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

Steen, Marc; Timan, Tjerk; van de Poel, I.R.

DOI
10.1007/s43681-021-00063-2

Publication date
2021

Document Version
Final published version

Published in
AI and Ethics

AI and Ethics
https://doi.org/10.1007/s43681-021-00063-2

ORIGINAL RESEARCH

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

Marc Steen[1] · **Tjerk Timan**[1] · **Ibo van de Poel**[2]

Received: 5 January 2021 / Accepted: 10 May 2021

Abstract
The collection and use of personal data on citizens in the design and deployment of algorithms in the domain of justice and security is a sensitive topic. Values like fairness, autonomy, privacy, accuracy, transparency and property are at stake. Negative examples of algorithms that propagate or exacerbate biases, inequalities or injustices have received ample attention, both in academia and in popular media. To supplement this view, we will discuss two positive examples of Responsible Innovation (RI): the design and deployment of algorithms in decision support, with good intentions and careful approaches. We then explore potential, unintended, undesirable, higher-order effects of algorithms—effects that may occur despite good intentions and careful approaches. We do that by engaging with *anticipation* and *responsiveness*, two key dimensions of Responsible Innovation. We close the paper with proposing a framework and a series of tentative recommendations to promote anticipation and responsiveness in the design and deployment of algorithms in decision support in the domain of justice and security.

Keywords Responsible innovation · Algorithms · Decision support · Justice and security · Higher-order effects

1 Introduction

In this paper we are concerned with the utilization of algorithms in the domain of justice and security. The usage of algorithms is associated with promises to increase efficiency and effectiveness of, e.g., law enforcement or the judiciary process. Using algorithms can, however, also cause significant harms, e.g., in terms of bias, unfairness and discrimination [9, 12, 13, 32], and can fall short regarding transparency and accountability [5]. Our aim is to explore both positive and negative views on algorithms more specifically, we will explore potential, unintended, undesirable, higher-order effects of algorithms—effects that may occur despite good intentions and careful approaches.

We will focus on algorithms that are used for *decision support*. For example, a police officer may use an algorithm for risk assessment while engaging with citizens. This algorithm may show red flags behind a specific citizens' names. Officers will typically interpret such flags as advices and interpret them based on their professional perception, understanding and judgement of the situation at hand. This is their professional, discretionary competence.

We can contrast this usage of *algorithms as tools* for decision support, with the usage of *algorithms as* agents,[1] e.g., in systems that autonomously make decisions without human intervention. This contrast, however, is not a sharp one and rather describes a continuum. Imagine that police officers always follow the algorithm's advice. They never neglect or overrule the algorithm's advice. Do these police officers then act like *agents*? Or more like a cogs in a machine? Or, conversely, imagine that police officers *never* follow the algorithm's advice. They always neglect or overrule it. What is then the algorithm's *function* or added value? In reality,

✉ Marc Steen
marc.steen@tno.nl

[1] TNO, The Netherlands Organisation for Applied Scientific Research, The Hague, The Netherlands

[2] Delft University of Technology, Delft, The Netherlands

[1] This conceptualization is disputed; many have argued that so-called *autonomous* systems cannot, or should not, have autonomy in an ethical sense, i.e. in terms of moral agency and responsibility [2, 6, 14, 19, 36, 45].

Published online: 29 May 2021

many police officers will do something in-between: they will use their professional, discretionary competence and sometimes follow the algorithm's advice, and sometimes neglect or overrule it. We understand the agency of police officers in the context of a sociotechnical system [25], in which their agency is connected to other persons, e.g., to supervisors or peers, or to reports in the police organization, and to various machines and processes to which they delegate some tasks and parts of their agency.

The design and deployment of algorithms in the domain of justice and security raises a range of ethical questions [29], values like fairness, autonomy, privacy, accuracy, transparency and property are at stake [20]. We can further extend this list with values from, e.g., the tradition of *Value Sensitive Design* [17][2] or the European Commission's High-Level Expert Group on Artificial Intelligence [21].[3] Hayes et al. [20] argue that some values have intrinsic value, e.g., fairness, autonomy and privacy, and that other values have instrumental value. The latter function as means to support other values, e.g., accuracy and transparency are typically instrumentally needed to support fairness. Or the other way around: a *lack* of accuracy or of transparency can negatively impact fairness. Another example is autonomy: police officers' autonomy is dependent on the accuracy of the data that go into algorithm ('garbage in, garbage out') and on the transparency of the algorithm. Can they trust the system and explain what they do, e.g., to citizens they are engaging with? In more general terms, the human decision-maker's autonomy depends on the system's accuracy and transparency.

These relationships between values complicate matters considerably. Some values may be needed to support other values, like accuracy and transparency are needed to support fairness and autonomy. Or values can conflict: for example, ownership and property rights of parties that develop or sell algorithms can come into conflict with other parties' values regarding transparency. Another example is the potential conflict between citizens' privacy and accuracy, since the accuracy of machine learning algorithms is typically related to having a lot of data points on a lot of people, to 'train' the algorithm.

This knowledge, about values and interdependencies and conflicts between values, can support people in the design and deployment of algorithms. They can identify and discuss values that are at stake in their project. It remains challenging, however, to identify which specific values are relevant in a specific case, and to find *appropriate* balances between conflicting values. One way to deal with is to resort to general principles. One can, e.g., follow the principles that the European Commission's High-Level Expert Group on Artificial Intelligence [21] put forward: respect for human autonomy, prevention of harm, fairness, and explicability. They will, however, typically need to do some heavy lifting to 'translate principles into practices' [30]. Merely being aware of ethical principles will not bring about beneficial algorithms [28]. Or worse, just *signalling* of ethical principles can be unethical [15].

2 A twofold view and methodology

It occurred to us that *negative examples* typically receive ample attention, both in academia and in popular media. Think of the books and articles about algorithms that propagate or exacerbate biases, inequalities or injustices, e.g., in predictive policing or in the judicial process. This work is very valuable indeed for many reasons. In a relatively young and dynamic field like data science, which deals with emerging and consequential technologies, critical studies are direly needed. On the other hand, we believe that it can also be instructive to share *positive examples;* examples of 'Responsible AI' [11]. We can learn from projects in which people worked with good intentions and with careful processes. We therefore follow a twofold methodology in this article. We first focus on *positive* or responsible examples. We discuss the principles put forward by the *High-Level Expert Group* (2019) and provide some context for the design and deployment of algorithms in government (in The Netherlands; where the authors reside). We then discuss two positive examples (from the Netherlands; for pragmatic reasons), to discuss the application of these principles in design and deployment 'on the ground'. After that, we shift gears and adopt a more *critical* view. We engage with *anticipation* and *responsiveness*, two key dimensions of Responsible Innovation [40], and explore several potential, *unintended, undesirable,* higher-order effects of algorithms. We believe that, even if it might not always be possible to prevent such negative effects from happening, it is at least useful to anticipate such effects, to be responsive, and attempt to mitigate their impacts, as much as possible. We close the paper with suggestions to further promote this type of anticipation and responsiveness.

[2] They mention: human welfare; ownership and property; privacy; freedom from bias; universal usability; trust; autonomy; informed consent; accountability; courtesy; identity; calmness; and environmental sustainability.

[3] They mention the following 'requirements': human agency and oversight; technical robustness and safety; privacy and data governance; transparency; diversity, non-discrimination and fairness; societal and environmental wellbeing; and accountability.

3 Algorithms and ethical principles: a European perspective

In April 2018, the European Commission stated its aspiration to ensure that Artificial Intelligence (AI) applications are based on values and benefit individuals and society.[4] One year later, in April 2019, the European Commission's High-Level Expert Group on Artificial Intelligence [21] put forward *Ethics Guidelines for Trustworthy AI*, with four 'ethical principles' for the design and deployment of 'lawful, ethical and robust' AI systems (pp. 11–13): respect for human autonomy; prevention of harm; fairness; and explicability.[5] Below, we relay their descriptions and discussions of these principles (pp. 12–13; in *italics*), and add several comments for our discussion of algorithms in decision support.

3.1 Respect for human autonomy

'Humans interacting with AI systems must be able to keep full and effective self-determination over themselves […]. AI systems should not unjustifiably subordinate, coerce, deceive, manipulate, condition or herd humans. Instead, they should be designed to augment, complement and empower human cognitive, social and cultural skills. The allocation of functions between humans and AI systems should follow human-centric design principles and leave meaningful opportunity for human choice.' This principle advocates carefully allocating (distributing or delegating) functions (or tasks or control) between people and AI systems. Regarding this issue, we may turn to Ben Shneiderman [38], who recently conceptualized agency as an interplay between *computer automation* and *human control*. He advocated viewing computer automation and human control *not* as opposites on one axis, but as two perpendicular axes, and combining *high computer automation* and *high human control* to make AI 'reliable, safe and trustworthy'. This conceptualization with two axes can help to go beyond 'algorithm as tool' versus 'algorithm as agent' dichotomy and explore ways to use algorithms under human control.

3.2 Prevention of harm

'AI systems should neither cause nor exacerbate harm or otherwise adversely affect human beings. This entails the protection of human dignity as well as mental and physical integrity. […] Particular attention must also be paid to situations where AI systems can cause or exacerbate adverse impacts due to asymmetries of power or information, such as between employers and employees, businesses and consumers or governments and citizens. Preventing harm also entails consideration of the natural environment and all living beings.' Additionally, we can understand the prevention of harm, like Floridi et al. [16] did, as preventing the *overuse* or *misuse* of technology, e.g., devaluing human skills, removing human responsibility, reducing human control, eroding human self-determination. Interestingly, they propose to also think about the *seizing* of specific opportunities, e.g., enabling human self-realisation, enhancing human agency, increasing societal capabilities, cultivating societal cohesion, because *not* seizing such opportunities can cause more harm than seizing them. This reasoning is relevant in the domain of justice and security. The general public tends to expect of organizations like the police or the judicature that they do use sophisticated technologies. On the other hand, the general public can be quick to criticise mistakes. It can then be challenging to organize innovation. A possible way out, is organizing careful experimentation and learning. We will return to this in the second half of this paper, under *anticipation*.

3.3 Fairness

'The development, deployment and use of AI systems must be fair [where fairness is understood as having] both a substantive and a procedural dimension. The substantive dimension implies a commitment to: ensuring equal and just distribution of both benefits and costs, and ensuring that individuals and groups are free from unfair bias, discrimination and stigmatisation. […] Additionally, … practitioners should respect the principle of proportionality between means and ends, and consider carefully how to balance competing interests and objectives. The procedural dimension […] entails the ability to contest and seek effective redress against decisions made by AI systems and by the humans operating them.' This description clarifies that fairness refers not only to algorithms in a narrow sense, but also to the processes and organizations in which algorithms are utilized. Moreover, fairness is embedded in the rule of law, e.g., in the *European Convention on Human Rights* (ECHR) and in the *General Data Protection Regulation* (GDPR), which

[4] https://ec.europa.eu/digital-single-market/en/news/communication-artificial-intelligence-europe.

[5] We chose to work with the High Level Expert Group's High four 'ethical principles', rather than with their seven 'key requirements' (*Human agency and oversight*; *Technical robustness and safety*; *Privacy and data governance*; *Transparency*; *Diversity, non-discrimination and fairness*; *Societal and environmental wellbeing*; *Accountability*). We do, however, address most elements of the 'key requirements' in our discussions of the 'ethical principles', as follows: we discuss *Human agency and oversight* under Respect for human autonomy; *Technical robustness and safety* and *Societal and environmental wellbeing* under Prevention of harm; *Diversity, non-discrimination and fairness* under Fairness; and *Transparency and Accountability* under Explicability. *Privacy and data governance* our outside our paper's scope, as mentioned in the introduction of the case studies.

enables member states' citizens to challenge the utilization of algorithms, to demand inspection of their personal data, and, if necessary, to seek correction and redress.

3.4 Explicability

'Explicability is crucial for building and maintaining users' trust in AI systems. This means that processes need to be transparent, the capabilities and purpose of AI systems openly communicated, and decisions—to the extent possible—explainable to those directly and indirectly affected. […] The degree to which explicability is needed is highly dependent on the context and the severity of the consequences if that output is erroneous or otherwise inaccurate.' Crucially, this description refers not only to the explicability of the algorithm itself, but also to processes in which algorithms are used, the capabilities and purposes of systems in which algorithms are used, and to communication about these processes, capabilities and purposes. We saw this broader perspective also under procedural fairness. Furthermore, it is critical to recognize that algorithms can vary regarding their inner workings and in their associated explicability. One can imagine a spectrum, with on the one end a simple decision tree ('if x, then y'), with a handful of branches with clear and stable cut-off values, so that its inner working is transparent and easy to understand or explain, and on the other end a complex neural network ('deep learning') with thousands of variables that interact with each other dynamically, so that its inner working is opaque and hard to understand or explain.

The *High-Level Expert Group* also discuss tensions between these principles, e.g., between prevention of harm and respect for human autonomy. They mention 'predictive policing' and argue that it may help to reduce crime, and also that it may bring risks related to individual liberties and privacy (2019, p 13). This means that different values need to be carefully considered and balanced.

4 Algorithms for decision support in government

Before we move to the case studies, we need to provide some context for the design and deployment of algorithms in decision support by the government in The Netherlands.

Over the last 3 years (2018–2020) The Netherlands ranked 4th in the yearly European *Digital Economy and Society Index* (DESI),[6] which can be seen as an indicator of digital maturity. A recent strategy for digital innovation in the Dutch public sector advocates protecting fundamental rights, increasing accessibility and making personalized services,[7] and the Dutch government presented a strategic action plan for the application of AI.[8] Interestingly, the actual usage of algorithms for decision support in the Dutch government is currently rather limited; most instances are in experimental phases, rather than in implementation phases [44].

In the domain of justice and security, as in other sensitive domains, such as healthcare and education, the government approaches the design and deployment of AI with utmost care. This, however, brings them in a catch-22 position. Socio-technical trends lead to expectations that government agencies use state-of-the-art technologies; at the same time, these agencies are bound by legal and moral boundaries for using novel technologies. There have been a series of incidents in The Netherlands, in which political and societal forces pushed back on the application of algorithms, insisting on the protection of fundamental rights. One case received international attention: the court decision, in February 2020, to forbid usage of the SyRI algorithm, which the Ministry of Health, Welfare and Sport used to detect citizens' welfare fraud.[9] Such pushback motivated the Dutch government to produce a *Toolbox for Ethically Responsible Innovation,*[10] *to* help navigate between innovation and experimentation, and protecting public values and fundamental rights, like human dignity, human autonomy, fairness and privacy.[11] Moreover, a series of case studies of legal aspects algorithms in decision making was recently published [24].

5 Case studies

We selected two cases[12] for further study. Both deal with the design and deployment of an algorithm in decision support in the domain of justice and security. In both cases the

[6] https://ec.europa.eu/digital-single-market/en/scoreboard/netherlands. The index comprises measures for connectivity, human capital, use of internet services, integration of digital technology, and digital public services.

[7] https://www.nldigitalgovernment.nl/digital-government-agenda/.

[8] https://www.government.nl/documents/reports/2019/10/09/strategic-action-plan-for-artificialintelligence.

[9] https://www.rechtspraak.nl/Organisatie-en-contact/Organisatie/Rechtbanken/Rechtbank-Den-Haag/Nieuws/Paginas/SyRI-legislation-in-breach-of-European-Convention-on-Human-Rights.aspx. Other recent examples of push back against algorithms in government agencies involve the Tax and Customs Administration (Belastingdienst) and the Employee Insurance Agency (UWV).

[10] https://www.digitaleoverheid.nl/overzicht-van-alle-onderwerpen/nieuwe-technologieen-data-en-ethiek/publieke-waarden/toolbox-voor-ethisch-verantwoorde-innovatie/ (in Dutch).

[11] https://www.digitaleoverheid.nl/overzicht-van-alle-onderwerpen/nieuwe-technologieen-data-en-ethiek/publieke-waarden/ (in Dutch).

[12] Our approach consists of studying two instances of a 'revelatory case', which refers to having the opportunity 'to observe and analyse a phenomenon previously inaccessible to scientific investigation' [46], p 40.

people involved aim to contribute to a societal or common good, and in both cases the people involved act carefully[13]:

A. *Fine collection by phone to prevent debts*, by the Central Judicial Collection Agency
B. *Risk Assessment of 'violent behaviour'*, by the National Police

Our case studies are based on interviews with the people involved in these initiatives, and on analyses of both internal documents and publicly available documents. Below, we first discuss the algorithm's goals and the current situation. We follow the *High-Level Expert Group*'s four principles: respect for human autonomy; prevention of harm; fairness; and explicability (2019). With regards to *human autonomy* we focus on the role of civil servants who *use* algorithms, rather than on citizens as implicated actors ('data subjects') [20]. The perspectives and experiences of citizens are paramount in our discussions of *preventing harm* and of *fairness*. Our discussions of explicability deal with perspectives of both civil servants *and* citizens; both are involved in understanding and explaining the algorithm at hand.

Please note that we will *not* discuss privacy and data protection. These issues are, of course, at play. We chose, however, *not* to discuss these because of two reasons: these issues would take us deep into legal territory, away from ethical issues; and, in these cases, the people involved do take care of privacy and data protection; they are well aware of relevant legislation and comply to it.

6 Case A: fine collection by phone to prevent debts

The Central Judicial Collection Agency (CJIB; part of the Dutch Ministry of Justice and Security) is responsible for collecting a range of different fines, such as traffic fines and punitive orders. The CJIB aims to collect fines *responsibly*; to balance *doing justice to vulnerable citizens* and *following the rule of law*. Our case study focuses on one initiative: to approach selected people by phone to remind, motivate or enable them to pay their fines and thus prevent them from falling (further) into debt.

Suppose, e.g., that one receives a speeding ticket of €100 and fails to pay.[14] The CJIB is then legally required to send an exhortation of €150. If one again fails to pay, another exhortation is sent, of €300. If one still does not pay, the CJIB can send a bailiff, the costs of which will also need to be paid. A fine of €100 can thus multiply to hundreds of euros. Still failing to pay the fine may result in coercive measures, such as seizure of one's car or, in extreme cases, imprisonment for (a maximum of) seven days.

For this initiative, the CJIB developed an algorithm that uses data *from only the CJIB* (the current fine, other open fines, and payment history; no data from criminal law or administrative law are used) to produce a list of people who would probably be *willing* to pay and can be *incentivized*, by providing them a reminder or a payment provision.[15] The fact that they use only their own data is noteworthy and commendable; it sets them apart from notorious cases of using data from multiple sources (e.g., [32]. Agents of the CJIB interpret the algorithm's output and decide whom to call from this list they make phone calls to individuals to discuss the situation and offer provisions, if needed.[16] Preventing these people from falling (further) into debts is beneficial to them, to their family members and friends, and to society at large.

6.1 Respect for human autonomy

With regard to respect for human autonomy, we will focus on the autonomy of the people who make the phone calls to these citizens. As noted (above), we do *not* focus on the autonomy of the citizens who are at the receiving end of this initiative. However, this initiative's aim is to support citizens in maintaining a certain level of financial autonomy. In that sense, respect for *citizens'* autonomy drives this initiative.

[13] These criteria are phrased rather informally. If you imagine a horizontal axis of possible cases, with on left projects that aim to *extract value*, e.g., from the environment, from workers, customers or citizens, at their expense, and in which the people involved do *not engage* in any ethical deliberation, and on the right projects that aim to *create promote positive values*, e.g., promote people's wellbeing and justice, and in which the people involved act conscientiously and carefully, then the cases we selected are on the right side of that axis.

[14] What follows is a simplified version of the process; in reality, there are many details and nuances, e.g., regarding the ways in which judges weigh perspectives and interests before they decide to use coercive measures.

[15] See: https://www.cjib.nl/en. The CJIB uses a matrix with four quadrants: one axis distinguishes between people who are *willing or unwilling* to pay; another axis distinguishes between people who are *able or unable* to pay. Our case focuses on people who are *willing* to pay and are *able or currently unable* to pay. For people who are *unwilling* to pay, several different law enforcement measures are available, which are outside our current scope.

[16] When it becomes clear that a person has problems with debts, the CJIB sends their name to respective agencies on the level of municipalities; they then approach these people to provide follow-up actions and support.

The agents who make these calls received a training program,[17] which emphasized that the call should not only focus on debt collection, but also make room for offering help in debt relief. At the moment, these agents' autonomy is not strongly affected by the system because they have professional competence to follow or not follow the list with names of people to call. In addition, the agents can ignore the algorithm's outcomes and they can add information to specific cases, e.g., specific situation about cases' context. In other words, the algorithm combines *computer automation* and *human control* [38].

6.2 Prevention of harm

The system's main goal is to identify and approach citizens who are at risk from going into (further) debts and offer them reminders or payment provisions. The algorithm will, however, produce not only *true positives* (people to approach), *true negatives* (people *not* to approach), but also *false positives* and *false negatives*. *False positives* refer to calling people and then finding out that they will *not* pay, which can be seen as a waste of the agents' time; a sort-of harm to the CJIB and its purposes, because time wasted in one place cannot be put to good use elsewhere. Furthermore, *false negatives* refer to people *not* being called, and therefore missing out on reminders or payment measures, which are actually in need of. They miss out on a service that was intended to benefit them. At the moment, the CJIB tends to find false negatives more problematic than false positives; their aim is to approach and support people in paying their fines and to 'better safe than sorry' (to accept false positives).

Both types of errors are hard to prevent from happening. What an organization can do, to reduce errors over the course of time, is create *feedback loops* that enables the people who use the algorithm's output and make the phone calls to report errors to the people who develop and control the algorithm, so that they can make corrections or modifications. The CJIB has created such feedback loops, including efforts to examine and understand *negatives*; to find out whether these are true negatives or false negatives.

6.3 Fairness

The algorithm does not lead to a differences in legal treatment of citizens; it is legally fair. Delving one spade deeper, however, we can discuss the fairness of treating different people differently. The algorithm points at people who may need incentives or support to pay their fines. At the moment, this service is perceived as additional, not as essential. It remains to be seen, however, how people, over the course of time, will perceive this; how will they perceive *false negative* errors (maybe as missing out on an essential service?) and *false positive* errors (maybe as a nuisance and a waste of time?).

Another dimension of fairness refers to procedural fairness. Are people at the receiving end of this process able to contest and seek effective redress? At the moment, the CJIB has various processes in place, via which people can express disagreement or complain about the process.[18] They are currently looking into ways to enable citizens to make legal objections. They want to have that in place before implementation.

6.4 Explicability

To promote the algorithm's explicability, i.e. its understandability and explainability, for audiences both inside and outside the organization, the CJIB put several measures in place. First, a general description of their initiative is publicly available, online.[19] In addition, the CJIB chose to use a relatively simple *classification* model[20] instead of, e.g., a relatively complex algorithm, such as *random forest,* which combines multiple regression analyses.[21] The former is like a decision tree, with stable and transparent cut-off points (e.g., 'if total amount of fines is higher than x, then offer measure y'); in contrast, the latter, more complex types of algorithms, can change over time and is therefore less transparent.

Those more complex algorithms can fall short regarding explicability; it can be hard for programmers to inspect and modify them, and it can be hard for agents to understand and explain them. On the other hand, these more complex algorithms *can* perform better than the simpler ones. Looking ahead, the CJIB are exploring ways to improve the algorithm's outcomes, e.g., using specific, trustworthy data sources outside the CJIB,[22] or by a using a *random forest* algorithm that is evaluated and updated every six months.

[17] Training program 'Motiverend Incasseren' ('Motivational Collection'), developed by Nadja Jungmann [22].

[18] See: https://www.cjib.nl/en/i-disagree-my-fine and https://www.cjib.nl/ik-heb-een-klacht (in Dutch).

[19] https://www.cjib.nl/innen-incasseren.

[20] https://www.wodc.nl/onderzoeksdatabase/2947-regulering-van-algoritmen-die-zelfstandig-besluiten-nemen.aspx.

[21] Using *one* regression analysis (*one* tree) means that one set of variables may influence the decision unevenly, because the analysis optimizes for this set of variables, which poses risks of 'overfitting' or 'variable bias'. In contrast, using a *random forest*, which combines multiple regression analyses, entails different variables, which can minimize these risks.

[22] See for instance: https://www.cjib.nl/nieuws/cjib-onderzoekt-mogelijkheden-om-oplopende-schulden-te-voorkomen.

They want to combine maintain or even improve explicability and improve performance—and, of course, keep working in compliance to legislation.

7 Case B: risk assessment of violent behaviour

Since a number of years, the Dutch Police has embraced a data-driven approach [7]. Their ambition is to utilize data they *already have* collected during police work, notably the data in their *incident registration system* (but *not* to collect more data on citizens), following principles of legality and proportionality. We will focus on one initiative: the development of an algorithm that assesses the likelihood of violent behaviour of specific people.[23] Such assessments can help police officers to work more effectively and efficiently; e.g., approach potentially violent people carefully and appropriately. For the design and deployment of this algorithm, they use not only *structured* data in their system, e.g., a code for a specific type of incident, but also *unstructured* data, e.g., notes made by police officers; these typically contain more specific and detailed information.

The algorithm was developed mainly using experts' knowledge. Domain experts generated a series of terms that relate to violent behaviour, e.g., *confused, crisis, psychotic* and *alcohol*. They chose *not* to use machine learning to populate or expand this list of terms. They used these terms to build an algorithm that analyses police reports associated to a specific person (using data that cover a period 5 years), and produces a 'score' to represent the likelihood of this person expressing violent behaviour. They then invited another group of domain experts to test and evaluate the algorithm. This enabled them to modify and fine-tune the algorithm, e.g., assign different weights to different terms.

7.1 Respect for human autonomy

The police envision two use cases for this algorithm: in 'slow' or 'cold' processes, e.g., to study violent behaviour as a phenomenon and to generate information for briefing meetings at the police station; or in 'fast' or 'warm' processes, e.g., while police officers are dealing with an incident, 'on the ground'. For now, they envision using the system in 'slow' or 'cold' processes. For the future, however, they envision using it also in 'fast' or 'warm' processes. For example: there is a report of an incident and police officers hurry to the incident; meanwhile, their colleagues at the Real-Time Intelligence Centre (RTIC) use various sources, *not only* the algorithm's output, to assess risks and provide advice to the police officers. This way, there is a 'human in the loop': people at the RTIC use their professional discretionary competence to interpret the algorithm's output. The goal of the RTIC is to have sufficient, relevant and reliable information available to handle the situation at hand—*not* to automize decision making.

This 'human in the loop' approach is motivated by a respect for human autonomy of RTIC operators and of police officers on the street. Both can use their professional discretionary competences and combine the algorithm's output, relevant protocols and their own, professional judgement and discretion. This way of deploying the algorithm combines *computer automation* and *human control* [38].

7.2 Prevention of harm

Overall, the system is meant to prevent harm: to identify people who may behave violently, so that police officers can approach them appropriately and prevent that person from causing harm to other people or to themselves. There are protocols for police officers to prevent risks of bias, stigmatization or discrimination in their engagements with citizens, and the police are continuously working on these issues. Moreover, one may argue that *not* utilizing data that the police already have can cause more harm. *Not* utilizing these data may lead to police officers approaching people unaware of their potential to behave violently, which may cause more harm.

In addition, the division of labour between people at the RTIC and police officers on the street can help to prevent or mitigate risks for misusing or overusing the system. The people at the RTIC dedicate their energies at interpreting various sources of data, including the algorithm's output; they are enabled to reflect on their usage of the system, which prevents them from misusing or overusing it. At the same time, the police officers on the street can focus on the situation at hand and the people they interact with; they do not need to bother with the algorithm and there is little risk that they misuse or overuse the system.

7.3 Fairness

The developers chose to follow a *cross-industry standard process for data mining* (CRISP-DM), to work systematically and carefully. This entails organizing an iterative process for modelling, development, deployment and evaluation, and processes to better understand both the business in which the algorithm is deployed and the data that are used. This process enables the people involved to create feedback

[23] The initiative is part of a larger program to develop an architecture for diverse types of risk assessment; the goal is to align existing models and algorithms, and to develop new models and algorithms within one architecture.

loops that can help to identify incorrect or unfair outcomes of the algorithm and to mitigate these.

At the moment, the system is in an experimental phase; it is not yet operational. Questions regarding *procedural* fairness are therefore currently open. Such questions relate, e.g., to citizens' abilities to question, contest and seek redress against decisions that are based on a combination of the outcomes of the algorithm and the interpretations of the people who use the algorithm.

7.4 Explicability

The choice to start the development of the algorithm with knowledge of domain experts yields promotes the system's explicability. These experts produced a list of terms, that others can inspect and modify if necessary. Furthermore, the algorithm is transparent in the sense that it uses an explicit and limited list of terms and weights, which can be inspected. Alternatively, they *could* have chosen to use, e.g., a *deep learning* algorithm that is fed with labelled data so that it can 'learn'. These algorithms are notoriously less transparent; it is often and typically hard to inspect variables and weights in these algorithms, which makes them hard to understand and explain, and hard to correct or modify.

In the current set-up, police officers on the street do not need to understand or explain the algorithm to, e.g., citizens whom they are engaging with. They receive information from the RTIC and combine that with their professional perception and judgement.

8 Summary and limitations

These two cases by and large follow the ethical principles that the High-Level Expert Group on AI (2019) put forward: respect for human autonomy; prevention of harm; fairness; and explicability.

The CJIB agents who make the phone calls (Case A), and the RTIC operators and the police officers (Case B) are required to use their professional discretion, which respects *human autonomy*. Both cases have as primary aim to *prevent harm*: to prevent people from falling (further) into debt (Case A) and to approach citizens with appropriate care (Case B). Both cases aim to uphold and promote *fairness*; we discussed various measures to maintain substantive fairness. It is, however, too early to discuss procedural fairness in full detail, since both cases are currently in experimental phases, and processes are under development. Regarding *explicability*, we saw explicit design choices for algorithms and ways of working that are understandable and explainable: a simple standard classification model instead of, e.g., a complex random forest algorithm (Case A); and an algorithm that starts with experts' knowledge rather than with, e.g., machine learning (Case B).

Our case studies have several limitations. First, they deal with algorithms that are *currently* being used in Dutch government agencies. So, the algorithms we found and studied are relatively simple [44], compared to systems with advanced algorithms, e.g., with deep learning or reinforcement learning. We did, however, discuss potential, future developments of more advanced algorithms, e.g., with regards to their explicability. In addition, our focus on algorithms that are used by the Dutch government meant a focus on government agencies and civil servants. We can assume that they have fairly benign intentions and fairly careful ways of working. Our findings could have been rather different if we had studied algorithms that are being developed and deployed in another domain, e.g., a domain driven by short term financial profits or a company with little interest in social responsibility.

Furthermore, the algorithms studied were in early and experimental phases [44]. As a consequence, the empirical parts of our case studies focused on impacts that are closely related to the design and application phases. We did, however, explore less immediate impacts, which may happen after some time. Our explorations are similar to the cases discussed by O'Neil (2016), e.g., how the deployment of algorithms may propagate and exacerbate existing inequalities and harm those with less power disproportionally. In general, it remains challenging, however, to anticipate higher-order impacts, like the corrosive effects that online social networks during can have on people's news consumption, on political polarization and on behaviours during elections.

9 Unintended, undesirable, higher-order effects

Above, we looked at two cases in which the people involved aim to contribute to a societal or common good and work carefully. They followed principles like respect for human autonomy, prevention of harm, fairness and explicability [21]. This can be challenging and complex enough. Below, we will explore one further level of complexity, namely potential *unintended, undesirable*, higher-order effects of using algorithms.

Such effects can occur, regardless of good intentions and a careful approach. That is why we call them 'unintended, undesirable'. We propose that such an exploration is needed if we take Responsible Innovation (RI) seriously. Stilgoe et al. [40] argued that RI entails four key dimensions: anticipation, responsiveness, inclusion and reflexivity.

We will focus on the first two: *anticipation* and *responsiveness*.[24] 'Anticipation', they write, 'prompts researchers and organisations to ask 'what if...?' questions …, to consider contingency, … [it] involves systematic thinking aimed at increasing resilience, while revealing new opportunities for innovation' (*op. cit.*: 1570). Anticipation involves exploration and speculation, envisioning various potential scenarios of what *might* happen. Not only of problems, by the way, also of opportunities. Regarding responsiveness, Stilgoe et al. comment that it refers to 'a capacity to change shape or direction in response to stakeholder and public values and changing circumstances' (*op. cit.*: 1572). Responsiveness is a necessary supplement to anticipation; one needs to be able to *respond* to issues that one anticipates. In addition, responsiveness requires anticipating changing circumstances to be able to respond adequately.

For our exploration of potential *unintended, undesirable*, higher-order effects, we will stay in the domain of algorithms that are used for decision support in the domain of justice and security. Let us further assume that the people involved in the design and deployment of these algorithms have good intentions and work carefully. They will typically focus on *intended* and *desirable* effects that the algorithm can help to realize. They will also make efforts to anticipate unintended, undesirable, *first-order* effects that can happen, e.g., when different values conflict in a relatively direct manner. When it is obvious that a lack of transparency can negatively impact fairness, people will work to improve transparency, to promote fairness.

They will, however, by definition, find it hard to anticipate unintended, undesirable, *higher-order* effects.

Now, what do we mean with *higher-order* effects? One way to understand these comes from *systems thinking*. In *systems thinking* one views different phenomena as parts of a larger system and looks at the relationships between these phenomena [27]. The feedback loop is a key concept here: imagine that A influences B, then information about the status of B regulates the influence of A on B. Feedback loops can be *balancing*; they steer parts of a system to some dynamic equilibrium. Think of the balancing feedback of a thermostat in a heating system. Or they can be *reinforcing*; they make parts of a system go increasingly up or increasingly down. They are sometimes referred to as virtuous cycles or viscous cycles. We are often able, to some extent, to anticipate *first-order* effects, like the influence of A on B. We are, however, much less able to anticipate *higher-order* effects, like the behaviour of a system that has multiple elements, multiple relationships and multiple balancing and reinforcing feedback loops.

We can illustrate what we mean with *unintended, undesirable*, higher-order effects with the anecdote of the *Cobra effect*.[25] In the time of the British rule of colonial India, the British wanted to get rid of venomous cobras in Delhi and offered a bounty for every dead cobra. The (first-order) intended and desirable effect was that people killed snakes for this reward. People, however, also began to breed cobras to claim their rewards. The government found out and stopped the reward program. The breeders then let their cobras go free, which were now worthless—which worsened the cobra plague. Officials could maybe have anticipated this effect if their analysis of the system had included variables and feedback loops for supply and demand, and for motivation and behaviour.

Now, if we envision a situation in which people design an algorithm. They focus on values A and B, and carefully combine and balance A and B. But then, after the system has been in use for a while, value C pops up, in a very disturbing way, seemingly out of nowhere. Or they create a careful balance between values P and Q, and then, as people use the system in ways slightly different from what they had intended, value Q goes off the rails, unexpectedly, and the balance between P and Q is gone.

A similar effect is known as Goodhart's law,[26] named after economist Charles Goodhart: 'When a measure becomes a target, it ceases to be a good measure.' Informally, this is known as the *KPI effect*: when an organization introduces a *Key Performance Indicator*, the people on the floor find (creative) ways to satisfy this KPI—sometimes, however, in ways that hamper what the KPI tries to achieve. The organization meets its KPIs, but fails to realize the underlying goals. This draws attention to the need to be very careful when articulating, quantifying and measuring the intended, desirable outcomes one wishes to achieve.

Below, we will explore several potential unintended, undesirable, higher-order effects that may occur in the design and deployment of algorithms for decision support in the domain of justice and security. Our exploration is based on several general findings from the cases discussed above. (Please note, however, that our exploration is *not* intended

[24] Our choice to focus on these two of the four dimensions is mainly pragmatic. Our case studies would have doubled in size if we had included *inclusion and reflexivity.* Also, we assessed that we would be better able to study *anticipation* and *responsiveness* from an outsider perspective and with a descriptive approach, e.g., by looking at the outputs that the people in the case studies produced, compared to *inclusion* and *reflexivity*, which would have required an insider perspective and a more participative approach, e.g., by attending meetings and involving the people in the case studies in our research. Topics like *inclusion*, diversity and gender are, of course, very relevant and topical indeed, with the recent firings of Timnit Gebru and Margaret Mitchell by Google (https://www.theguardian.com/technology/2021/feb/19/google-fires-margaret-mitchell-ai-ethics-team; https://www.washingtonpost.com/technology/2020/12/23/google-timnit-gebru-ai-ethics/).

[25] https://en.wikipedia.org/wiki/Cobra_effect.

[26] https://en.wikipedia.org/wiki/Goodhart's_law.

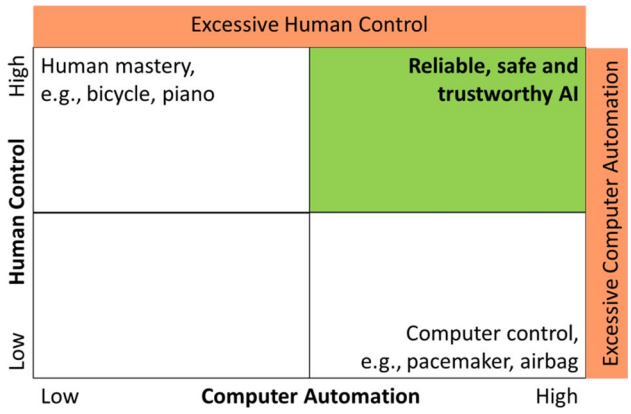

Fig. 1 'Reliable, Safe, and Trustworthy' AI requires appropriate levels of computer automation and human control (adapted from [38]

as an assessment of what might happen in these particular cases.) We will, again, follow the principles of the High-Level Expert Group on Artificial Intelligence [21].

9.1 Respect for human autonomy

The deployment of algorithms can alter people's ways of working and thus affect their autonomy. In the introduction we referred to algorithms in decision support, where agents need to combine the algorithm's output with their professional discretion. The extremes of *always* strictly or slavishly following the algorithm and of *always* neglecting or overruling its output are rather ineffective or inefficient.

We may be able to anticipate unintended, undesirable, *first-order* effects and respond appropriately, e.g., by enabling people to inspect, question, modify or correct its functioning. But how may we anticipate and respond to *higher-order* effects? Shneiderman [38] provided a diagram that may be helpful—see Fig. 1. He advised exploring the top-right quadrant of *high computer automation* and *high human control* to create 'reliable, safe and trustworthy' AI systems (although there may be good reasons, in specific cases, to go for other quadrants: for example the combination of high human control and low computer automation for piano playing ('human mastery') or the combination of high computer automation and low human control for airbags ('computer control').

One way to explore potential, unintended, undesirable, *higher-order* effects, is to explore various ways in which the system and its usage may, over time, unintentionally, move across the plane in Fig. 1, away from the top-right quadrant. It may drift to *excessive* human control, where people need to micro-manage the system, which could be very inefficient, or even dangerous, e.g., in a situation in which a self-driving car very suddenly requires the driver, who is busy doing something else than driving, to take control of the steering wheel. Or towards *excessive* automation, where too many tasks are delegated to the system, so that people can no longer monitor its functioning in any meaningful way, or the system performs tasks that do require human perception, discretion and judgement. Computers are notoriously bad at the latter; they cannot take into account *context* and they lack *common sense* [26, 35].

Or the system may drift away towards *too little* human control; this can also happen because of people's evolving practices, e.g., when people have learned to follow the algorithm's output unthinkingly and routinely 'click the okay button'. Or it may drift towards *too little* computer automation; people then need to perform too many routine tasks and effectively waste their time and energy—or worse, start to make mistakes, e.g., because of reduced concentration.

Another issue regarding respect for human autonomy is the combination of *explicit knowledge* and *tacit knowledge*. The former refers to data that is used as input for an algorithm and is associated with computer automation. The latter refers to information in people's minds and bodies, which is associated with human control. Both types of knowledge are relevant for algorithms in decision support. Imagine an organization that procures such a system. They will typically want to realize benefits that outweigh the costs associated with using the system. Over time, they may unintentionally slide towards preferring *explicit knowledge* and *computer automation* over *tacit knowledge* and *human control*—it would be silly to buy an expensive system and not use it. There is ample (anecdotal) evidence of people feeling unhappy when their tacit knowledge, their abilities, skills, expertise, experience, are not valued and replaced by automation. This may even lead to an unintentional, undesirable focus on means, and losing sight of ends. Choosing for automation brings risks for the unintended and undesirable effect—over the course time—of prioritizing explicit knowledge and computer automation at the expense of tacit knowledge and human control.

9.2 Prevention of harm

A first step in anticipating and preventing potential harms involves assessing and evaluating the different pros and cons of *using* an algorithm (a future situation) in comparison to *not using* an algorithm (the current situation). It is indeed possible that *not using* the algorithm causes more harm than using it. This would be an argument in favour of deploying this algorithm. In addition, one would need to design and deploy measures to increase its benefits and to decrease its drawbacks.

Another way to anticipate and prevent potential harms of algorithms, is to create an error matrix. Such a matrix plots true positives, true negatives, *false positives* and *false negatives*. These errors can be viewed as *first-order* unintended,

undesirable effects. One way to anticipate *higher-order* unintended, undesirable effects is to explore how this error matrix may evolve, as it is modified and fine-tuned over the course of time, either by people or 'by itself', in cases of *machine learning*.

For example, there may be unintentional incentives to promote the occurrence of *false negatives*. If we look back at Case A, this would refer to a person *not* being offered support to pay their fines, whereas this person would *actually* need such support. For the sake of argument, let us assume, in more general terms, that *false negatives* refer to advice that 'no action' is needed, which typically costs less money and time than action. Moreover, *false negatives* are likely to stay undetected; they typically do not appear in weekly, quarterly or yearly reports. This may unintentionally nudge the organization, over time, to modifying the system towards producing more *false negatives*—which may hamper the organization's overall goals.

Alternatively, a system may unintentionally evolve, over time, towards producing more *false positives*. If we look back at Case B, this would refer to a person *incorrectly* receiving a high likelihood of violent behaviour, and an advice to act cautiously and carefully. For the sake of argument, let us assume, in more general terms, that *false positives* entail taking action, which costs money and time. Organizations typically steer away from costs and, unintentionally and over time, may modify the system towards producing less false positives. Having fewer errors does not need to be problematic, of course. It can, however, be a problem if the modified algorithm's reduction of false positives leads to more false negatives. In Case B, this would refer to predictions of non-violent behaviour for people who will actually behave violently—an unintended and undesirable effect.

There remain questions regarding a fair balance between having *false negatives*, which may hamper the organization's main goals, or *false positives*, which may involve wasting money and time, and harms.

One particular example of an unintended, undesirable, higher-order effect could be a drift towards 'low hanging fruits'. An organization may gain insights in 'what works best' and drifts towards prioritizing cases that are *very clearly* true positive. In Case A, this would refer to people who are very willing and very able to pay their fines. One might say that they do not *really* need to be offered support. The organization, however, can be very successful if it targets them. Such a priority for 'low hanging fruits' may lead to a neglect of people who are less clearly *true positive*, who will miss out on the support they actually need.

Another higher-order, unintended harm can manifest when organizations collect and use data from multiple data sources, especially if these data sources pertain to different domains. Imagine an insurance company collecting data about their customers' life styles. Or a care provider collecting data on their patients' finances. Combining data from different sources is not necessarily always a bad idea. There are situations in which the public expects that different public service organizations collaborate and share information—of course following principles of legality and proportionality. The public will criticize the organizations involved if *not* collaborating and *not* sharing information resulted in harm that could have been prevented precisely by collaborating and sharing information.

9.3 Fairness

It almost goes without saying that we expect government organizations to comply to legislation and to ensure *substantive* fairness, an 'equal and just distribution of both benefits and costs' and *procedural* fairness, which refers to people's abilities 'to contest and seek effective redress' [21], p 12). Making the algorithm *substantively* fair is necessary but not sufficient. The organizations involved will also need to organize *procedural* fairness, e.g., by organizing processes via which people at the receiving end of the decisions supported by the algorithm, are able to critique these decisions. We will further discuss this topic below, under *explicability*.

Our discussion of harms (above) focused on individuals. Harms can, however, also affect *groups* of people. In such cases, we can view these harms as systemic *unfairness*. There are ample examples of unfairness being repeated, propagated or exacerbated through the usage of algorithms [32]. Imagine an algorithm that puts a specific label on a relatively large number of people in a specific socio-economic or cultural group, then this may lead to stigmatization or discrimination of that group. There is a risk that people are reduced to labels and that the labels get reified. In very general terms, we can point at four sources for such unfairness:

- the data that the algorithm uses as input—these data may refer to current unfair situations; when these data are used to train an algorithm, these unfair situations are likely to be repeated;
- the algorithm itself, which can function unfairly—intentionally or unintentionally;
- the process in which the algorithm's output is used—which affects *procedural* fairness;
- or the feedback loop, which feeds back information about the application of the algorithm's output, so that the algorithm, or processes around it, can be corrected and modified.

This feedback loop is critical. 'Without feedback', Cathy O'Neil argued, 'a statistical engine can continue spinning out faulty and damaging analysis while never learning from its mistakes' (2016: p. 7). She stressed the need for properly functioning feedback loops, otherwise we risk 'confusing

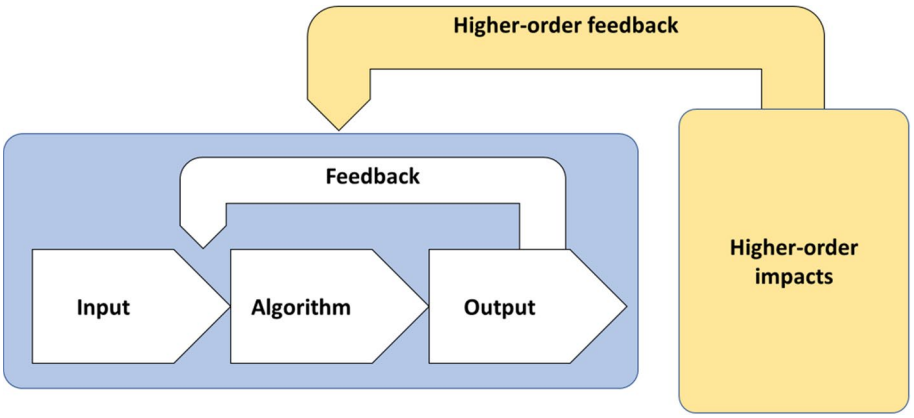

Fig. 2 Framework for promoting anticipation and responsiveness regarding potential unintended, undesirable, higher-order effects of algorithms

[algorithms'] findings with on-the-ground reality' (*ibid*: p. 12).

Finally, we must look at the larger picture. Promoting fairness in the design and deployment of algorithms must go hand in hand with questioning and critiquing the larger context in which these algorithms are used [3–5]. For example, in the infamous COMPAS case,[27] of an algorithm that assesses the likelihood of recidivism, one needs to make the algorithm more fair (or less unfair), but also make room to question and critique the role of racial discrimination in the judiciary system, and the larger systemic, racial inequalities and injustices in society.

9.4 Explicability

Explicability can be understood as having *instrumental* value in that it contributes to other values or principles. According to Hayes et al. [20], explicability (or in their words: 'accountability/transparency') contributes to *autonomy* and to *fairness*. Autonomy, of both those who use the algorithm ('human decision makers') and those at the receiving end ('data subjects'), critically depends on their abilities to understand and explain the algorithm's functioning. Moreover, explicability is critical for people's abilities to question and critique the algorithm's fairness: to find an appropriate balance of agency between people and technology; to inspect and evaluate the various types of errors; and to organize processes via which people can critique and provide pushback, and seek correction and redress. There are several domains of knowledge dedicated to promoting explicability, e.g., XAI (Explainable AI) and FAT (Fairness, Accountability and Transparency; which focuses on more than explicability)—a discussion of which is outside our scope.

Two issues, which we also encountered in our cases, are, however, worth mentioning. First, there is the 'problem of many hands' [8]. This refers to the problem that in a complex system, with many actors and many moving parts, it can be hard to attribute responsibility. If we want to explore unintended, undesirable, higher-order effects, we need to look at the larger processes in which algorithms are used, at organizations that use the algorithms. A decision based partially on an algorithm's output can only be understood and explained if the processes and organization are understandable and explainable. One will need to avoid situations where citizens' questions get a reply like: 'Computer says no. I don't know why. You will need to go elsewhere. I don't know where.'

Second, there are different types of algorithms with different properties and levels of explicability. In the two cases, we saw that the people involved chose for a simple decision-tree rather than a complex deep-learning (Case A), and for using 'expert knowledge' rather than 'data mining' (Case B); in both cases, they chose the former because it typically provides better explicability than the latter. Looking forward, it would be wise to keep an eye on developments in FAT and XAI; these fields may provide solutions that combine autonomy, fairness, accuracy and privacy. Organizations need to explore, innovate, experiment and learn.

10 Discussion and conclusion

We looked at two positive examples of using algorithms in decision support systems in the domain of justice and security. We discussed the ways in which these initiatives followed the principles put forward by the High-Level Expert Group on Artificial Intelligence [21]: respect for human autonomy, prevention of harm, fairness, and explicability. We then explored potential, unintended, undesirable higher-order effects. We identified and discussed a range of such effects—*which might occur, despite good intentions of the people involved and their careful approaches*. Based on our exploration, we speculate that, in more general terms, such effects become, to some

[27] https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm.

extent, better accessible for anticipation and responsiveness, if we *zoom-out*. Zooming-out refers to both space and time: we need to zoom-out to see the algorithm not in isolation, but within its context, within the process and organization in which it is used; and we need to zoom-out to envision the algorithm and its usage develop over time, while it is deployed in practice.

We can visualize the main findings of our exploration—see Fig. 2. The blue box represents the algorithm's functioning and immediate effects and feedback loop; this is typically 'in scope' of the people involved in design and implementation. The orange box and arrow represent the usage of the algorithm's higher order impacts and higher-order feedback loops; these are not always 'in scope' and are likely to happen over the course of time and therefore not easy to anticipate and respond to. This figure meant as a tentative framework to explore higher-order effects—as a reminder for people involved in the design and deployment of algorithms to follow the four principles put forward by the *High Level Expert Group on AI* (2019): respect for human autonomy, prevention of harm, fairness, and explicability.

Here are several tentative recommendations, based on the findings from our exploration—needless to say: there may be many other issues in other cases, which may lead to different recommendations:

- Input: one can promote *fairness* by preventing *bias*, both in the data that are used in training the algorithm and in the data that the algorithm uses in deployment; one can promote *explicability* by being transparent about the data that go into the algorithm; more generally, one needs to look at the larger picture of data collection, e.g., which data sources are used and which are *not* used—it is wise to question assumptions about the system's boundaries: which data are included and which are excluded;
- Algorithm: one can promote *respect for human autonomy* by considering to use experts' knowledge in the design of the algorithm, rather than machine learning; one can promote *prevention of harm* by analysing false positives and false negatives and their respective harms; one can promote *explicability* by carefully choosing a specific type of algorithm (they can greatly differ regarding explicability); and by explaining the algorithm's functioning in a vocabulary that is appropriate for the addressee;
- Output: one can promote *respect for human autonomy* by enabling the people who use the algorithm's output to use also their professional discretion, in combination with the algorithm's output; promoting such human autonomy in the process ('human in the loop') is also critical to *prevent harm, to promote fairness, and to promote explicability,* assuming that people are able to check and mitigate harmful outcomes and impacts of the algorithm.
- Feedback: one can *prevent harm* and promote *substantial fairness* by organizing processes around the algorithm's deployment that enable both agents ('human decision makers') and people at the receiving end ('data subjects') [20] to provide pushback, i.e. to make corrections or modifications, when necessary; this needs to goes hand in hand with promoting *procedural fairness* and *explicability*, e.g., by organizing processes that enable agents and citizens to engage in a fruitful dialogue, if needed.
- Higher order impacts: one can monitor potential, unintentional or undesirable effects as they happen over the course of time. This would require zooming-out to see not only the algorithm's immediate effects, but also the effects it has on processes in the wider organization that deploys the algorithm, and the broader impacts this has in society.
- Higher order feedback: one can promote anticipation and responsiveness by putting mechanisms in place that feedback information on these higher-order impacts to the organization that deploys the algorithm. One can do this, e.g., by organizing continuous improvement (CRISP-DM) or frequent, critical reviews of objectives and realized outcomes.

To some extent, such high-order, unintended, undesirable consequences *can* be anticipated, e.g., by exploring possible future scenarios [31, 34]. Anticipation, however, remains notoriously difficult [43]. In addition, one can move to *responsiveness*, 'a capacity to change shape or direction in response to stakeholder and public values and changing circumstances' [40], p. 1572. The time dimension is critical here since higher-order effects typically happen over the course of time. One way to promote responsiveness is by organizing small-scale experiments, e.g., 'testing zones' or 'living labs', to try out and evaluate technologies and applications. These experiments can involve diverse actors, e.g., developers, suppliers, customers, users and societal stakeholders, and cover different domains, e.g., technology, ethics, organizational culture, societal expectations and norms, and economics. Moreover, such experiments will need to be designed and executed with care, with bespoke conditions and for a limited period of time [41].

Despite such experimentation, *some* high-order effects will only materialize once systems are operational, or after being in operation for some time [42]. Dealing with such effects requires continued monitoring of potential unintentional, undesirable high-order effects and procedures to respond to them, e.g., through redesign of the system or adapted use of the system. De Reuver et al. [10] propose an adaptation to the traditional Value-Sensitive Design (VSD) methodology to better address unanticipated effects by extending VSD to the full life cycle for digital platforms that might also be useful for algorithms.

Above, we did already mention the Cobra effect and Goodhart's law: an organization puts measures in place to realize a certain objective, then things take an unexpected and different turn, and the measures backfire and work against that very objective. We hope our exploration and tentative recommendations can support people who are involved in the design and deployment of algorithms to promote *anticipation* and *responsiveness*. Besides these, Responsible Innovation also requires *inclusion*, e.g., by creating more diverse project teams or by involving stakeholders [39], and *reflexivity*, i.e. 'holding a mirror up to one's own activities, commitments and assumptions' [40], p. 1571). Looking forward, we can imagine further research into potential, unintended, undesirable higher-order effects by drawing from traditions like Technology Assessment [18, 23, 33] and Organizational Learning [1, 37, 43].

Declarations

Conflict of interest The authors declare that they have no conflict of interest.

References

1. Argyris, C.: On organizational learning. Blackwell Publishing, Cambridge, Massachusetts (1992)
2. Arnold, T., Scheutz, M.: Against the moral turing test: accountable design and the moral reasoning of autonomous systems. Ethics Inf. Technol. **18**(2), 103–115 (2016). https://doi.org/10.1007/s10676-016-9389-x
3. Barabas, C.: Beyond bias: "Ethical AI" in criminal law. In: Dubber, M.D., Pasquale, F., Das, S. (eds.) The Oxford Handbook of Ethics of AI. Oxford University Press, Oxford, UK (2020)
4. Barabas, C., Doyle, C., Rubinovitz, J.B., Dinakar K.: "Studying Up: Reorienting the study of algorithmic fairness around issues of power." ACM Conference on Fairness, Accountability, and Transparency, January 27–30, 2020, Barcelona, Spain (2020)
5. Binns, R.: Fairness in machine learning: lessons from political philosophy. Proc. Mach. Learn. Res. **81**, 149–159 (2018)
6. Bonnemains, V., Saurel, C., Tessier, C.: Embedded ethics: some technical and ethical challenges. Ethics Inf. Technol. **20**(1), 41–58 (2018). https://doi.org/10.1007/s10676-018-9444-x
7. Brinkhoff, S.: Big data data mining by the Dutch police: criteria for a future method of investigation. Eur. J. Secur. Res. **2**(1), 57–69 (2017). https://doi.org/10.1007/s41125-017-0012-x
8. Coeckelbergh, M.: Artificial intelligence, responsibility attribution, and a relational justification of explainability. Sci. Eng. Ethics **26**(4), 2051–2068 (2020). https://doi.org/10.1007/s11948-019-00146-8
9. Danaher, J.: The threat of algocracy: reality, resistance and accommodation. Philos. Technol. **29**(3), 245–268 (2016). https://doi.org/10.1007/s13347-015-0211-1
10. de Reuver, M., van Wynsberghe, A., Janssen, M., van de Poel, I.: Digital platforms and responsible innovation: expanding value sensitive design to overcome ontological uncertainty. Ethics Inf. Technol. (2020). https://doi.org/10.1007/s10676-020-09537-z
11. Dignum, V.: Responsible artificial intelligence: how to develop and use AI in a responsible way. Springer Nature, Cham, Switzerland (2019)
12. Eubanks, V.: Automating inequality. St. Martin's Press, New York (2017)
13. Ferguson, A.G.: Policing predictive policing. Washington Univ. Law Rev. **94**(5), 1109–1189 (2017)
14. Floridi, L.: The ethics of information. Oxford University Press, Oxford, UK (2013)
15. Floridi, L.: Translating principles into practices of digital ethics: five risks of being unethical. Philos. Technol. **32**(2), 185–193 (2019). https://doi.org/10.1007/s13347-019-00354-x
16. Floridi, L., Cowls, J., Beltrametti, M., Chatila, R., Chazerand, P., Dignum, V., Luetge, C., Madelin, R., Pagallo, U., Rossi, F., Schafer, B., Valcke, P., Vayena, E.: AI4People—an ethical framework for a good ai society: opportunities, risks, principles, and recommendations. Mind. Mach. **28**, 689–707 (2018)
17. Friedman, B., Kahn, P., Borning, A.: Value sensitive design and information systems. In: Zhang, P., Galletta, D. (eds.) Human-computer interaction and management information systems, pp. 348–372. M. E. Sharpe, Armonk, NY (2006)
18. Grunwald, A.: Technology assessment for responsible innovation. In: Van den Hoven, J., Doorn, N., Swierstra, T., Koops, B.-J., Romijn, H. (eds.) Responsible innovation 1: innovative solutions for global issues, pp. 15–32. Springer Science+Business Media, Dordrecht, The Netherlands (2014)
19. Gunkel, D.J.: Mind the gap: responsible robotics and the problem of responsibility. Ethics Inf. Technol. (2017). https://doi.org/10.1007/s10676-017-9428-2
20. Hayes, P., van de Poel, I., Steen, M.: Algorithms and values in justice and security. AI Soc. **35**, 533–555 (2020). https://doi.org/10.1007/s00146-019-00932-9
21. High-Level Expert Group on Artificial Intelligence: Ethics guidelines for trustworthy AI. European Commission, Brussels (2019)
22. Jungmann, N., Madern, T.: Basisboek aanpak schulden, Eerste druk Noordhoff Uitgevers, Groningen (2017)
23. Kiran, A.H., Oudshoorn, N., Verbeek, P.-P.: Beyond checklists: toward an ethical-constructive technology assessment. J. Responsib. Innov. **2**(1), 5–19 (2015)
24. Kulk, S., Van Deursen, S., Boekema, M., Breemen, V., Heeger, S., Philipsen, S., Snijders, T., Wouters, A.: Juridische aspecten van algoritmen die besluiten nemen: Een verkennend onderzoek. Boom Juridisch, Den Haag (2020)
25 Latour, B.: On recalling ANT. Sociol. Rev. **47**(1_suppl), 15–25 (1999). https://doi.org/10.1111/j.1467-954X.1999.tb03480.x
26 Marcus, G., Davis, E.: Rebooting AI: building artificial intelligence we can trust. Pantheon, Toronto, Canada (2019)
27. Meadows, D.H.: Thinking in systems: a primer. Chelsea Publishing, White River Junction, Vermont (2008)
28. Mittelstadt, B.: Principles alone cannot guarantee ethical AI. Nat. Mach. Intell. **1**(11), 501–507 (2019). https://doi.org/10.1038/s42256-019-0114-4

29 Mittelstadt, B.D., Allo, P., Taddeo, M., Wachter, S., Floridi, L.: The ethics of algorithms: mapping the debate. Big Data Soc. (2016). https://doi.org/10.1177/2053951716679679

30. Morley, J., Floridi, L., Kinsey, L., Elhalal, A.: From what to how: an initial review of publicly available AI ethics tools, methods and research to translate principles into practices. Sci. Eng. Ethics (2019). https://doi.org/10.1007/s11948-019-00165-5

31. Nathan, L.P., Klasnja, P.V., Friedman, B.: Value scenarios: a technique for envisioning systemic effects of new technologies. In: CHI '07 Extended Abstracts on Human Factors in Computing Systems. ACM, San Jose, CA (2007)

32. O'Neil, C.: Weapons of math destruction. Penguin, London (2016)

33. Rip, A., Robinson, D.K.R.: Constructive technology assessment and the methodology of insertion. In: Doorn, N., Schuurbiers, D., Van de Poel, I., Gorman, M.E. (eds.) Early engagement and new technologies: opening up the laboratory, pp. 37–53. Springer Science+Business Media, Dordrecht, The Netherlands (2013)

34. Rip, A., Te Kulve, H.: Constructive technology assessment and socio-technical scenarios. In: Fisher, E., Selin, C., Wetmore, J.M. (eds.) Yearbook of nanotechnology in society, pp. 49–70. Springer, Berlin, Germany (2008)

35. Russell, S.: Human compatible: AI and the problem of control. Allen Lane, London, UK (2019)

36. de Sio, F.S., Van den Hoven, J.: Meaningful human control over autonomous systems: a philosophical account. Front. Robot. AI **5**, 1–15 (2018)

37. Senge, P.: The fifth discipline: the art and practice of the learning organization. Doubleday, New York (1990)

38. Shneiderman, B.: Human-centered artificial intelligence: reliable, safe & trustworthy. Int. J. Human Comput. Interact. **36**(6), 495–504 (2020). https://doi.org/10.1080/10447318.2020.1741118

39 Steen, M., Nauta, J.: Advantages and disadvantages of societal engagement: a case study in a research and technology organization. J. Responsib. Innov. (2020). https://doi.org/10.1080/23299460.2020.1813864

40. Stilgoe, J., Owen, R., Macnaghten, P.: Developing a framework for responsible innovation. Res. Policy **42**, 1568–1580 (2013)

41. van de Poel, I.: An ethical framework for evaluating experimental technology. Sci. Eng. Ethics **22**(3), 667–686 (2016). https://doi.org/10.1007/s11948-015-9724-3

42. van de Poel, I.: Society as a laboratory to experiment with new technologies. In: Bowman, D.M., Stokes, E., Rip, A. (eds.) Embedding new technologies into society: a regulatory, ethical and societal perspective, pp. 61–87. Pan Stanford Publishing, Singapore (2017)

43 van de Poel, I., Asveld, L., Flipse, S., Klaassen, P., Kwee, Z., Maia, M., Mantovani, E., Nathan, C., Porcari, A., Yaghmaei, E.: Learning to do responsible innovation in industry: six lessons. J. Responsib. Innov. (2020). https://doi.org/10.1080/23299460.2020.1791506

44 van Veenstra, A.F., Grommé, F., Djafari, S.: The use of public sector data analytics in the Netherlands. Transform. Gov. (2020). https://doi.org/10.1108/TG-09-2019-0095. (**ahead-of-print**)

45 Van Wynsberghe, A., Robbins, S.: Critiquing the reasons for making artificial moral agents. Sci. Eng. Ethics **25**(3), 719–735 (2019)

46. Yin, R.: Case study research, 2nd edn. Sage, Thousand Oaks (1994)