important component of the EU's strategy of ensuring a values-driven approach to AI is to develop ethical frameworks for responsible AI. To define the underlying principles and requirements, the European Commission set up an independent High-Level Expert Group on AI (AI HLEG) (see box).[14] In addition, many EU Member States and parties in the industry have developed their own ethical frameworks in an effort to support the responsible development of AI. Because ethical frameworks consider the whole system from the vantage point of human rights, they can provide a better understanding of self-learning systems that take decisions (in part) without human intervention.

Established under the auspices of the European Commission, the AI HLEG has drawn up a set of ethical guidelines consisting of seven requirements for trustworthy AI. The seven requirements are based on four ethical principles: respect for human autonomy, prevention of harm, fairness, and explicability. The seven requirements are:

1. human agency and oversight;
2. technical robustness and safety;
3. privacy and data governance;
4. transparency;
5. diversity, non-discrimination and fairness;
6. environmental and societal well-being;
7. accountability.

An analysis of fifteen commonly used ethical frameworks (see box), however, shows that they incorporate many different ethical standards but do not clarify who decides how the standards are applied. They also fail to explain how choices can be made between conflicting values. That is why these frameworks are generally better suited to an ex-post assessment of the impact of AI systems. Another problem is that AI developers often find ethical frameworks abstract and difficult to apply when they are developing systems. In other words, important questions remain. Which applications should be subject to these principles? Who should apply the principles? How do we deal with conflicting values and dilemmas? At what point in its development should an AI system be tested against these principles?

14 European Commission High-Level Expert Group on AI, Shaping Europe's digital future (europa.eu).

Fifteen ethical frameworks considered in our analysis:

1. AI Guidelines – Deutsche Telekom
2. Everyday Ethics for Artificial Intelligence – IBM
3. Principles for Accountable Algorithms and a Social Impact Statement for Algorithms – Fairness, Accountability and Transparency in Machine Learning (FATML)
4. Artificial Intelligence and Machine Learning: Policy Paper – Internet Society (ISOC)
5. Ethically aligned design – Institute of Electrical and Electronics Engineering (IEEE)
6. ITI AI Policy Principles – Information Technology Industry Council (ITI)
7. Top 10 Principles for Ethical Artificial Intelligence – UNI Global Union
8. Recommendation of the Council on Artificial Intelligence – OECD
9. Charlevoix Common Vision for the Future of Artificial Intelligence – Leaders of the G7
10. Montréal Declaration for Responsible Development of AI – Université de Montréal
11. AI in the UK: ready, willing and able? – UK House of Lords, Select Committee on AI
12. An Ethical Framework for a Good AI Society – Floridi et al.
13. Preparing for the future of Artificial Intelligence – US National Science and Technology Council, Committee on Technology
14. How can humans keep the upper hand? Report on ethical matters raised by AI algorithms – French Data Protection Authority (CNIL)
15. Automated and Connected Driving: Report – Federal Ministry of Transport and Digital Infrastructure, Ethics Commission

The answers often involve combining ethical frameworks with specific methodologies aimed at designing or testing AI systems, such as the guidance ethics approach, 'by-design' methods, or impact assessments. The guidance ethics approach addresses the question of how to develop and use technology by engaging in an open dialogue with the relevant actors about the possible effects and values at play.[15] The developers then incorporate the results of this dialogue into the technology. Then there are several 'by-design' methods that focus on how we can incorporate or 'build in' values when designing a new digital service or application. They include 'privacy-by-design',[16] which emphasises data protection and cryptography, and 'value-sensitive-design',[17] which stresses human rights and public values and tests these in the design and development process.

15 Verbeek, P-P. & Tijink, D. (2019). Guidance ethics approach. An ethical dialogue about technology with perspective on actions. ECP | Platform voor de InformatieSamenleving.
16 TNO (2021), Privacy by design: data combineren voor betere overheidsdienstverlening | TNO.
17 Van de Poel, I. (2013). Translating values into design requirements. In Philosophy and engineering: Reflections on practice, principles and process (pp. 253-266). Springer, Dordrecht.

Impact assessments are used to determine impacts or to assess risks. The actual impact of using a technology is usually assessed ex-post, for example using the audit framework developed by the NCA,[18] but there are also ex-ante risk impact assessments. Some assessments are geared towards identifying the risks of using a specific technology,[19] while others focus on the risks that use might pose to a fundamental right. For example, the GDPR requires a data protection impact assessment (DPIA) when data are processed on a large scale and when processing is likely to result in a high risk for individuals.[20]

The question is whether using such methodologies to apply an ethical framework will prevent the adverse consequences identified. While the above tools do test whether new technologies are responsible, three challenges remain when it comes to their usefulness in the specific case of AI-driven algorithmic decision-making:

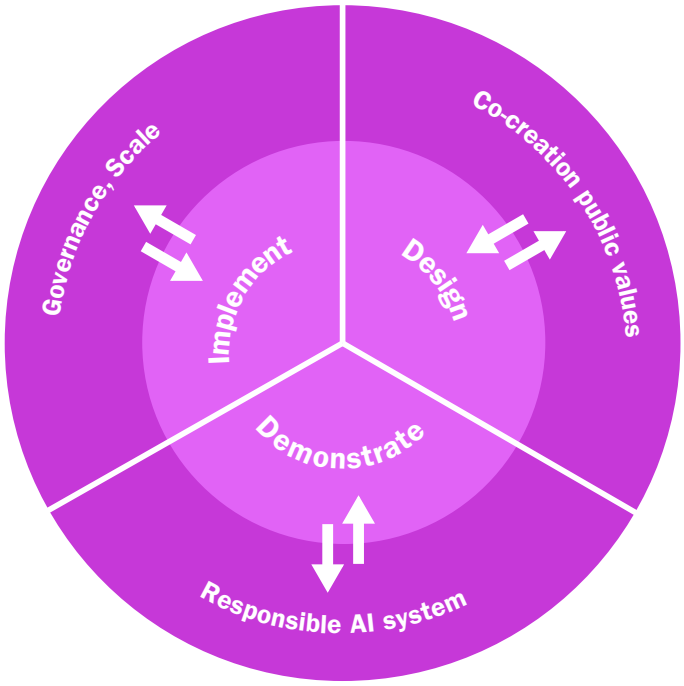
1. Many ethical frameworks address AI developers rather than AI users and developers tend to be commercial parties that are only involved in certain parts of the decision-making process. Ultimately, the responsibility for the system often lies with the parties that use the AI systems in their automated decision-making. Ensuring that the outputs of algorithmic decision-making are transparent would require them to make organisational and process-related changes, and that means considering the entire system in which AI is applied.
2. Explicability of algorithms and involving parties in the use of AI. As described above, the adverse consequences of AI for individuals and society can be profound. Under the GDPR, data processors must notify individuals that their data are being processed. This is all the more important in the case of self-learning systems that learn patterns or make assessments themselves, and whose operations are difficult to explain. In the case of AI, the rules must therefore enforce transparency about who is using the data to train algorithmic models and in what context the outcomes will impact both the user and the individual.
3. Much remains unclear about the long-term consequences of algorithmic decision-making. For instance, what are the secondary effects? As we saw in the example of the algorithm used to score A-levels in the UK, one effect might be a widening education gap between groups of pupils. The question is whether these effects will be revealed in an ex-ante impact assessment. More long-term and dynamic oversight and monitoring of AI systems is therefore advisable.

Ethical frameworks, especially those linked to the guidance ethics approach, values-sensitive design methodology or impact assessments, can thus support the development of responsible AI. Even so, we need to consider (1) the whole system in which AI is used, 92) involving the various stakeholders more closely, including the public, so that their interests are represented, and (3) using the frameworks more dynamically.[21]

18 Netherlands Court of Audit (2021), Understanding algorithms | Report | Netherlands Court of Audit.
19 ECP (2018), Artificial Intelligence Impact Assessment
20 Autoriteit Persoonsgegevens (Dutch Data Protection Authority), Data protection impact assessment (DPIA) | Autoriteit Persoonsgegevens.
21 TNO is working on the above-mentioned challenges in the AI Oversight Lab (www.appl-ai-tno.nl)

4. TESTING GROUND FOR AI SYSTEMS

Our conclusion is that ethical frameworks meant to protect public values must not only provide guidance when an AI system is under development but should also be in place when that system is in use, based on the interests of all the stakeholders. To ensure that this happens, AI in algorithmic decision-making should initially be trialled in an experimental environment, a testing ground for responsible AI based on ethical frameworks and impact assessments. Assessing AI in such an environment makes it possible to identify in time the risks arising in the different stages, from data generation to data collection, data processing and AI modelling all the way up to the output stage. A testing ground is also ideal for inviting input from different stakeholders, including the public. One example of a methodology used to test responsible AI systems in cooperation with partners is TNO's Dynamic Impact Assessment (see box). This methodology consists of three phases with key questions that are addressed in three different stages of AI-driven algorithmic decision-making.



Dynamic impact assessment methodology for responsible algorithmic decision-making

The dynamic impact assessment methodology consists of three phases: (1) involve stakeholders and the public to ensure that the design reflects differing interests (design), (2) test whether the systems are responsible (demonstrate), and (3) use AI systems for algorithmic decision-making (implement).

Design:

1. System analysis: **explore** the system in which the AI would be used and, together with the stakeholders involved, capture this system in a common model that gives rise to specific sub-questions to be answered by data. *Key questions:* What are the social issues that the AI application is meant to address? What legitimate interest is being served by using data and AI?
2. Experimental environment: if AI offers an appropriate solution, then **set up** a controlled environment for experiments. This makes it possible to identify unwanted adverse consequences early on and to take appropriate action. *Key questions:* Which stakeholders are needed to design an AI application? Which legislation applies? Which of the ethical frameworks is appropriate?

Demonstrate:

3. Co-creation and explicability: **consider the vested interests** of the public and other stakeholders in using AI algorithms in a system. *Key questions:* Is the public sufficiently aware of how the system works and what impact it can have on their lives? Who is the creator, owner and/or operator of the AI system?
4. **Adapt and use AI** systems for algorithmic decision-making responsibly. *Key questions:* Are the datasets correct or biased or do the algorithms discriminate or exclude? What do transparency and explicability mean in the context of this policy item or decision-making process? How is human oversight organised?

Implement:

5. Testing of responsibility: **test** whether the system is a responsible one, including whether it is explicable. *Key questions:* What method is suitable for testing the system? How can we explain how AI systems work to people who are impacted by algorithmic decision-making?
6. (Long-term) impact: **identify and monitor** systemic effects that may result from using AI systems. *Key questions:* How should such systems be monitored to prevent systemic effects from occurring, and by whom? What proven technological and/or organisational measures can be used for this purpose?

This methodology calls for a transdisciplinary approach that goes beyond involving AI developers and legal practitioners to incorporate the views of policy officials, regulatory bodies and the public. Public organisations must also be proactive about asking themselves the above questions when using AI in algorithmic decision-making. This is especially true for the high-risk AI systems identified in the European Commission's proposed Regulation laying down harmonised rules on artificial intelligence. An example is described in the box below.[22]

22 Steen, M., Timan, T. & van de Poel, I. (2021) Responsible innovation, anticipation and responsiveness: case studies of algorithms in decision support in justice and security, and an exploration of potential, unintended, undesirable, higher-order effects. *AI Ethics*.

Central Judicial Collection Agency debt relief assistance – secondary and long-term effects of AI

A recent evaluation of how ethical standards from the AI HLEG's ethical framework play out in real-world situations concerns an AI system used by the Netherlands' Central Judicial Collection Agency (CJIB) to keep people from falling further into debt. It does this using historical payment behaviour to identify people with outstanding traffic fines who might be at risk. The idea is to enable CJIB agents to intervene promptly and effectively by identifying those who are willing but unable to pay so that a payment arrangement can be made for them. During the design phase, it was decided to make use of a limited but more easily explicable algorithm to make this identification. While the AI system was developed according to 'ethics by design' principles and seemed to work well, its real-world use revealed 'secondary' effects, such as the unfairness of treating some people differently from others. It also generated unforeseen extra work for the agents, who wasted time acting on the AI system's incorrect recommendations at the expense of more meaningful action. This shows that assessing the long-term effects of AI systems should always be part of the monitoring system.

The core of the approach is that the various stakeholders are all involved in using the AI system during the experiment, that oversight of the system's use is dynamic, that the system is modified where necessary, and that its results are communicated in an intelligible way. The (anticipated) risks and impact are reported during implementation, with a standard information leaflet being development that is easy for ordinary people to understand.

5. CONCLUSION

The use of AI in algorithmic decision-making is on the rise but can have unwanted adverse consequences. To ensure that AI behaves responsibly in algorithmic decision-making, policymakers are examining existing and new legislation governing data privacy, product liability and other matters. They are also considering ethical frameworks and values-driven methodologies based on human rights, specifically with respect to high-risk AI systems. It will take more than ex-ante ethical frameworks and ex-post testing to prevent unwanted adverse consequences, however. What is important is to test and monitor such systems in real-world situations in consultation with various parties, including the public. AI algorithms also require dynamic human agency and oversight, including of the long-term effects of AI. This paper's main recommendation is therefore to set up experimental environments in which AI systems for algorithmic decision-making – or at the very least high-risk AI applications – can be tested in a series of steps.

Main authors
Anne Fleur van Veenstra
Tjerk Timan

Additional authors
Gabriela Bodea, Cass Chideock, Ilina Georgieva,
Claudio Lazo, Mathilde Theelen

Contact
Babette Bakker, Strategy and Policy
Location The Hague – New Babylon
babette.bakker@tno.nl
+31 621 137 231

TNO innovation for life

TNO.NL