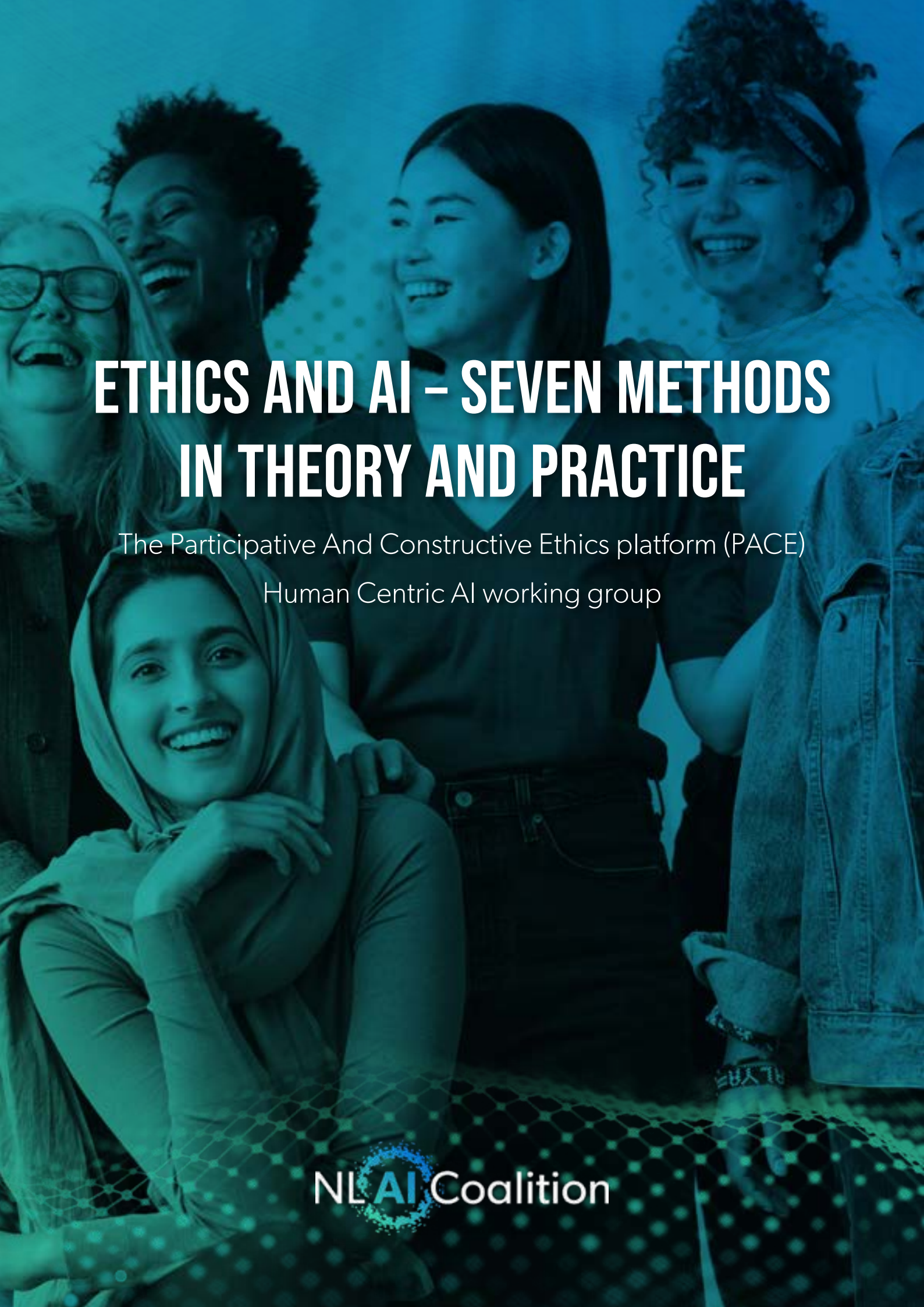

ETHICS AND AI – SEVEN METHODS IN THEORY AND PRACTICE

The Participative And Constructive Ethics platform (PACE)

Human Centric AI working group

NL AI Coalition

TABLE OF CONTENTS

ETHICS AND AI: METHODS IN THEORY AND PRACTICE

OBJECTIVE

This publication covers some practical ethics methods that can be used when implementing technology and algorithms. They are intended in the first instance not only for the AI fields of application and the AI hubs but also for the various ELSA Labs of the Netherlands AI Coalition (NL AIC). We think that these include good approaches for any technique in any context. Our aim is to develop these methods further. We would therefore like to encourage you to use these existing methods and share your experiences with us. The majority of the methods have been tried and tested in practice; they have solid foundations and are well-respected. If we use the same kind of methods for ethics, it will in fact become possible to scale up and iteratively improve these approaches significantly. We can also learn from the effects in various sectors and organisational environments. Using these approaches is in line with the Netherlands' strong tradition of practical technology ethics. We are one of the pioneers in Europe in that regard. This experience will be highly valuable, especially when discussions about abstract frameworks start to matter on the ground in practice. We hope to help you develop and implement AI applications constructively and ethically, and we hope you will help us by sharing your experiences of using these methods to help make them stronger and even better suited to your practice.

BACKGROUND

The Netherlands AI Coalition focuses on human centric AI. This applies to all AI applications and all application areas. A horizontal building block has therefore been defined for this field, collectively giving shape to dialogue about ethics, law and society. ELSA Labs are being set up that rely on this three-pronged approach (Ethical, Legal and Societal Aspects) that are thus also products of that building block. For the ethics part, a platform has been designed in which various organisations from society, science, government and the commercial sector come together. That platform chooses participatory and constructive ethics and has therefore been named PACE (the platform for Participatory And Constructive Ethics). This also deliberately uses the English word pace, in the sense of a step, representing progress and achieving better (i.e. more ethical) uses of AI applications. The focus and activities of PACE have been set out in a position paper[1].

PRACTICAL AND CONSTRUCTIVE APPROACHES TO ETHICS

Various methods have been discussed within the platform for ways of putting ethics and AI into practice. From the perspective of the participants in the Netherlands AI Coalition (NL AIC), there is a need for getting acquainted with some of these practically useful methods. That is why we looked at methods with a certain 'street credibility' – not methods that are scientifically perfect, but in particular methods that are genuinely used in practice. We have maintained that practical angle for this publication as well. It is not an exhaustive overview with in-depth comparisons between the methods. Above all, we are offering a description of some interesting ethics-oriented practices. They differ significantly in various aspects, such as the context in which they are used, what deploying the method requires from an organisation, what results the method produces, and so forth.

SETUP

In this publication, we look successively at seven methods that consider ethics and AI, describing not only the content of each method but also a case study for each that one of the working group's members has written up for the method in question. We conclude with an overview based on various relevant criteria plus a few recommendations, making it clear which method is best to use in which situations. This publication therefore deliberately aims not to describe one method as better than another, but instead to provide an understanding of the pros and cons, based on the idea that several methods can be used side by side within a single organisation, or even applied to the same case. There are also references to contacts and/or websites for more information.

1. https://nlaic.com/wp-content/uploads/2022/08/Position_Paper_PACE_EN_June_2022.pdf

METHOD 1: GUIDANCE ETHICS APPROACH (GEA)

WHAT IS IT?

Guidance ethics is a practically-based approach for holding a constructive dialogue and ensuring that innovations are ethically responsible. The approach has been developed because abstract value frameworks are particularly difficult to convert into practical steps. Both the technology and the application context keep turning out to be too specific. The Guidance ethics approach arose from the need for a more active, innovative form of ethics that is appropriate for the issues raised by technological developments. The philosophy of Prof. Peter-Paul Verbeek (University of Twente) on guidance ethics provided the inspiration. The focus here is not on assessing the ethics of a technology but on guiding that technology within society. The Guidance ethics approach was developed in a broadly composed ECP working group led by Prof. Verbeek and Dr Daniël Tijink.

Keywords here are connecting to real-world practice (bottom-up), input from stakeholders (the public, professionals, policy and technology), bottom-up and a focus on what we do want rather than what we don't want (generating options for action). Based on this, the Guidance ethics approach was created that consists of a workshop guided by moderators. In this, the various stakeholders engage in a dialogue about how to apply a real-world technology in a specific context so that they can jointly produce several specific options for action for doing so in an ethically responsible way. Listing the options for actions to take helps ensure that it is not merely talk and that the technology concerned becomes embedded in a genuinely value-based way in the day-to-day running of the organisation or within a network. The workshop's programme consists of the following components:

- Technology in context: joint discussions of actual or possible uses of a specific technique in a specific context based on input from an expert.
- Dialogue: dialogue with relevant stakeholders (or dialogue from their perspectives) about values and the effects of the technology in that context.
- Options for 'ethical' actions: jointly coming up with options for actions that make ethical application a possibility. These can be options that may relate to technology (ethics by design), the environment (ethics for organisational and other contexts) and the user/individuals (ethics for human behaviour).

This approach has been used dozens of times, some of them dealing with the use of AI applications, for example in healthcare, the police and in public administration. The approach is very much appreciated, both in terms of the substantive output and the process (getting people involved). Because of the social objectives of ECP, efforts are being made to accumulate and disseminate knowledge jointly. There are several publications on the website www.begeleidingsethiek.nl (in Dutch), as well as reports on the sessions, by the parties that have given their consent. There are quite a lot of them, which is unique for an ethics approach. A short training course has also been developed for Guidance ethics approach moderators, so that people within organisations or consultants can also use it themselves.

CASE STUDY: U-PREVENT

U-Prevent provides an interactive evidence-based risk prediction model for individual patients with cardiovascular disease. U-Prevent helps estimate the individual effect of different types of medication to determine which drug or combination of drugs would be best for each patient. To do that, the patient's data is compared against data from individuals of similar weight, gender, age, lifestyle and so forth for whom the progression of the disease is known. This assists joint decision-making by the doctor and patient in the consulting room. The workshop was organised by the Netherlands Patients' Federation (NPF) and the two parties behind U-Prevent (UMC Utrecht and Ortec) as part of the Health and Care working group of the NL AIC, and it was supervised by ECP. The sessions included several stakeholders, a patient, people from the Heart Council, UMC Utrecht and the supplier, someone from the healthcare institute, a cardiac nurse and doctors from UMC Utrecht and UMC Amsterdam.

The session yielded a broad set of positive and negative impacts that could be translated into values. The group chose quality of care, quality of life and autonomy as the most important ones. This identified numerous courses of ethical actions for U-Prevent, the organisation and the people affected. You can read all about this in the report on the session[2]. In an interview one year later, the initiator stated that the session had made a serious impact. "The core of it is that the Guidance ethics session helps implement innovations successfully. There's a lot of talking about innovation, but we invest very little in implementation. That session does help. You're investing in a dialogue with the various people who are involved in the implementation: patients, care providers, designers and policymakers – and that simply helps create a better product that is better embedded. We've already taken several of the suggestions on board."

Contact for GEA:
Daniel Tijink, ECP
(Daniel.Tijink@ecp.nl)

2. https://begeleidingsethiek.nl/wp-content/uploads/2021/06/Verslag-workshop-Begeleidingsethiek-U-Prevent-14042021.pdf

METHOD 2: DATA ETHICS DECISION AID (DEDA)

WHAT IS IT?

The Data Ethics Decision Aid (DEDA) is a tool and a method for ethics-based deliberations, decision-making and documentation associated with data projects. It was developed by the University of Utrecht over the course of 2016[3]. Several researchers from Utrecht Data School were involved in it, including Aline Franzke, Dr Mirko Tobias Schäfer and Iris Muis. Development was driven by the need for ethical frameworks for Big Data research, which were not available at the time[4]. An initial public version was published in 2017 in cooperation with the Municipality of Utrecht, aimed specifically at use by governmental organisations. Since then, DEDA has been used by many government organisations and several updates of DEDA have been published, adapted to reflect user experiences and advances in legislation and technology.

DEDA consists of a large poster, A0 size, that can be used by project teams working on a specific data project. This may be an algorithm but could equally be the creation of a dashboard, the use of drones or other ways of developing and deploying digital technology. It is a tool for interdisciplinary teams: to get the broadest possible picture of the potential ethical pitfalls of a project, it is essential that people from different functions and backgrounds are present during the DEDA discussions. The tool involves asking questions on various topics that are answered by the team together. The topics discussed can for instance be bias, data sources, access, responsibilities, privacy, etc. The team runs through the questions, discusses and answers them, and documents the answers. When finished, the team should have identified the ethical bottlenecks in a project, discussed them and made decisions about them. The ethical considerations will have been documented so that they can be used for justification. During the corona virus pandemic, the DEDA poster was converted into an interactive PDF so that the method can also be used during online meetings. There is a manual to help you use DEDA. The tool is available in Dutch, English, German, Swedish and Finnish.

Several studies have been carried out into the use of DEDA and its effectiveness. In 2019, a Master's student investigated the effect of a DEDA intervention[5]. The conclusion was that DEDA made a genuine contribution to ethical use of data within the governmental organisations studied. A detailed account of the development and use during the early years was published in 2021 in the leading scientific journal Ethics and Information Technology[6]. A scientific article about DEDA's participatory and ethnographic research method was published in 2022[7]. It also describes how using DEDA raises the level of ethical awareness associated with data projects within an organisation.

.

3. All DEDA materials are available on https://dataschool.nl/deda/.

4. DEDA was later the inspiration behind the international ethical guidelines of the Association of Internet Researchers – see Franzke, Aline Shakti; Bechmann, Anja; Zimmer, Michael; Ess, Charles and the Association of Internet Researchers (2020). Internet Research: Ethical Guidelines 3.0. https://aoir.org/reports/ethics3.pdf

5. Hans van Wijk. "Wanted: Data Ethics Assistant (fulltime, public sector)". Public Administration, 2019. http://hdl.handle.net/2105/47341

6. Franzke, A.S., Muis, I. & Schäfer, M.T. Data Ethics Decision Aid (DEDA): a dialogical framework for ethical inquiry of AI and data projects in the Netherlands. Ethics Inf Technol 23, 551–567 (2021). https://doi.org/10.1007/s10676-020-09577-5

7. Siffels, L., van den Berg, D., Schäfer, M.T., Muis, I. (2022). Public Values and Technological Change: Mapping how Municipalities Grapple with Data Ethics. In: Hepp, A., Jarke, J., Kramp, L. (eds.) New Perspectives in Critical Data Studies. Transforming Communications – Studies in Cross-Media Research. Palgrave Macmillan, Cham. https://doi.org/10.1007/978-3-030-96180-0_11

CASE STUDY: PERSONALISATION OF A GOVERNMENTAL AUTHORITY WEBSITE

In September 2018, a large governmental organisation contributed a case study for a DEDA workshop. It involved modernising a governmental website. The project was not initially seen as controversial or ethically complex. There were two directions that the modernisation of the website could have taken: either personalising the website, so that each member of the public would get to see a different variant entirely adapted to their needs, or not adapting the website, so that every visitor would see the same page and the same information from the authorities. The DEDA workshop participants soon related this choice to the public values of equality and equal access to public information. Personalising governmental websites would mean that not all members of the public would have exactly the same access to governmental information. And who would determine what an individual got to see and what they didn't? How could that process be made transparent? The workshop participants concluded that they did not have a mandate to make such choices and that this discussion should be conducted in the political sphere, by people with a political mandate. Ultimately, a decision was taken not to personalise the website.

Contact for DEDA:
Iris Muis, Utrecht University
(i.m.muis@uu.nl)

METHOD 3: FUNDAMENTAL RIGHTS AND ALGORITHMS IMPACT ASSESSMENT (FRAIA)

WHAT IS IT?

Technology and algorithms offer numerous opportunities for companies and authority bodies to set up their operations more efficiently and more effectively. This can also involve pitfalls, though, particularly when algorithm-based decisions affect people. These can impinge upon human rights. Injustices associated with algorithms have come to light in recent years and are frequently discussed in the media. FRAIA was developed to detect and overcome such wrongs associated with algorithms at an early stage. The Fundamental Rights and Algorithms Impact Assessment – also known in Dutch as IAMA (Impact Assessment Mensenrechten en Algoritmes) was developed by a team at Utrecht University (Prof. Janneke Gerards, Dr Mirko Tobias Schäfer, Arthur Vankan and Iris Muis), on instructions from the Ministry of Internal Affairs. FRAIA has been published on the national government's website (as IAMA) and it is accessible to the public[8]. FRAIA is only the acronym used for the English version: Fundamental Rights and Algorithms Impact Assessment[9].

What fundamental rights are being infringed? How likely is it that this could occur? What would the impact be for someone? Is that impact proportionate to the purpose of the algorithm? And are those considerations – i.e. whether it is or is not acceptable – transparent enough and sufficiently explainable? In a nutshell, that is the mirror that FRAIA holds up to the users of an algorithm. FRAIA is an interactive PDF with various components:

- The reasons for developing and using an algorithm, and the principles behind it;
- The technology: nature and quality of the input and the algorithm itself;
- The implementation: the context in which the algorithm is used, the communications strategy, etc.;
- Human rights considerations: do the benefits of the algorithm outweigh the negative impact on fundamental rights?

An interdisciplinary team goes through those components answering the questions together and documenting those answers. Ultimately, using FRAIA is intended to produce a well-considered assessment and decision about the justification for using algorithms.

A motion was introduced in the Dutch parliament in March 2022, calling for IAMA (FRAIA) to be mandatory for algorithms used by the government[10]. That motion was passed but has not yet been implemented. So we have to wait and see whether (and in what form) using FRAIA will be made mandatory in the future.

8. Gerards, J., Schaefer, M., Vankan, A., & Muis, I. (2021). Impact Assessment Mensenrechten en Algoritmes. https://www.rijksoverheid.nl/documenten/rapporten/2021/02/25/impact-assessment-mensenrechten-en-algoritmes

9. Gerards, J., Schäfer, M. T., Muis, I., & Vankan, A. (May 2022). Fundamental Rights and Algorithms Impact Assessment (FRAIA). National government. https://www.government.nl/documents/reports/2021/07/31/impact-assessment-fundamental-rights-and-algorithms

10. See https://www.tweedekamer.nl/kamerstukken/moties/detail?id=2022Z06024&did=2022D12329

CASE STUDY: SCANNER VEHICLES FOR PARKING CHECKS

In the Municipality of Rotterdam, scanner vehicles are being used that drive around the city to detect cars that have been parked without paying and issue fines[11]. Initially, Rotterdam reckoned that this was a high-risk algorithm, as personal data could be involved. The algorithm behind the scanner vehicles was evaluated using FRAIA. After running through the tool, it transpired that the risks of the scanner vehicle posed no problems. After all, they only look at the licence plates. Various items of personal data can become involved, but only if there has been a parking violation or if no parking fee (or not enough) has been paid and the car owner also does not have a parking permit; a fine is issued in those cases. The results of the FRAIA assessment have been used to tighten up the accountability for the scanner vehicles in Rotterdam's algorithm list. That accountability is publicly accessible[12].

Contact for FRAIA:

Iris Muis, Utrecht University
(i.m.muis@uu.nl)

11. See the detailed interview with the Municipality of Rotterdam and Utrecht University about this case study and FRAIA in general: https://www.mensenrechten.nl/actueel/toegelicht/interviews/2022/laten-we-nou-vooral-leren-van-gemaakte-fouten-en-kijken-of-we-algoritmes-wel-verantwoord-kunnen-inzetten

12. See https://www.rotterdam.nl/bestuur-organisatie/algoritmeregister/

METHOD 4: TECHNOLOGY IMPACT CYCLE TOOL (TICT)

WHAT IS IT?

The Technology Impact Cycle Tool (TICT) was developed by a multidisciplinary team at Fontys[13]. It is a toolkit created to make students (and professionals) think more carefully about the impact of technology on society and give them the chance to develop or deploy a product for which the ethical aspects have been considered more carefully.

As with the Guidance ethics approach, which originated at more or less the same time, the ideas of Peter-Paul Verbeek[14] (University of Twente) were the basis, although they were also combined in this case with a theory of Ibo van de Poel[15] (TU Delft). We have adopted two of Verbeek's ideas. Firstly, that it is a good idea to guide technology rather than condemn it 'a priori'. Second, that technology is often not ethically neutral, as humans and technology have been intimately interlinked for decades; we are 'Homo technologicus'. The idea we took from Ibo van de Poel is that it is in developers' nature to consider mainly internal values (does the product work, are all the bugs ironed out?), which can make them overlook external values (what kind of society are we actually building? How does this affect human values – equality? Bias? Inclusivity? Privacy? Sustainability? etc.)

Achieving ethically responsible and meaningful innovation will therefore also need external values to be internalised, and this will have to be seen and experienced as a shared responsibility: a multidisciplinary approach therefore seems advisable here. Moreover, these values should preferably be taken into account by the team involved throughout the process of design or implementation. New understandings can sometimes lead to new outcomes and, in addition, the context can also change in the meantime. That meant that it seemed sensible to return to the ethics evaluation several times and to use the tool cyclically. This was partly because there is often no time, space, energy, budget or desire left to make changes if ethical issues are only examined at the end of the process.

The toolkit does not give you ready-made answers: it primarily contains questions (divided into ten categories) that encourage you to think more carefully about the impact you are trying to achieve with the product. It also encourages reflection on human values such as inclusiveness, transparency, sustainability and privacy. Stakeholders, data and bad actors are also addressed. On top of that, there is a category asking you questions about what could potentially happen in the future when the real product is launched in society and/or if it were to be applied differently than originally intended.

The toolkit consists of a quick scan, a full scan and an improvement scan; it also includes comprehensive masterclasses for each category in which you can build up more knowledge about that particular category. Additionally, you will also find references at the end of each category to other models, toolkits and literature that you can use to learn more. Many of the other methods mentioned in this publication can therefore also be found in this method.

The toolkit can be used by all, free of charge: ww.tict.io.

13. Kamp J.M. and Vorst, R.M.C.M. (June 2022) Moral impact (pp. 43–62); In: Wernaart, B. (ed.). Moral design and technology. Wageningen: Wageningen Academic Publishers.

14. Verbeek, P., & Jong, de, R. (1 June 2017). MOOC Philosophy of Technology and Design — University of Twente. Consulted on 21 October 2019, from https://www.futurelearn.com/courses/philosophy-of-technology

15. van de Poel, I. (2015). Values in Engineering and Technology. In: Gonzalez, W. (ed.) New Perspectives on Technology, Values, and Ethics. Boston Studies in the Philosophy and History of Science, vol. 315. Springer, Cham. https://doi.org/10.1007/978-3-319-21870-0_2

CASE STUDY: BABY DON'T CRY

Young parents are often uncertain whether they are doing everything right for their newborn child. So how easy would it be if there was an app that could recognise and analyse the sounds of their crying baby and tell the parents that the baby is tired or hungry or in pain? After all, each sound is different and an AI can be trained to distinguish between them. That is the hypothesis, at any rate. But what kind of society would we then be creating? If we hand that responsibility over to a system, are we rendering ourselves redundant? Will we stop trusting our own intuition? Will we become dependent on an app? And can we trust the results absolutely?

Will that kind of technology be made available to everyone in due course? Or only the happy few who can pay for it? Will some people be excluded? Or conversely, might the app be made mandatory and will society or your health insurer blame you for not listening to the app and instead using your own instincts (or turning the app off to save energy)? Will you then be deemed to have acted irresponsibly? And what should you do if your baby keeps on crying anyway? Is that your fault?

And what does it actually mean for a child if its behaviour is captured as data and quantified from a young age? Are they then still able to develop into an autonomous individual? And what happens to all that data? Does it go to a commercial company? Will it be stored on a server in a different country, subject to a different legal system? Can that data be used against you at a later date? Can it be hacked? Is it even justifiable to start creating such an app? If so, what requirements must it comply with? And how can we then make it 'better'?

Contact for TICT:
Jo-An Kamp, Fontys University of Applied Sciences
(j.kamp@fontys.nl)

METHOD 5: DATA GOVERNANCE CLINICS (DGC)

WHAT IS IT?

The data governance clinics approach was developed by an interdisciplinary team of researchers from the Tilburg Institute for Law, Technology and Society at Tilburg University[16]. As an approach, it focuses on setting up governance structures for data-driven innovative projects in the public domain. It focuses in particular on projects within public organisations or public-private partnerships, although it can also be used at commercial or civil society organisations. Like the methods mentioned earlier, this approach assumes that technology is not neutral and that specific core values are reflected not only in the technical design but also in the organisational and social structures surrounding it. The data governance clinics approach therefore assumes that the public interest must be safeguarded when developing these technologies, without adopting a position on what that public interest involves.

The challenge for developers of digital data-driven systems in the public domain, such as crowd monitoring systems or school allocation systems, is to develop the system in line with the relevant core values. However, which values matter and how they should be interpreted can vary from one context to the next. On top of that, there will all too often not be a consensus about them. Who ultimately ought to be taking the decision and how can we ensure that the system complies with this decision in reality? Another challenge is that new questions and dilemmas continually arise as new technologies are developed that cannot always be envisaged in advance. A one-off reflection session about ethical questions at the start of the development process is soon seen to be insufficient.

The researchers wanted to use this approach to go a step further than identifying and operationalising ethical values once only or translating them into technical designs: they wanted to focus on how continuous ethical reflection about public values and interests can mould innovative development practices. They see the governance structures surrounding the development of new systems as a starting point for this. The approach starts from existing practices and aims to embed continuous ethical reflection as part of governance practices that go beyond compliance and take public interests in the broader context as their starting point.

Together with the moderators who run the clinic, the project team goes through an iterative process that first identifies possible pain points related to public interests and then considers how existing governance structures can address those points and where the gaps are. Perspectives for the actions to take are then formulated based on that analysis. Such action perspectives could include e.g. expanding or adapting the role of supervisory bodies, assigning responsibility to a person in development practices for actively highlighting ethical considerations during the development process, as well as putting broader social consequences of the project on the agenda at the local political level. An important role of the clinic moderators, in addition to process guidance, is to question implicit assumptions, offer alternative perspectives and make ethical choices explicit.

16. Jameson, S., Taylor, L., & Noorman, M. (2021). Data Governance Clinics: a new approach to public-interest technology in cities. https://pure.uvt.nl/ws/portalfiles/portal/57039352/Data_Governance_Clinics_2021.pdf

CASE STUDY: CROWD MONITORING

The Municipality of Amsterdam is very much aware of problems that may be associated with the new data-driven technologies it hopes to deploy and has taken various steps to address some of these issues. The actions taken by the municipality include defining a set of values and guidelines for data-driven systems, where core values such as openness and transparency are central. Those values have to be translated into practice, though. What do these values mean within specific innovation projects? Where do conflicts arise between the values? And how can translations of abstract values be aligned with the public interests? Who decides what interpretation should be adopted or how the priorities between competing values should be set?

These questions came to the fore in one of the clinics for the Municipality of Amsterdam. That clinic was about a crowd monitoring project in the city of Amsterdam, based on a large amount of data[17]. Within the project, innovations relating to mobility were applied experimentally in the public domain. On the one hand, this yields real-world, practical questions about the ethical aspects of these innovations, and on the other more strategic questions about how the decision-making about such questions should be set up. Such strategic questions include e.g. issues about the democratic legitimisation of decisions within public-private partnerships and the relationship between the innovation centre and the democratic authorities. Together with researchers and external experts, the project team further described these two types of questions during the clinic session and drew up a list of various possible actions for addressing both the ethical and strategic issues.

Contactpersoon DGC:
Merel Noorman, Tilburg University
(m.e.noorman@tilburguniversity.edu)

17. https://openresearch.amsterdam/nl/page/87977

METHOD 6: ASSESSMENT LIST FOR TRUSTWORTHY AI (ALTAI)

WHAT IS IT?

The Assessment List for Trustworthy Artificial Intelligence (ALTAI) was developed by the High-Level Expert Group on Artificial Intelligence on instructions from the European Commission[18]. It is what is known as a 'self-assessment' that can be used by developers, development teams and the organisations they work for. It helps these groups estimate the extent to which their intended AI product is in line with the 'Ethics Guidelines for Trustworthy AI', developed by the same High-Level Expert Group[19].

According to these guidelines, reliable AI must (1) be lawful under all applicable legislation and regulations, (2) be ethical and thus respect ethical principles and values, and (3) be robust from both a technical point of view and in the way it responds to the social surroundings. They then put forward seven requirements that trustworthy AI must comply with, namely:

1. *Human Ownership and Supervision.*
   AI systems should enable people to make informed decisions. Moreover, AI systems should promote fundamental human rights. At the same time, proper monitoring mechanisms should be put in place, approaches such as a human in the loop, a human on the loop and human in command.

2. *Technical robustness and safety.*
   AI systems must be both resilient and safe. They should have contingency plans in case something goes wrong and must also be accurate, reliable and reproducible. That is the only way to make sure that unintended damage can be minimised and prevented.

3. *Privacy and data management.*
   As well as ensuring privacy and data protection, appropriate data management mechanisms also need to be guaranteed. This must take account of data quality and data integrity, and guarantee only legitimate access to data.

4. *Transparancy.*
   Data, systems and AI business models must be transparent. Traceability mechanisms can help with this. Moreover, AI systems and their decisions should be explained in a way that is suitable for the interested party (personalised explanations, for instance). People need to be made aware that they are interacting with an AI system and should be informed about the capabilities and limitations of the system in question.

5. *Diversity, non-discrimination and fairness.*
   Unfair biases must be avoided as they can have all kinds of negative consequences, from marginalisation of vulnerable groups to exacerbation of prejudice and discrimination. Furthermore, to promote diversity, AI systems must be accessible to everyone, regardless of disability, and must involve relevant stakeholders in decision-making processes about creating and deploying the AI systems in question throughout the lifecycle.

18. https://altai.insight-centre.org/
19. https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai

6. *Social and environmental well-being.*
   AI systems should benefit all people, including future generations, and steps must therefore be taken to ensure that they are sustainable and environmentally friendly. On top of that, they must take account of the environment, including other living beings, and their social and societal impacts must be carefully considered.

7. *Liability.*
   Mechanisms must be put in place to safeguard the responsibility and accountability for AI systems and their results. Auditability – which allows algorithms, data and design processes to be assessed – plays a key role in that, especially in critical applications. On top of that, an appropriate and accessible rationale must be provided.

ALTAI is a self-assessment list (that has incidentally also been made available digitally) that is intended to be run through and completed collectively by development teams and interested parties (as a multidisciplinary approach, i.e. AI designers/developers, data scientists, procurement specialists, end users, legal/compliance officers, managers, data subjects, etc.). It is used in the form of what is known as a 'do-confirm checklist', checking for potential risks and encouraging the team to come up with mitigations and solutions to avoid or minimise those risks. The team is encouraged to describe clearly how they are meeting each of the requirements listed in their design of the technology. In practice, you will see that teams do not complete the checklist all in one go, instead regularly revisiting the product and the questionnaire to bring the existing or intended product ever more into line with the intended goal on the one hand and the preconditions imposed by ALTAI on the other.

CASE STUDY: REVIEW TOOL

ALTAI was used in this case to evaluate several proofs of concept developed by the data team concerned, then using those evaluations to draw up an action plan to improve the proofs of concept as held up to the yardstick of the seven ethics guidelines. The reviews were done in two different configurations for each proof of concept: 1) the domain experts, focusing primarily on the problem, the application, the significance of the data used and the embedding in the organisation, and 2) the technical experts, focusing largely on the technical choices – for the design in particular – and mitigation options for bringing the design better into line with the guidelines. The review was held once or twice annually, also looking at the extent to which progress had been made with respect to the preceding review and the roadmap that it had produced.

Contact for ALTAI:
Marieke Peeters, Mooncake AI
(marieke.peeters@mooncake-ai.com)

METHOD 7: DATA PROTECTION IMPACT ASSESSMENT (DPIA)

WHAT IS IT?

The Data Protection Impact Assessment (DPIA) is not a rigidly defined instrument, but instead an evaluation that has been made mandatory. Organisations can complete it for themselves, although there is a template that has been drawn up by the United Kingdom[20]. The DPIA is obligatory under the General Data Protection Regulation (GDPR – Article 35), as well as the Police Data Act and the Judicial Data and Criminal Records Act[21,22]. A data protection impact assessment describes a process designed to identify risks arising from processing personal data and minimise those risks as much and as early as possible. DPIAs are important tools for mitigating risks and demonstrating GDPR compliance. A DPIA is required when processing personal data is likely to engender a high risk to the rights and freedoms of individuals. A DPIA is required at least in the following cases:

- Systematic and comprehensive evaluation of personal aspects based on automated processing, including profiling, and decisions that affect people based on such evaluation;
- Large-scale[23] processing of special personal data or criminal records;
- Large-scale and systematic tracking of people in publicly accessible areas (e.g. through camera monitoring).

National data protection authorities (DPAs) such as the Dutch Data Protection Authority[24] may, in consultation with the European Committee for Data Protection, provide lists of cases in which a DPIA is required. The DPIA must be carried out before the data processing and should be considered a living tool, not just a one-off exercise. If there are residual risks that cannot be mitigated by the measures taken, the DPA should be consulted before processing begins.

Generally, at least the following people are involved in the DPIA process: the product owner or innovation manager, the privacy officer, the security officer, the data analyst or data scientist, and the internal or external client or case owner. In some cases, a Data Protection Officer (DPO) is also appointed, who also monitors the DPIAs to ensure that the company or organisation in question complies with the GDPR. Because a DPIA is genuinely about compliance with legislation and regulations, the whole exercise is often seen as a time-consuming hurdle that simply has to be taken. In practice, the DPIA process therefore often becomes a matter of ticking all the boxes so that people can get on with the project. Furthermore, the GDPR – and thus the DPIA procedure – sometimes conflicts with other legislation and regulations, leaving those doing the implementation work (data scientists, developers) in an awkward spot: take money-laundering legislation for banks, for instance, which requires them to check more rigorously for suspicious transactions that might indicate money-laundering activity[25].

20. https://gdpr.eu/data-protection-impact-assessment-template/

21. https://ec.europa.eu/info/law/law-topic/data-protection/reform/rules-business-and-organisations/obligations/when-data-protection-impact-assessment-dpia-required_en

22. https://www.autoriteitpersoonsgegevens.nl/nl/zelf-doen/data-protection-impact-assessment-dpia

23. https://www.autoriteitpersoonsgegevens.nl/nl/onderwerpen/algemene-informatie-avg/algemene-informatie-avg#wat-ziet-de-avg-als-een-grootschalige-verwerking-van-persoonsgegevens-6019

24. https://www.autoriteitpersoonsgegevens.nl/nl/zelf-doen/data-protection-impact-assessment-dpia

25. https://www.banken.nl/nieuws/24105/bunq-wint-witwas-rechtszaak-tegen-de-nederlandsche-bank

CASE STUDY: FRAUD DETECTION

A few years back, the Dutch government wanted to introduce a new fraud detection system called SyRi. It linked together all kinds of data about members of the public from multiple government agencies to identify 'conspicuous' or 'suspicious' behaviour. Ultimately, the courts decided that the level of data collection, the way it was done and the processing of personal data were disproportionate to the objectives.

Companies and organisations will have to complete DPIAs for such plans to determine for themselves (sometimes under the watchful eye of the DPO and the DPA) whether certain personal data processing is legally proportionate, permissible and acceptable, given the purpose of that processing.

Contact for DPIA:

Marieke Peeters, Mooncake AI
(marieke.peeters@mooncake-ai.com)

RECOMMENDATIONS

This publication has listed seven methods, each with its own specific combination of characteristics, goals, strengths and weaknesses. To conclude, we would like to share a few recommendations that apply to all the methods and which will help organisations deploy the right method effectively at the right time.

Methods can be used side by side; knowledge of methods is required

The first conclusion is that the various methods are not mutually exclusive. We are seeing in current practice that organisations adopt a single method and then apply it regularly, thereby excluding other methods. The seven methods covered in this publication show that there can be different rationales for employing specific methods. When the organisation's aim is to bring in external perspectives and explore all possible values and options for action under the guidance of a moderator, for instance, the Guidance ethics approach is probably a suitable method. But if people prefer to learn a little about values independently and work – individually or in groups – on a particular case, the TICT method is somewhat more self-directing and therefore sometimes easier to use. When the objective is more about ring-fencing all aspects around the ethical use of data, DEDA is in turn more appropriate. FRAIA is a highly suitable method if algorithms are involved. In other words, each method has its own angle and its own strengths and weaknesses. This requires the relevant official within the organisation to be particularly knowledgeable about the methods, so that they understand that it may also not be appropriate to deploy a certain method in some cases.

Implementation is important; make clear where responsibility for this lies and thus ensure continuous attention to ethics

The second conclusion is that ethical technology implementation and development, in our experience, only begin when a method is applied. The pitfall here is that concentration on ethics slackens after a session or assessment, and the usefulness of applying the method fades with it. What helps here is appointing someone in charge within the organisation (or finding a volunteer) who continues to generate attention for ethics. This could be for a specific case, for instance as they continue to question the internal stakeholders involved about the follow-up of a method. Equally, it could be a focus on the ethics and governance of AI within the organisation, addressing e.g. the relationship between ethics and the organisation's strategy. This constant attention being paid to ethics is important for another reason too: digital technology, by its very nature, is constantly changing and therefore never finished. That also has implications when applying the methods described: applying a method at the beginning of the process may for instance yield outcomes that are no longer relevant after several rounds of development, for example because users use the product in a different way than had been expected beforehand. This requires constant monitoring of products to keep it clear when a particular method needs to be reapplied (or another method adopted). Applying the methods listed in this publication sometimes demands a lot of time and resources from the organisation and its stakeholders. We are therefore also in favour of flexibility when applying the methods, noting that regularly revisiting the outcomes in a smaller group can be enough to be useful.

Applying a method also raises ethical awareness.

Lastly, our final conclusion is that our comparison of methods has shown that each method produces specific outputs: in the case of the Guidance ethics approach, you get a report with options for actions to take; TICT produces a PDF listing trade-offs and improvements; DEDA yields a completed poster, and so forth. But in addition to those direct outputs, we also see when methods are applied that the mere act of applying a method creates awareness within the organisation – not only awareness of what ethical risks and opportunities there are, but also of what can be done about them and which people within the organisation can be involved. This is an outcome that is in principle independent of the method used, but it is nevertheless a result that should always be considered when weighing up whether utilising a method is worth the time and resources required. Applying a method to a specific case therefore also produces effects that are not related

directly to the case in question but are much broader. We therefore see that a choice is made in some organisations to implement the methods discussed more widely, for example by including them in the regular training that is made available within that organisation.

CALL FOR PARTICIPANTS

We welcome new participants in the PACE platform and are open to new knowledge and new perspectives. If you are interested in attending a meeting about participative and constructive ethics, if you have any suggestions or if would you like to work with us actively on this very relevant topic, take a look on the website[26] or contact us at communicatie@nlaic.com.

26. https://nlaic.com/en/

Editing
Human Centric AI Working Group

Contact
E-mail — communicatie@nlaic.com
Website — nlaic.com/en

December 2022


NL AI Coalition