AI & Ethics at the Police:

Towards Responsible use of Artificial Intelligence in the Dutch Police

Co-authored by:

Leiden University:	*TU Delft:*
Dr. Francien Dechesne	Dr. Virginia Dignum
Lexo Zardiashvili, LLM	Jordi Bieger, MSc

AI & Ethics at the Police:

Towards Responsible use of Artificial Intelligence in the Dutch Police

Co-authored by:

Dr. Francien Dechesne,
Dr. Virginia Dignum,
Lexo Zardiashvili, LLM
Jordi Bieger, MSc

This white paper is produced as part of research conducted by Leiden University Center for Law and Digital Technologies (eLaw) and TU Delft Institute of Design For Values, commissioned by the Dutch National Police. The research objective is to identify possible consequences of using AI for and by law enforcement and the ethical issues this may lead to. This white paper describes the state-of-the-art in AI, how it could benefit law enforcement, and what ethical concerns will need to be addressed in the use of AI in order to safeguard the legitimacy of and trust in the national police. It does not provide an ethical evaluation and assessment of police utilization of AI, nor does it claim to have analyzed all relevant factors for such evaluation.

The qualitative research for this white paper consists of two parts. (1) We present a review of literature on developments in AI technology and ethical considerations surrounding existing applications, both in general and specific to the police domain. (2) In the scope of the research we conducted interviews with relevant parties in the law enforcement chain: 4 employees of the Dutch Police, 1 external hire (Ordina), 1 Ministry of Justice, 1 Public Prosecution Service (Openbaar Ministerie, hereafter OM), 1 City of Amsterdam, 1 Dutch academic, 1 international body (UNICRI), 1 technology company (Sentient). The reports of these interviews provide a perspective on the experience with and the potential of AI for the police profession. In addition, in the scope of this research a questionnaire was developed for data scientists working (internally or externally) in the police (see Appendix 3).

Version 1.2. - Updated 22.03.2019

March, 2019

Leiden / Delft
The Netherlands

Executive Summary

Artificial Intelligence (AI) is increasingly being used to perform an ever-growing range of tasks that previously required human intelligence. Recognizing the many potential benefits, the Dutch National Police wishes to utilize this technology to continue and improve its ability to uphold the law. Given the police's special role and authority in society, which is highly dependent on societal trust, it is important to ensure that new technologies like AI are used in an ethical and responsible manner.

While AI is indeed a very promising technology that could provide many benefits to the police, it is important to be aware of its limitations, understand how they result in challenges for the ethical and responsible use of AI, and remedy the misconceptions people have. In this white paper we intend to give a realistic picture of AI and ethics in relation to the police practice for the upcoming years.

Ethics, as a set of accepted principles on what is morally right or wrong within a community, can be seen as foundational for frameworks on fundamental rights: the EU Charter is founded on the indivisible, universal values of human dignity, freedom, equality and solidarity. In this paper, we use "ethics" to refer to moral rights and expectations not otherwise fixed by laws or regulations. Based on our review of the relevant literature, we identify six principles for the responsible use of AI, resulting in requirements on the technical, individual and societal levels. These principles are Accountability, Transparency, Privacy and Data Protection, Fairness and Inclusivity, Human Autonomy and Agency, and (Socio-technical) Robustness and Safety.

We connect these principles to current visions on AI and the practice of AI use and development within the police on the basis of expert interviews and a small-scale survey among data scientists active for the Dutch police. Aside from strategies for addressing these specific principles, we recommend the general strategies of Design for Values, Regulation, Standardization, Awareness and Dialogue. More specifically, we recommend an internal ethical review board for AI, augmenting the police's code of conduct to include the ethical use of AI for employees working with AI, and continued and closer collaboration with law enforcement, academic and other expert partners. The recent launch of the National Police Lab AI is a great initiative in that way.

AI has many potentially beneficial applications in law enforcement including predictive policing, automated monitoring, (pre-) processing large amounts of data (e.g. from confiscated digital devices, police reports or digitized cold cases), finding case-relevant information to aid investigation and prosecution, providing easier to use services for civilians (e.g. with interactive forms or chatbots), and generally enhancing productivity and paperless workflows. It can be used to promote human dignity, freedom, equality, solidarity, democracy and rule of law. However, AI techniques are powerful and can pose challenges to the rights and principles we outlined. These challenges should be taken into consideration in the development and use of AI application by the police to prevent the societal trust on which their operations rely from eroding. The strategies and recommendations for responsible use of AI in this paper aim to contribute to retaining this trust.

Table of Contents

Introduction

Artificial Intelligence (hereafter AI) is increasingly used in situations that traditionally required human intelligence. It is embedded in all aspects of everyday life, re-shaping human interactions as well as our environment.

AI can help us in many ways: it can relieve us of hard, dangerous or boring work; it can help us save lives and cope with disasters; and, it can entertain us and make our daily life more comfortable. AI is very suitable to manage complex, data-intensive tasks, e.g. monitor credit card systems for fraudulent behavior, enable high-frequency stock trading, support medical diagnosis and detect cybersecurity threats. AI is soon to move and work among us in the form of service, transportation, medical and military robots (1).

The Dutch National Police intends to experiment with AI in several work processes, as it fits well within their philosophy of information-driven policing (2). Envisaged outcome of such experimentation is to help the police make use of these technological innovations to better carry out their tasks for the benefit the people: those in the service of the police – who will be working with those systems – as well as civilians, for whom the police organization fulfils such important and sensitive social task. To ensure that new technology like AI actually enables the police to do their job better and keep up with the communities that they serve, development and implementation of technology for use in law enforcement should happen in an ethical and socially responsible manner.

This white paper was commissioned by the Dutch National Police to identify the considerations for such ethical and responsible use of AI. Assuming that within the coming years, the police will deploy AI on a large scale, the research assignment is centered on the following overarching research questions:

(a) What consequences does the deployment of AI applications on a large scale in policing have for the police profession and the police domain?
(b) To which ethical issues with respect to the police profession and policing does this lead?

The use of AI systems by law enforcement in policing directly affects human life and notions of justice underlying the state. Therefore, while evaluating possible effects of AI on human beings and the common good, particular attention should be given to how AI use potentially contributes to and/or enforces asymmetries of power or information.

Our research methods include literature review, in-person interviews and a questionnaire. We reviewed the literature on state-of-the-art applications of AI with a special focus on possible connections to the law enforcement domain and the ethical considerations that arise. To learn more about the experiences and perspectives of the people and organizations involved in the use of AI at the police, we conducted 11 interviews with 4 employees of the Dutch Police, 1 external hire (Ordina), 1 Ministry of Justice, 1 Public Prosecution Service (Openbaar Ministerie, hereafter OM), 1 City of Amsterdam, 1 Dutch academic, 1 international body (UNICRI), 1 technology company (Sentient). Additionally, we conducted a survey among data scientists working for the National Police (internally or externally) to get both quantitative and qualitative responses from a group of experts in both data science[1] and the law enforcement domain.

PART I – Effects of AI Use in the Police

1. The role of the Police in the state

1.1 Under the Dutch Police Law (Politiewet 2012) the task of the Dutch police is two-fold: (1) to ensure maintaining the rule of law (law enforcement) and (2) to provide assistance to those in need (3). The police in the Netherlands do not stand alone in these tasks: they cooperate on law enforcement with several bodies in the national and local governments (“*ketensamenwerking*”), such as municipalities and Ministry of Justice and Security.

1.2 The police have a special role in society that involves a constitutional right to use violence for the enforcement of the law (4). For the police to function and realize its objectives, society has to deem the police as legitimate and trust that it is effective in its tasks (5). In order for the police to be trustworthy in their *efficacy*, they must continuously innovate to evolve with developments, stay ahead of criminals’ new strategies and capabilities, and utilize new methods and technology for the fulfillment of their tasks. In order for the police to be trustworthy in their *use of power*, the police must demonstrate good will and respect for the rights of civilians. The National Police greatly values the trust of Dutch citizens, which was measured to be the highest of any measured institution in 2017 (6). It is important to retain this trust, also when introducing new technologies such as AI that have a fundamental impact on the nature of their operations and interactions with society. This report reflects on ethical approaches to the use of AI in the police practice, to provide a basis for trustworthiness under these changes.

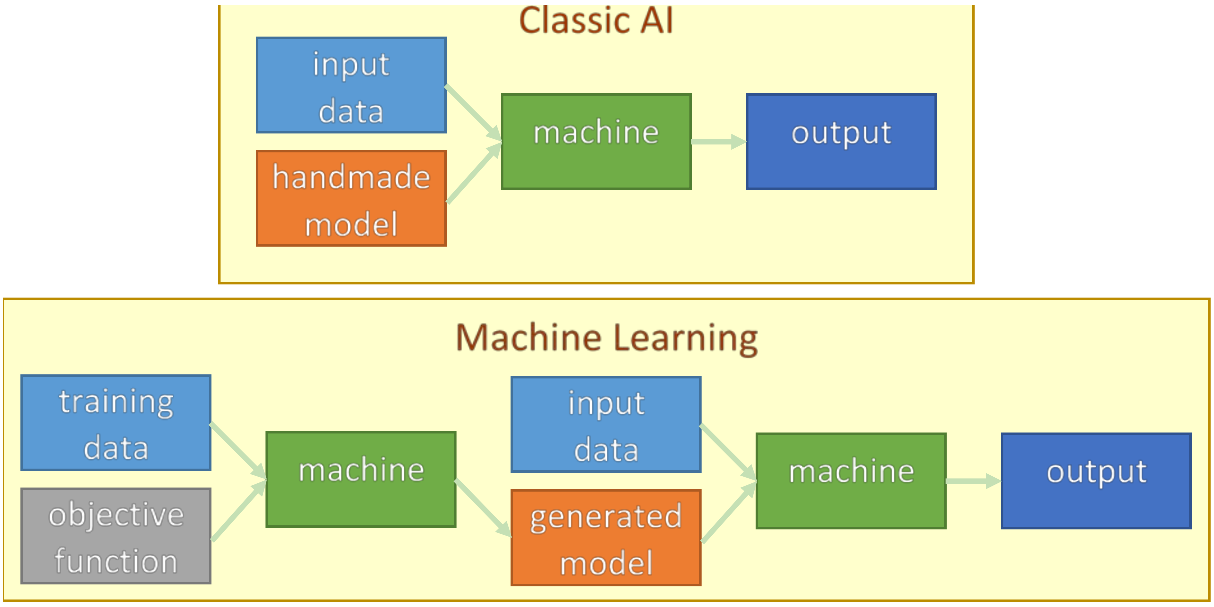
2. Artificial Intelligence

2.1 Artificial intelligence (AI) is a broad, multidisciplinary field of study that concerns itself with the construction of machines and systems that perform functions commonly associated with the cognitive capacities of the human mind. While there is no clear consensus on what systems should be regarded as AI, we take the term broadly to include machine learning (hereafter ML), (autonomous) decision making, search, planning, reasoning, perception, natural language processing, etc.[2]

[1] Data science is closely related to AI, and for the purposes of this research is considered a part of it.

[2] Due to natural cycles in buzzword popularity, we note that "AI" is currently being used to describe things that might previously have been called "data mining", "big data", "business intelligence", "data science", "deep learning" or many other things.

Figure 1: Illustration of the construction of AI systems.



2.2 AI systems convert input data into outputs (results, decisions, predictions or judgements) by applying a so-called *model* that describes its behavior (see Figure 1). In classic AI models are handmade by human programmers by imbuing them with explicit knowledge, often in the format of rules and elicited from domain experts. Such an explicit approach has pros and cons, as it both *allows* and *requires* ex ante specification of the formal instructions for what the system should do. This makes it hard to automate tasks that rely on tacit knowledge or that are otherwise ineffable.

Machine learning (ML) algorithms generate or fine-tune models by attempting to optimize a human-provided *objective function* – a function that indicates how good an outcome is – based on data it interacts with during a process called *training*. ML allows for the identification of patterns in the data, often unknown to humans, but tends to produce opaque[3] AI systems whose reasoning and decisions we do not completely comprehend, and it is often unclear how exactly the choices of training data and objective function affect this.

2.3 One category of AI applications supports our perceptive abilities, for example by transcribing handwritten, typed or spoken text; by detecting objects (e.g. guns), suspicious behaviors (e.g. violent/suspicious or lying) or salient events (e.g. trespassing); or by identifying people based on their faces, voice, fingerprint, DNA or other information. Natural language processing (hereafter NLP) can extract structured observations from unstructured text, search in and through documents, detect expressed sentiments, summarize texts, and translate between languages. Data science allows for the analysis and visualization of large amounts of data, the detection of patterns in space and time (e.g. crime or weather), including clusters and outliers (e.g. fraudulent credit card transactions), find similarities between events and people (e.g. crimes and criminals or victims), and generally uncover

[3] Opacity is the opposite of transparency. Often the term “black box” is used for opaque systems. Technically a system is a “black box” if we can observe its inputs and outputs, but we lack a meaningful understanding of its internal workings; i.e. how input is transformed into output.

complex correlations between features of the data on which predictions could be based (e.g. in predictive policing or risk assessments).

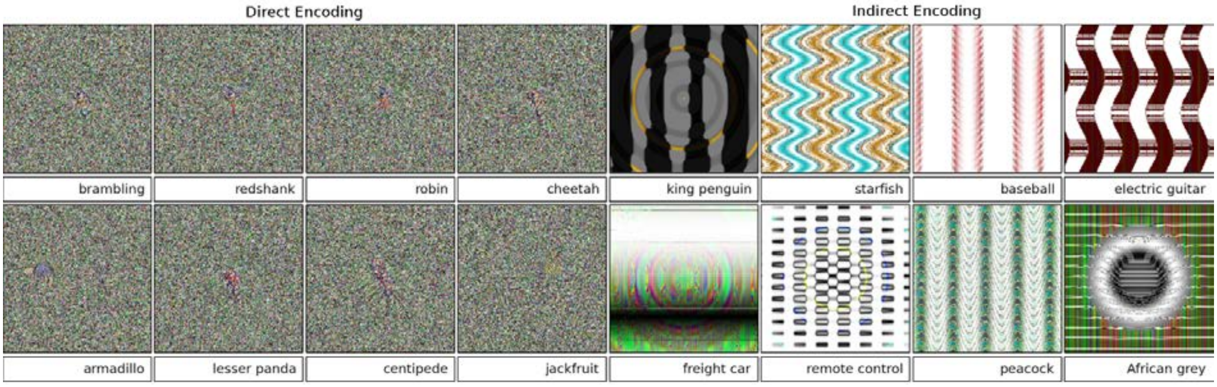


Figure 2: Evolved images that are unrecognizable to humans, but that state-of-the-art DNNs trained on ImageNet believe with ≥ 99.6% certainty to be a familiar object. This result highlights differences between how DNNs and humans recognize objects. Images are either directly (top) or indirectly (bottom) encoded. (8)

2.4 AI systems or agents can also directly act based on the information they receive or uncover. Embodied AI systems, such as robots, can take many physical forms, including humanoids, self-driving cars, drones or surveillance bots. Chatbots, or virtual assistants, can guide the user through a process (like reporting a crime) or help them achieve it (e.g. Siri putting an appointment in your agenda). AI techniques can help with scheduling, resource allocation, and route planning. Agent-based simulation enables the analysis of the effect of new policies and supports the development of strategies to increase security against criminals, e.g. applying game-theory approaches.

2.5 AI techniques are also applied to generate partially or entirely new instances of certain objects, such as pictures, videos, (news) stories, faces, voices, music, (digital) paintings or car parts. These can be applied for purely artistic purposes, but generative design can also serve to optimize certain aspects of a product (e.g. the strength and weight of car parts). Relevant for law enforcement is that these techniques can also be used to mislead: writings, voice recordings, pictures, and even video can be either fabricated “from scratch” or altered to appear like they feature other people. This may be used to commit crimes or to obfuscate evidence. On the other hand, AI techniques can also assist in effectively detecting fake material.

2.6 While the potential of AI techniques is impressive, there are also many challenges that require awareness and attention. AI systems currently still lack common sense and an understanding of context or the concepts they are dealing with, and – as computer programs – will literally execute the instructions in their programming. For instance, if you program an AI system to play a video game and penalize it for dying, it may simply pause the game, even though that’s not what the programmer wanted (7). Or an AI may be very good at object detection, but then get fooled by weird patterns in a way that no human would (see Figure 2) (8).

2.7 The behavior and reliability of ML-based systems depends in large part on the data that is available for training, which typically needs to be accurate and plentiful and requires a lot of computational resources. Another large part is played by the *objective function* that the ML algorithm is attempting to optimize. Specifying exactly what we want an AI system to do – whether that is through the direct programming of expert knowledge or behavior, the provision of data, or choosing the objective function – can be very difficult for humans. This can result in a lack of robustness, unintended outcomes and the inability to deal with unforeseen circumstances and a changing world (cf. Chapter 10).

2.8 Another challenge is the necessary monitoring, evaluation, updating, management and maintenance of AI systems. And where data is used, it needs to be carefully gathered, sanitized, entered into a database, (often) annotated, evaluated, securely stored, and used to (re)train the ML system or (re)evaluate predictions. Failure to do so can result in inaccuracies, unfair outcomes (cf. Chapter 8) and breached privacy (cf. Chapter 7). In the end, it is always a human or organization who is responsible for this, and who decides how AI is applied (cf. Chapter 5). The level of autonomy is not inherent to an AI system but determined by what it is allowed to do without human oversight or control (cf. Chapter 9). Making sure this oversight or control is meaningful can be challenging and may require a degree of transparency in how the AI systems works (cf. Chapter 6). This can be a problem, especially if the system is (partially) based non-monotonic techniques such as machine learning, where properties and behavior evolve rather than are designed or when dealing with large amounts or poorly legible data.

3. The use of AI in the Police organization

3.1 Rapid technological developments demand efforts from the police to stay up-to-date with the state-of-the-art to adequately position themselves in the midst of modern society. This is important to keep pace with criminals who utilize such new technology and to combat new forms of crime emerging as a result. Furthermore, the police have a moral obligation to society to carry out its task to serve and protect as well as reasonably possible. Technologies like AI-techniques can be used to improve the efficacy and efficiency of law enforcement processes (e.g. investigations), as well as for empowering civilians – e.g. by opening up additional communication channels with the police.

3.2 AI has many potentially beneficial applications in law enforcement including predictive policing, automated monitoring, (pre-) processing large amounts of data (e.g. from confiscated digital devices, police reports or digitized cold cases), finding case-relevant information to aid investigation and prosecution, providing easier to use services for civilians (e.g. with interactive forms or chatbots), and generally enhancing productivity and paperless workflows. AI also promises to enhance police operations through techniques such as image/face/behavior recognition. Having in mind the possible benefits that AI can provide in the law enforcement domain, the police cannot ignore this technology. The question is not *if* AI should be used, but *what* it is most suited for and *how* it can be properly implemented.

3.3 Currently the police in the Netherlands has been using AI in all of the applications mentioned in the paragraph 3.2. For example, the Crime Anticipation System (CAS) is an internally developed predictive policing tool that aims to predict crimes with statistics based on data from the various sources (9); Pro-Kid 12- SI (pronounced “Pro-Kid twelve-minus”) is a rule-based system for risk assessment on children aged between 0-12 years, used nationwide by the police to prevent children from being involved in a crime or anti-social behavior (10); the Online Fraud Report Intake System uses NLP techniques, computational argumentation (legal informatics) and reinforcement learning to assist civilians in reporting crime.

3.4 A large part of police work (for example, important parts of the investigation process and crime report intake, etc.) is done behind the desk and consists of bureaucratic work. Growing amounts of paperwork mean less resources for fieldwork and communication with civilians, which can be seen to cause a disconnect between police personnel and civilians. Such a disconnect might raise concerns of job satisfaction of the police personnel as well as their productivity, which can have a negative impact on the general efficiency of the police organization. The police organization sees AI as part of the solution to help them eliminate such a disconnect, as well as increase job satisfaction and general efficiency/effectiveness of the police (see for example Box 1).

3.5 The long-term vision of the organization involves building AI support for bureaucratic processes and the connection of these processes within the network of automated systems. This will amount to a semi-autonomous business process within the police organization. To reinforce this vision, AI scientists within the police are working on pilots for individual building blocks.

Box 1 – *Pilot: Online Fraud Report Intake System ("Fraud System")*

One pilot is the development of an AI-supported online fraud report intake system (hereafter the "Fraud System"). Using natural language processing (NLP) techniques, computational argumentation (legal informatics) and reinforcement learning, the police gets structured data out of the text of a civilian's complaint. A collection of AI systems is used to obtain "observations" from the input, which is the filled-out form – i.e. a (structured) collection of structured and unstructured data. Example observations might be "the product was delivered" and "a payment occurred". A domain-specific argumentation theory uses observations to reach a verdict on whether the report can be discarded or should be processed further. This can also be used to figure out which observations (and hence which information) is needed next from the reporter. If an observation cannot be obtained, the civilian or processing officer can be warned. Reported intake is part of a legal chain of police tasks – while dropping the case might result in a criminal achieving his goal, inaccurate acceptance of the complaint might invade someone's privacy. The Fraud System is already operational in its minimal version and it is planned to go fully operational by the summer of 2019.

3.6 Improper handling of the challenges inherent in the use of AI (cf. 2.6-2.8) may also threaten the legitimacy of and trust in the police. The opacity of reasoning, inherent in some AI techniques might decrease transparency and weaken human agency in the police's decision-making. Accountability and transparency are closely related to the perception of justice in public. (4) The police must be able to publicly answer the question of whether their duties are performed in accordance with the applicable rules. It must be possible to remove any doubts by providing openness and performing audits of the circumstances of police actions; not merely *explaining* the decisions, but also by making them *justified*. The inherent opacity of some AI techniques could be a hindrance for ensuring such auditability and transparency, and thereby accountability (effects on public trust in the police organization).

3.7 Implementing AI systems within the police poses a number of organizational challenges. As parts of the organization become automated, it is important to ensure the interoperability of systems so that continuity of the police processes is guaranteed. Development, acquisition, maintenance and use of these systems may require a differently skilled workforce that may not fit neatly in the traditional police job hierarchy. In addition to the creation of new jobs, AI can also change or in some cases displace existing jobs. AI can improve the well-being and job satisfaction of police personnel by relieving them of boring or repetitive (bureaucratic) work, rendering them more efficient, and allowing them more time to spend on other tasks (e.g. social tasks or fieldwork). However, care must be taken that AI systems don't negatively impact employees' sense of purpose, agency or safety, and if necessary adequate opportunities for re-training or transfer to other jobs within the organization should be explored (effects on the police personnel and police organization).

3.8 AI might also impact the general security of the country by increasing chances of successful police investigations. In addition, it can make it easier to start prosecution, as it may provide the prosecutor with better information and legal argumentation. However, AI might also trigger more investigations in less time. The capacity of police and OM should remain proportional to cases in question so that their resources are not disproportionately allocated on online fraud and therefore such a system does not have the unintended negative impact on general security system of the country (effects on the general security system of a country).

3.9 Benefits for the police personnel and general security of the country have indirect positive effects for the civilians as well, as their concerns will have more prompt (and arguably more effective) response from the police. AI systems will also have *direct* benefits on civilians, as they might enable more inclusion by providing another means for communication, convenience, feedback, etc. (effects on civilians).

Box 2: *Benefits of Fraud System for Civilians*

Often online fraud concerns a small amount of money and going to the police station requires more time and energy than civilians consider worthwhile. In addition, sometimes fraud cases go along with shame and embarrassment, which civilians often wish to avoid. The Fraud System aims to make reporting online fraud easier, more pleasant and less costly by allowing people to do so from the comfort of their homes and without needing to relay potentially embarrassing events in person, which will hopefully result in less fraud cases going unreported. However, these positive impacts are presumed and not yet tested in practice. Many civilians might prefer face-to-face interaction with a police officer as they might seek empathy from another human being.

4. Responsible use of AI

4.1 The role of the police and its operations within a society requires continuous ethical, legal and societal reflection, evaluation of the proposed solutions, and the embedding of accountability mechanisms. Based on the effects discussed in paragraphs 3.6-3.9, this holds in particular for the use of AI in policing.

4.2 The police organization in the Netherlands is committed to protect fundamental human rights and to ensure respect for the rule of law (3). The police is directly obliged to comply with domestic and international legal instruments that specify this commitment, like the national constitution, EU Charter, specific national legislative acts, and the EU directives and regulations like the General Data Protection Regulation (GDPR) or Law Enforcement Directive. Fundamental human rights can inspire new regulatory instruments and guide the rationale for the development, use and implementation of AI systems (11).

4.3 In a democratic state such as the Netherlands, compliance with holding laws and regulations must be seen as a given for any application of AI. Practical considerations of ethical use of AI should therefore focus on the spaces left open by the law. Legal compliance is necessary but not sufficient for ethical use of AI. Of course, the existing legal frameworks should also continuously be re-evaluated to see if improvements are possible (12). Such re-evaluation requires identification of moral principles that led to the current fundamental rights framework. Decades of consensual application of fundamental rights in the EU provide clarity and readability of these moral principles. For example, in the core of the right to “respect for human dignity” is the principle that all human beings have an inherent value just by the virtue of being human (regardless of their characteristics) (11). Human dignity encompasses the idea that every human being possesses an “intrinsic worth”, which should never be diminished, compromised or repressed by others – nor through the use of any technology, including AI. (13) Identifying this principle can help us look at the values that we need to uphold in the scenarios that are not specifically addressed by regulation.

4.4 In common use, the term “ethics” refers to a set of accepted principles on what is (morally) right or wrong within and for a certain community.[4] As such, it is a normative framework, which can be seen as foundational for more formalized frameworks on fundamental rights: the EU charter cites “the indivisible, universal values of human dignity, freedom, equality and solidarity” as principles on which it is founded. We use “ethics” here to refer to ‘soft’ moral rights and expectations not otherwise fixed by laws or regulations. They can help us understand how AI may affect different fundamental rights, as well as provide better guidance on what we *should* do with this technology for the common good rather than using it just because we *can* (11).

4.5 In order to use AI responsibly, general and abstract ethical principles need to be mapped into concrete requirements. Requirements for such an approach are grounded in three pillars of equal importance:

4.5.1 **Technical**. Requirements on the technical level concern properties of the AI system itself, such as system integrity, task efficacy, transparency and interoperability with other systems.

4.5.2 **Individual**: Requirements on the individual level concern the rights, agency and well-being of individual people who are affected by the AI system (e.g. civilians and police personnel).

4.5.3 **Societal**: Requirements on the societal level concern the effects of the AI system on society as a whole, including e.g. crime levels, clearance rate and perceptions of the police.

4.6 Based on the literature survey, interviews and questionnaire we conducted we have selected six concerns that the Dutch police must address to ensure that their use of AI upholds the high-level ethical principles enshrined in existing fundamental rights frameworks: (1) Accountability, (2) transparency, (3) privacy & data protection, (4) fairness & inclusivity, (5) human autonomy & agency, and (6) (socio-technical) robustness and safety. They will be outlined in the next Part of this white paper.

[4] Ethical principles are usually distinguished from “morals” in that the latter are taken to be individual sets of beliefs. As an academic field, ethics usually refers to the branch of philosophy that pertains the systematic study of what is (morally) right or wrong, which can be analytic, descriptive or normative.

PART II – Principles for Responsible Use of AI in the Police

5. Accountability

5.1 Accountability in the normative institutional sense is the means through which authority is "controlled" in order to render it "appropriately" exercised. (14) In the context of this white paper we use "accountability" to refer to the ability to hold people or groups accountable or responsible (or sometimes liable) for an action, choice or decision: who deserves credit or blame.[5] To ensure accountability, decisions should be derivable from, and explained by, the decision-making mechanisms used. It also requires that the ethical values and societal norms that inform the purpose of the system – as well as their operational interpretations – are explicit and open.

5.2 From an ethical and legal perspective, responsibility/liability must always be assigned to a moral agent or legal person: AI systems are neither.[6] However, the use of complex technology like artificial intelligence can lead to "attribution confusion", where it is not clear who, if anyone, should be held responsible. Typical candidates might be the owner of the technology, the creator, and the user. The situation becomes more complex if multiple technological systems and people are involved, whose collective actions resulted in a problem (15).

5.3 Assigning legal liability can in some ways be easier than assigning moral responsibility. It is (naturally) imperative that the police work within existing legal frameworks, regardless of what technologies (including AI) are involved. When an AI system is bought or rented from a third-party, contracts will need to carefully outline which organization is liable under what circumstances. If liability must be assigned within the police organization (e.g. if the AI application is developed in-house), it is important to take incentives and workability concerns into account: for instance, if there are potentially severe consequences for a programmer if their code has a bug, this may require (even) higher salaries (hazard pay) or significantly slow down development. Or if an analyst/user is held individually responsible, they may not feel comfortable working with (opaque) AI techniques.

5.4 Accountability is a particularly important requirement in the law enforcement domain, and has broader applicability than mere responsibility, answerability, liability. Police accountability involves holding both individual police officers and law enforcement agencies responsible for effectively delivering basic services of crime control and maintaining order while treating individuals fairly and within the bounds of the law (14). This means that the

[5] Accountability is sometimes also taken to refer to the requirement for the system to be able to explain and justify its decisions to users and other relevant actors. This important aspect of responsible use of AI is indeed essential for ensuring accountability, but is itself a different concept that is discussed under the header of transparency of decisions (specifically through explainability and interpretability) and discussed in Chapter 6 (specifically 6.2-ix).

[6] There have been discussions about legal personhood for AI, but the current situation is that AI systems are not legal persons in the Netherlands and virtually all other countries on Earth (Saudi Arabia granted citizenship to one robot). We will not speculate about the eventuality that this may change in the (far) future, except to say that legal personhood would only address issues of legal liability and not moral responsibility.

police are under constant observation from the public (their activities are also monitored by Ministry of Justice and Security) who demand that they respect laws regarding due process, search and seizure, arrests, discrimination, equal employment, sexual harassment, etc. Such oversight is important for maintaining the public's "faith in the system" on the societal level. On the individual level, civilians must have some rights of recourse if they feel like they were mistreated.

5.5 For the Dutch police to ensure such accountability of newly developed AI systems as well, proper oversight of the systems must be ensured. Such oversight can be two-fold: first, through reviewing AI systems through the lens of ethics by an *internal AI review board* and second, by providing third-party expert opinion on the systems in question in the form of, for example, external audits. Current efforts of the police to involve academia for external research on identifying ethical concerns and the impacts of AI systems in the police are welcome and must be seen as steps towards ensuring accountability.

5.6 Such oversight, on the other hand, can only be adequate and meaningful if the systems can be reviewed (auditability), and if the decisions that they make explained and justified (explainability) on the technical level. This places requirements on transparency (see chapter 6) and reproducibility. Independent evaluations should be able to verify and reproduce the AI-system’s behavior in all situations. However, this can be complicated by the complexity, non-determinism and opacity of many AI systems, together with their sensitivity to training/model building conditions. There is an increased awareness within the AI research community that reproducibility is a critical requirement in the field (16).

5.7 While AI developers currently are quite aware of ethical issues in developing AI systems for the police, the lack of structured oversight and governance of their work suggest the necessity for more comprehensive training. This will contribute to the necessary skills and knowledge of AI developers to take on the responsibility reasonably assigned to them. The police organization also has to guarantee that third parties or their employees are able to report potential vulnerabilities, risks or biases, and put clear processes in place to handle these reports (for example, install a single point of contact for dealing with concerns arising from using AI systems).

6. Transparency

6.1 Transparency is an important component in ensuring trust and figuring out who or what is accountable for potential problems with AI systems. With transparency, we must always ask 1) about what, 2) to whom and 3) how much transparency should be provided, and of course to what end.

6.2 <u>*Transparency about what?*</u> – Transparency can be provided regarding different aspects of an AI system or process: e.g. about the rationale for development and use, about the development process and design decisions, about evaluations of the system, about the used data, about the system's code and inner workings, or about individual decisions that the system produces. The AI literature pays special attention to explicating and understanding how an AI system might come to a certain outcome in branches studying interpretable/explainable systems (17) (18).[7] This results mainly in requirements on the technical level (See 4.5.1).

i) People – Giving transparency about who is involved in the financing, management, development, operation and maintenance of the AI systems can help determine whether effective protocols were followed, who is accountable for what, and where biases or conflicts of interest might crop up. This involves identifying who owns what (algorithm, data, process, etc.) and who is responsible for what. Transparency can be given at the organizational or individual level, but the privacy and rights of involved parties should always be taken into account.

ii) Rationale – At the highest level of abstraction one can be transparent about the rationale for using an AI system or process. What are the goals the AI techniques are meant to (help) achieve, and why is this technology the best tool for the job? What were the requirements for the system? What are the envisioned and acceptable costs and benefits, and to whom do they fall primarily? What is the scope of operation?

iii) Development – Transparency about the development process can involve information about used methodologies, technologies and protocols, and the justification for their use. What design decisions were made, and what compromises? What is the purpose or intent of the system, who was involved on determining that and how is this purpose being enforced?

iv) Operation – How does the AI system operate in practice? What do police personnel (have to) do to make the system operate optimally? How does the application of AI interact with other parts of the police organization or police work? How does it change employees' jobs, and does it change them for the better?

v) Analyses/Evaluations –AI systems and their use should be analyzed and evaluated, and the results can be made public. This can involve e.g. results about accuracy, affected parties, operational costs, how the police use it and how happy employees are.

vi) Data – Data is one of the most important factors that determine the behavior of AI applications. This goes for both the data that's used to develop an AI system (e.g. training, validation and test data) and the input data that it's fed during operation. How was the data collected, stored, cleaned, anonymized[8], de-biased[9],

[7] The operation of *interpretable* AI systems is directly meaningfully understandable to humans. *Explainable* AI attempts to meaningfully explain AI behavior that is typically not directly understandable. Rule-based systems are typically interpretable, while e.g. neural networks require an explanation for why they arrived at their output (e.g. by showing which inputs contributed most to the decision).

[8] Please note that perfect anonymization is often impossible, due to inferences that can be made using external data sources. See Chapter 7.

[9] De-biasing aims to remove certain biases from data, but 100% removal is typically beyond the state of the art. See Chapter 8.

preprocessed and/or presented to the AI system? Transparency can be achieved to different degrees by e.g. open sourcing or giving away the data, or by simply giving high level descriptors or answering questions about it.

vii) Implementation – Transparency about implementation describes how the system works. This can be done at a high level (in e.g. a research paper), by showing design documentation (e.g. UML diagrams), or by showing the code. Naturally, transparency can be provided about an entire system or only parts of it.

viii) Executable – A special kind of openness might result from making (part of) an AI system available as an executable or library. This will allow beneficiaries to use or test the system for themselves. It should be noted that this can have the side-effect of also publishing the source code, because executables can often be "decompiled" (sometimes even if code obfuscation software is used to prevent this).

ix) Decisions – Finally, people may require a meaningful understanding of how individual decisions or judgments are made with AI (especially when it concerns police decisions). This can help uncover whether something has gone wrong (e.g. a bad data source was used, a reasoning error was made, or bias was present in the system), or find the source if it has. In so-called interpretable AI systems, developers can look into the AI system itself, trace what happens when a certain input is given, and understand at a high level why the output was produced. In other AI systems this would not lead to meaningful insights, e.g. because they are composed of many interconnected small components that hold no meaning to humans. In such a case we might try to get explanations. Ante-hoc explainability is built into an AI system, and can affect how it works (it can e.g. provide an extra output that signifies the system's confidence). Post-hoc explainability methods are used to "explain" things about existing AI systems. For example, they might be able to tell you how much influence each input feature had on the system's output (19), (20). A bad way of approaching this is to run a high-performance opaque system side-by-side with a simpler interpretable system and explain the decisions of the opaque system based on how the interpretable system decided.

6.3 *Transparency to whom?* – On the individual and societal level different people may require different kinds of transparency. A judge potentially needs to be able to fully understand and reproduce the outcome of an AI system in a way that doesn't require detailed technical knowledge. Developers and IT experts within the police may want more technical detail. For the media and the general public, it depends on what the goal is: an open source code base will not be readable to the majority of civilians, but it may enhance trust to know that full transparency is given to journalists and experts.

i) Courtroom – Judges must be able to have full access to the decision-making and judgements that are relevant to the prosecution of criminals. Furthermore, a defendant's lawyer also has extensive rights to interrogate these decisions and request a great amount of transparency. This may be granted by the judge. Like with most non-developers, they will primarily require explanations that are relatively simple and understandable without extensive technical knowledge. This could be accomplished through high-level explanations and justifications of decision-making, visualizations and expert witnesses.

ii) Police organization – The police organization itself needs to have insight into the systems that it is using and developing. Documentation should exist for most types of transparency. Care should potentially be taken that whoever requests certain information is authorized to receive it, especially with sensitive data and proprietary algorithms.

iii) The public – The media, journalists or individual civilians may request transparency on each of the above criteria, but the amount of transparency given should be carefully considered. Individuals may have a right to explanation about the decisions that affect them. On the societal level transparency can (be necessary to) build trust, but once something is out in the open, it cannot be undone. No information should be published that allows significant misuse or gaming of the system. Privacy should be considered at all times.

iv) Third parties – Other third parties will typically fall under the category of "the public" but may occasionally have a special status. For instance, collaborators may require access to more information. It's important to ensure that they keep any sensitive secrets and that they don't stand in the way of further transparency by the police. For instance, the systems the police use should always be transparent to the police, and if necessary, to a judge. It cannot be the case that civilians' lives are affected through an AI system that is a proprietary secret belonging to the company who developed it.

v) Government – Other government parties may furthermore have more need for transparency, either as collaborators or as watchdogs. The police organization should be able to work together with other government agencies, but also to be answerable to the public. Audits may be conducted to this end, where it is ensured by non-police government that the police are filling its role responsibly.

vi) Developers – The developers of an AI system will naturally have access to the system(s) they are themselves developing. However, in many cases new software will have to work together with other software. Furthermore, data may come from other sources, and in collaborations there may be other parties who develop a part of the system. Developers will typically need technical documentation for all systems they need to work with, and at the very least high-level information about their data. Finally, individual judgements of AI systems may not always be transparent even to the system's developers, because many AI algorithms are very opaque. This can make debugging, maintenance and further innovations more difficult, and is another reason that interpretability or explainability should be strived for.

vii) Users – The users of an AI system primarily need to know how to operate it, but might also want to know why, how it works, and how well it works. Users might be police employees, but in publicly facing systems they might also be civilians. It is important to explain in a simple and intuitive way how they should interact with the system and what they can do to get help.

6.4 *How much transparency?* – Another question is about *how much* transparency suffices. Perhaps giving everyone full access to everything is not productive, and it can even be dangerous if it lets bad actors find ways to exploit or circumvent the police's AI.

Transparency is a gradual matter, and the same holds for explainability and interpretability: we have to take into account that only parts of a decision may be interpretable, or that explanations only give a rough idea of what happened.

***Box 3 –** Argumentation in the Fraud System*

In the Fraud System (Box 1), for instance, one of the main observations is whether a person has paid or not. This can be identified in the document, as well as another major observation, for example, whether a person has received a product or not. Such observations are put into a argumentation engine that makes recommendations on how to follow up: e.g. ask for more information, discard the report, or send it to a police officer for processing. This argumentation engine is fully rule-based, and its behavior is easily interpretable by experts. The observations that feed into it, on the other hand, can be extracted from the structured and language-based fields of the filled-out form using opaque NLP techniques. This results in a sort of hybrid, partial transparency of the Fraud System's decisions.

6.5 AI scientists within the police regard providing argumentation for system decisions as a central legal issue in the application of AI. The Dutch Police cares about being transparent, as almost all interviewees in the chain of cooperation for law enforcement (hereafter: *ketensamenwerking*) mentioned the importance of transparency in enhancing trust in public for AI use in the Police. The purpose of AI systems deployed in the police is clear for their designers and usage scenarios for users are clearly communicated.

6.6 In order for transparency to be operationalized within the police organization, auditability of AI systems should be ensured by providing traceability mechanisms (documenting the method of building the system). It should be clarified how the system was programmed, and if the system is learning-based, clarification of the method, algorithms and data used for training is also needed. Importantly: how was the data obtained, selected, processed and stored, and was personal data used (11)? For ensuring auditability of AI systems, testing methods should also be documented and cover a wide range of scenarios and metrics, including ones for explainability, privacy, fairness, performance, safety and security.

7. Privacy and data protection

7.1 Privacy is the ability of an individual (or group of individuals) to seclude oneself into a state of not being observed or disturbed by other people, and "an integral part of human dignity" (21). Such seclusion also includes the ability to seclude information about themselves. We refer to this informational dimension of privacy as *data protection*.

7.2 While maintaining order and guaranteeing security, the police often requires infringements on individual privacy (e.g. asking for a driver's license). In the legal doctrine this tension between security and privacy has been balanced by the notion of "reasonable expectations of privacy" (22). For example, a person has a reasonable expectation to be private in their residence or hotel room, while such expectations may not exist in more public areas, e.g. garbage containing sensitive information left in a public place, or when smells can be detected by a drug-sniffing dog. However, where civilians can reasonably expect to be private is being altered by the current technology that allows personal data from many different spheres to be processed on an unprecedented scale, also for law enforcement purposes (e.g. prevention, investigation, detection or prosecution of criminal offences).

7.3 AI can increase the information-gathering capabilities of the police, because of its ability to combine and analyze vast quantities of data from different sources, and therefore has an immense impact on privacy. This impact is heightened because of the speed at which AI can autonomously do computations on a large scale. Although, data processing in the law enforcement domain is closely regulated (23) (24) (Police Data Act), the characteristics of AI raise several issues that go beyond the laws and need ethical considerations.

7.4 Machine learning results may become better by training them with bigger quantities of (appropriate and high quality) data. The principle of data minimization requires that the data use be adequate, relevant and limited to what is required for achieving the intended purpose. This raises a question what is "adequate", "relevant" and "limited to the purpose". It is difficult to give general guidelines for designers how to strike these balances.

7.5 While using ML to find patters in criminal careers, for instance, police must ensure that used data is not older than what the police is allowed to store by data retention rules (see Police Data Act/Wet Politiegegevens) (25). Anonymization of data might seem to provide a solution when training AI to identify general patterns, but it has to be noted that it is usually not enough to prevent privacy violations, as anonymized data almost always allows for re-identification (through combination with other available information). Anonymization as a solution for storing data longer is especially challenged if related (newer) data exist that do not (yet) have to be anonymized.

7.6 AI becomes increasingly effective in identification methods such as voice recognition and facial recognition. These methods have the potential to severely compromise anonymity in the public sphere. For example, law enforcement agencies could use facial recognition and voice recognition to find individuals without probable cause or reasonable suspicion. Also, while there is no great expectation of privacy when walking around a city – we don't necessarily mind that someone who runs into us might know one place where we've been – mass camera surveillance combined with facial recognition makes it possible for companies or the police to track our every move.

7.7 Sophisticated machine learning algorithms may infer or predict sensitive information from non-sensitive forms of data. For instance, someone's keyboard typing patterns can be utilized to probabilistically estimate their emotional states such as nervousness, confidence, sadness, and anxiety (26). Also, sensitive characteristics or dispositions, such as political views, ethnic identity, sexual orientation, and even overall health are inferred on the basis of correlations with data such as activity logs, location data, and similar metrics (27). The person in control of such information could use this to his/her advantage. The police must be mindful of how data is used and might impact users, and ensure full compliance with the Police Directive (23), GDPR as well as other applicable regulations dealing with privacy and data protection throughout the entire life cycle of the AI system.

7.8 It is also worth noting that there can be a tension between privacy and providing transparency about used data to the public or third parties. Furthermore, data protection and storage measures can complicate system development. For instance, there are rules that say certain data can never leave a particular production server, which means that AI developers must either experiment with improvements to their AI system in the production environment (which is not intended or equipped for this), or that they must experiment on their own computers without the real data.

8. Fairness and Inclusivity

8.1 The EU Charter of fundamental rights includes rights of human beings that ensure dignity, equality and solidarity within the union (Article 20-38). By this framework, everyone is equal before the law and any discrimination based on any ground (such as sex, race, color, etc.) is prohibited (Article 20 – Non-discrimination). Moreover, the Charter recognizes and respects the right of persons with disabilities to benefit from measures that are designed to ensure, among other things, their independence and participation in the life of the community (28). This places technical requirements on usability and fairness of systems and data in order to provide equal and non-discriminatory services for individuals, and ensure no groups are marginalized on the societal level.

8.2 To conform with the EU Charter and Dutch legislation (Constitution, the Equal Treatment Act, the Equal Treatment of Disabled and Chronically Ill People Act, etc.) on non-discrimination, it is imperative that the police treats all civilians equally. We interpret this widely as an imperative to inclusivity: to make police services accessible to all, and for the police to treat all civilians they interact with fairly.

8.3 AI systems can play an important role in the inclusivity and accessibility of police services. Take the reporting of a crime for instance. This will be accessible to more people if more reporting methods are available: e.g. in person at a police station, by phone and online. Within the online version, intelligent chatbots can help the user fill in the report by only asking the most relevant questions. This makes reporting crimes more accessible for some by increasing user friendliness and catching errors that might otherwise be made on static forms. See also Box 2.

8.4 One should however be careful that the range of methods offered is indeed usable by all, including e.g. blind people or (computer) illiterate people. If this is not feasible for the main method, alternatives should (continue to) be provided. AI can also increase usability by e.g. adding speech recognition functionality (which can help people who can’t type text).

8.5 Adding more methods is not always better. While overall accessibility may increase, the police have to be careful not to serve one segment of the population more than others. For instance, if the method for reporting crime A (e.g. fraud) – but not crime B (e.g. blackmail) – is improved, the number of reports for A relative to B might increase, and the police’s attention may be allocated more to typical victims of A than to typical victims of B.

8.6 **Bias/Fairness.** In the pursuit of justice, it is of primary importance for the police to treat people fairly and equally. For instance, ethnic profiling is a central point of attention that should be avoided. However, bias is inherent in human judgement and hard to eradicate. And there are cases where some bias and disparate treatment are arguably reasonable: e.g. a young man is statistically more capable of causing physical damage than an elderly lady.

8.7 Given that computers run by objectively executing their instructions, it is often falsely believed that decisions they produce are free from human biases. But even if machine learning algorithms and rule-based systems do not make direct references to human traits (including race, gender, religion, etc.), biases may still slip in through selections made in the data and choices made in the system design. For instance, if there is a rule that depends on how often a person’s family member is named in crime reports (29), this disparately affects people based on the size of their family, which in turn may for example be correlated with their cultural, national or ethnic background. Bias can also easily slip in to AI systems constructed with machine learning based on a dataset with training data. In cases where humans have to explicitly label the data, their personal prejudices and biases may slip in to the AI system. This can only partially be mitigated by having multiple people label the same data, but systematic biases in the population of labelers would persist. The same problem can occur when the AI system is explicitly created to take over a task from humans, where their prior (biased) performance is taken as the ground truth. This particular problem can be solved by creating an AI system that doesn't mimic humans but directly tries to optimize some performance measure, but then it must be ensured that this performance measure is unbiased.

8.8 In some cases, the data represent (a part of) reality exactly as it is, but the resulting model can still be considered biased. For instance, language models often learn to convert words to somewhat meaningful feature vectors in a way so that the distance between these vectors reflects the relatedness of the words. It is found in these cases that words like "doctor" are more closely associated with male words, while words like "nurse" are more closely reflected with female words. Some of this may be due to biased use of language by humans, but it has also been shown that the degree of association roughly matches the actual gender ratio in these professions (30). In such cases one must still ask whether such "biased" data is an appropriate basis for making decisions, especially if we feel that the world as it is, is not necessarily the same as the world as we would like it to be.

8.9 It is also important to look out for sampling bias and interactions with interventions. For instance, an AI system may be used to predict the amount of contraband carried in each section of a city. If the police then send more officers to the high-likelihood areas, it is likely that more contraband will indeed be found there by virtue of the fact that 10 police officers will find more than 0 police officers, *regardless of how much contraband there actually is in each area*. If the new data is then used to inform new estimates, it is essential that it is processed adequately (or that "special" data is omitted) and "feedback loops" avoided. This requires careful logging of not just where the police were sent, but also where they actually went and what they did there.

8.10 Bias may also result from unbalanced or otherwise corrupted data sets. For instance, facial recognition software often works better on white than on black faces. This may be 1) partially due to the fact that cameras are usually configured to work optimally for white faces, 2) partially due to the relative prevalence of white vs. black faces in the data, and perhaps 3) partially due to actual physical characteristics such as lower contrast. The first issue can be mitigated when you are in control of the camera settings (e.g. when collecting data or making mug shots). The second issue can be addressed through the use of more diverse training sets. However, here questions may arise about what is fair: for instance, is it fair to have just as many black and white faces, or should the ratio be proportional to society (with the consequence that the trained AI system might work less well for any minority group)? The third issue cannot be solved directly, but by using more data from the more difficult classes or scoring failure in these cases more harshly, the learning algorithm can be influenced to do better on these harder cases (possibly at the expense of doing worse on easier ones, raising another question of what is considered fair). Another example of measurement error resulting in bias is if the data contains e.g. the locations at which people realized that their wallet was missing, which may not be the locations where it was stolen/pickpocketed.

8.11 While machine learning algorithms themselves are pure applications of mathematics, their implementations may indirectly encode human prejudices about e.g. race and gender.

Inductive bias and training procedures affect what kind of models can be learned and which are most likely to be learned. For instance, if the learned system is constrained to be very simple, this likely means that it will perform relatively well on average cases and relatively poor on exceptional cases (which might correspond to minorities). Or when e.g. hiring someone, a choice might be made between an AI system that measures a candidate's conceptual distance to an idealized candidate (which may look a lot like the average employee), and an AI system that estimates how a candidate would score on some performance measure.

8.12 Many methods have been proposed for detecting and mitigating bias in AI system (e.g. IBM's open source AI Fairness 360 toolkit implements metrics and algorithms from 10 research papers (31)). But what the appropriate measure is for fairness (or bias) depends on the situation and on the interests and perspective of whom you ask. Research on fairness in machine learning has identified over twenty formal sensible measures of fairness, and also that these are mutually incompatible: it is provably impossible for a classifier (whether human or AI) to satisfy all (or even a large number) of fairness definitions simultaneously (32). As such, the police will need to decide (or ask society to decide) what tradeoffs are appropriate to make, and what kind(s) of fairness are the most relevant for each domain-specific application of AI.

8.13 **Predictive policing** practices are aimed at predicting crimes, often followed by an attempt to prevent or mitigate them as much as possible through a variety of measures. Predictive policing is not new and has existed for as long as police officers have had intuitions about crimes that might occur in the future. AI enables a scaled-up data-driven approach to this that can take into account huge amounts of data to, hopefully, deliver more accurate results.

8.14 The most common role for AI in predictive policing is to help in the *predictive* part (as opposed to the *policing* part), but even here it is important to acknowledge the importance of human oversight and expertise. AI is void of common sense and context awareness and may lack information about events that interfere with the predictions. For instance, there may be no method for making the AI system aware of special circumstances, such as a big festival taking place, so it cannot predict the high incidence of pickpocketing at that time and place that would be obvious to any human expert. Or it may indicate a very high probability of encountering contraband in a certain area, which a police officer might realize is due to a stop-and-frisk action in that location in the prior week.

8.15 It is also important to acknowledge that while (other) AI decision support systems could aid in suggesting a course of action in response to the prediction, the actual response remains firmly in the hand of the police professionals who use the system. This is important to acknowledge, because aside from bias and fairness issues discussed above, the main ethical questions surrounding predictive policing have to do with the follow-up. What interventions can the police ethically perform when no actual wrong has been done, but an individual is predicted (exceeding a probability threshold) to commit a crime in the future?

8.16 Interviewees and survey respondents have indicated that it is not desirable to harshly punish people based on what they are predicted to do in the future. The police have instruments of varying severity to use. When the predicted crime is severe enough, preventing its possibility (*pre-empting* the crime) may weigh heavier than the inconvenience caused to the suspected civilian. At the same time, accuracy and validation of predictions for exactly the highest severity crimes such as terrorist activities may be an intrinsic problem because of their relative infrequency. Transparency plays an important role here to inform this highly sensitive decision-making process (see Chapter 6, especially 6.2-ix).

8.17 A complicating factor in this kind of predictive/pre-emptive policing is that interventions can make it difficult to evaluate the AI's accuracy (33). If the police succeed in preventing crime in an area where it was predicted to be likely, this may either mean that the prediction was incorrect, or that the police's prevention efforts were successful – but it will be impossible to tell which. It is also difficult to assess whether crime was actually prevented overall, or whether crime has just moved over to the next neighborhood (which may happen especially if criminal manage to get information on the predictions too). Extensive logging of planned and actual police activities can help with this. AI could potentially help with that as well, although the police employees' privacy should also be taken into account.

9. Human Autonomy and Agency

9.1 Autonomy is the ability of a person or entity to make decisions for themselves. There is a tendency to divide AI systems into a class of autonomous agents that perceive and act in an environment to which they are connected, and a class of passive AI systems that only conduct analyses. However, the AI developers we interviewed (see Appendix 2) did not find this a useful distinction. Autonomy should not be considered as a binary property that is either present or not, but as a property that exists on a scale. It is furthermore not an *inherent* property of an AI system, but heavily dependent on the *context* in which it is deployed, and what decisions we allow it to make without human oversight or control.

9.2 All AI systems are autonomous in some sense: they all "decide" what their output is (under a generous reading of the notion of "decision"). For instance, a speech recognition system "decides" autonomously that a certain audio segment corresponds to the word "bomb". The (perceived) degree of autonomy of the system then depends on what happens after this judgement is made: automatically dispatching the bomb squad to a location (a physical action) would be considered more autonomous than simply displaying a warning on the screen (a speech act).

9.3 Delegating decisions to automated systems can be a great source of efficiency gain, but it is important to consider carefully in which situations it is acceptable. Humans still outperform AI systems in many areas, including commonsense reasoning and taking into account special circumstances and exceptional information. It is often necessary to check if the assumptions of the system hold. The higher the impact of a decision, the more important it is to get it right, and to treat it with respect (see Figure 3). It's important that a human (or organization) always remains accountable, and discrepancies between legal accountability and moral responsibility should be avoided (i.e. it's undesirable to put someone in a position where they are held accountable for decisions, even though they have insufficient knowledge and/or control over them). Transparency in an AI system's decision making can help here, but it is clear that in many cases meaningful human control or oversight is needed.

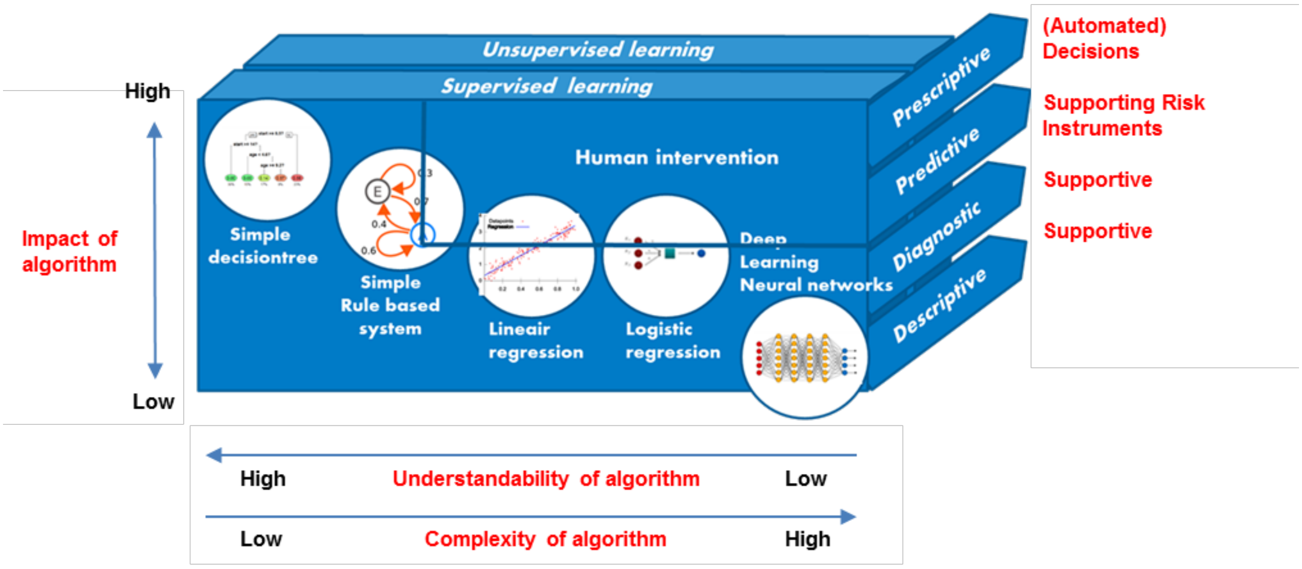


Figure 3: AI Autonomy and Human Intervention (provided by Ministry of Justice and Security)

9.4 Processes with a human in the loop (HITL) involve a sequence of tasks, some of which must be carried out by a human. This means that a human is always there to stop or (partially) correct the process if they are not satisfied with the output of other systems in the process. One downside of this is that having a human in the loop creates a speed bottleneck as computers can perform some tasks much faster, and more at the same time, than humans. Processes with a human on the loop (HOTL) involve a human who can monitor the process and stop (or correct) it if needed. In both situations a human is ostensibly in control, but the question is whether this control is *meaningful*.

9.5 For the control of an AI system to be meaningful, it must meet the technical requirements to provide a human with the ability and knowledge to intervene (34). If a "controlled" AI system is operationally opaque, a human overseer may lack the knowledge to understand how it came to its decision, which impairs their ability to meaningfully evaluate it. It is also important that there is enough time; an advantage of HOTL processes is that they can move faster than HITL, but they cannot move so fast that there is no time for intervention. Automated requests for intervention can help here but must in turn be trusted to operate "autonomously".

***Box 4 –** Control-loop in the Fraud System*

The Fraud System enables a form of meaningful human control because there is always a civilian in the loop during the operation of the AI system. The system asks for the most relevant information, but the civilian is always in control of the information they provide and are asked to verify. Furthermore, the creation of a "panic button" is considered which would let the civilian opt-out of the use of AI and automated decision-making and instead give their crime report to a human police officer.

Based on the impact of the Fraud System on individuals and society, the National Police Lab AI (as well as researchers from Risbo[10]) have determined that there does not need to be a human police officer in the interaction loop between the Fraud System and a civilian. The Fraud System's output does eventually move to police officers who should ensure that it is correct. If functionality is added in the future that lets the Fraud System undertake sorting or filtering (e.g. based on automatically assessed importance) this may mean that not all fraud reports are viewed by human eyes, in which case constant oversight would be diminished. Before such functionality is fully utilized, it is important to thoroughly tested, fair and accountable, and occasional spot checks should be conducted throughout the system's operation.

9.6 Preserving the human sense of agency is mainly an individual-level requirement to realize the EU Charter's universal values of dignity and freedom and should help with both job satisfaction and the ability to provide meaningful human control. Problems can occur with decision support systems that recommend a course of action that must then be evaluated by a human operator. People are increasingly willing and expected to delegate decisions and actions to machines (e.g. recommender systems, search engines, navigation systems, virtual coaches and personal assistants). A possible consequence of working with AI systems is the loss of a sense of agency: the ability to act freely. Especially with systems that are very accurate in some respect, human operators may be "nudged" to act upon the outcome of the system without further critical deliberation (for example, to assess whether a particular case is an exception).

Box 5 – *Human Control in Automated Driving Systems*

The problem of meaningful human control has also been salient in the development of automated driving systems (ADS)[11]. The Society of Automotive Engineers (SAE) has outlined 6 levels[12] of driving automation (35). Many car manufacturers have moved from level-0 (no automation) to level-1 (e.g. adaptive cruise control) to level-2, where the car can steer, accelerate and brake, but constant vigilance and action is required on the part of the human driver (e.g. to detect and respond to objects and events in traffic). However, many car manufacturers have stated that they will skip level-3 and intend to move straight to level-4.

At level-3 the car can mostly drive itself, but it requires human control in exceptional situations. If a car drives itself most of the time but warns the human driver to intervene 1

[10] **Risbo** is an independent institution for research, training and advice, linked to the Erasmus School of Social and Behavioral Sciences. Currently researchers from Risbo are conducting a research to identify impacts of using Fraud System in the police force.

[11] Automated driving systems (a.k.a. autonomous vehicles or self-driving cars) are not currently used by law enforcement, but when the technology matures it could be of obvious use as it lets police officers focus on important other tasks instead of driving.

[12] The SAE's 6 levels of driving automation start from level 0 (no automation), and add functionality at each subsequent level: 1) accelerating/braking *or* steering in limited situations (e.g. adaptive cruise control), 2) accelerating/braking *and* steering (e.g. lane keeping without obstacle avoidance), 3) mostly automated driving but user must intervene in emergencies, 4) fallback requirement is lifted, 5) situation limitation is removed.

second before a crash, this is very likely too late. Similarly, aircraft pilots often prefer to land manually, because the vigilance required to monitor the autopilot and the speed with which they would have to intervene in an emergency are more taxing. A proposed solution for level-3 ADSs is to tell the driver to be constantly vigilant and have their hands on the wheel. This might allow for meaningful human control in theory, but is problematic in practice, because it is incompatible with human psychology: nobody can pay attention for hours without ever having to actually do anything and be ready for a split-second intervention.

For this reason, many car manufacturers want to skip over level-3 autonomy to level-4, which does not require emergency handovers of control: when the ADS is on, there is no human driver – there are only passengers. This also illustrates that there are cases where human control is not the answer (i.e. if the AI system can do better on its own), and that autonomy is not an all-or-nothing affair. Case in point: humans are still in control of the choice of where to drive at all envisioned levels of vehicle autonomy.

9.7 In general, the Dutch police strives to govern AI autonomy by keeping decisions under the overall responsibility of human beings and limiting the systems that do *not* rely on human oversight or control to scenarios were such oversight or control is not at all necessary. Here is a relation with accountability as well: it is important that within the organization a specialized department other than the National Police Lab AI, for example the Ethics Commission, is made responsible for verifying that AI systems can and will be used in a manner in which they are properly governed and under the ultimate responsibility of human beings.

10. (Socio-technical) robustness and safety

10.1 In the discussion of the responsible use of AI, issues of efficacy are sometimes overlooked, because they present themselves as engineering values rather than ethical values. Nevertheless, adequate performance (10.2), robustness (10.4), security (10.6) and safety (10.9) are requirements for the responsible use of AI and upholding ethical principles like fairness, privacy, human autonomy, etc. AI systems must be developed and deployed with an awareness of the risks and benefits of their use, and an assumption that despite ample preventative measures, errors *will* occur. They must be robust to errors and/or inconsistencies in their design, development, deployment and use phases, and degrade gracefully in extraordinary situations, including adversarial interactions with malicious actors. Errors and malfunctions should be prevented as much as possible, and processes should be in place to cope with them and minimize their impact. (11)

10.2 **Performance** – Efficacy has many dimensions which depend on the specific application. Accuracy pertains to an AI system's ability to make decisions, judgements or predictions that are *correct*. Subjecting people to faulty systems can be unethical. For instance, an AI could erroneously discard someone's (valid) report of online fraud, or falsely accuse someone of a crime. If errors are biased against specific groups, or the system is unreliable or inconsistent, this can also lead to (unfair) disparate treatments.

10.3 The police will need to consider what level of efficacy is good enough to ethically deploy an AI system, and what their obligations are towards improving it, depending on the task(s) it performs. A rule of thumb might be that if the AI is taking over a job that was previously done by a human, it should be at least as good. In cases where the AI is slightly worse in some aspects (e.g. context-awareness or foreign language crime reports) and better in others (e.g. in speed, enabling human workers to carry out other tasks), the tradeoff will be more difficult to assess, but the maxim remains that civilians and society should not be worse of when subjected to the AI system. Additionally, there is a moral obligation to improve outcomes if the cost is sufficiently low in order to avoid needlessly subjecting people to faulty decisions. The police or government (possibly in a dialog with society) should decide on the details of this obligation on a case-by-case basis: e.g. if AI system X's accuracy can be improved from 90% to 95% at a certain financial cost, ought the police spend this money to make those improvements?

10.4 **Robustness** – While measuring accuracy is a standard part of the development process of any AI system, testing for robustness is slightly more complicated. Robustness concerns how well an AI system can deal with novel situations and violations of the assumptions on which it was built. For instance, an AI model that only memorizes its training data might only work well on new inputs that are somehow in-between the known data-points and perform very poorly on inputs that are somewhere outside of that space. This can be partially mitigated by having a large data set that is representative of the application domain.

10.5 However, the world is constantly changing: the prevalence of different crimes may rise and fall, criminals may innovate their methods or adapt to police action, and even our language evolves. The assumptions underlying an AI system may become outdated and are sometimes violated outright because the system's designers failed to anticipate some (details) of use cases. It is typically not realistic to expect an AI system to never make any mistakes, but we would like the quality of decisions to degrade gracefully (i.e. make small understandable errors rather than large erratic ones). Robustness can be enhanced by improving a system's ability to generalize to new situations, or by letting it calculate a confidence score about its decisions so it can alert a human (although these confidence scores may themselves be less reliable in novel situations), or by otherwise mitigating the impact of errors. Regular maintenance and updates can avoid AI systems becoming outdated, but it's important to always test new versions thoroughly.

10.6 **Security** – AI systems, like all software systems, can include vulnerabilities that can allow them to be exploited or disabled by adversaries. Malicious actors who engage in hacking may gain access to an AI system, allowing them to monitor or alter its behavior, or to steal or corrupt data or code. Other attacks might simply disable the system or cause it to malfunction. Poor security – by which it becomes possible for unauthorized entities to gain access to AI systems or to (intentionally or unintentionally) tamper with the data – can also result in discrimination, erroneous decisions, or even physical harm. A common argument against self-driving cars is that it would be very dangerous if someone could hack and gain control over them. The police should consider this possibility for all of their (AI) systems.

10.7 Efforts should be taken both to prevent and mitigate the damage of successful attacks. This may in some cases mean that full transparency cannot be given to the public, or that the AI system should not interoperate with other highly sensitive systems. Security can also be enhanced by having (external) safeguards that monitor the AI system for security breaches and enable a preconceived fall-back plan that e.g. alerts humans and/or systems it interacts with, switches the system off, or otherwise changes its behavior (e.g. if the AI's statistical procedures have been corrupted, it could switch to a rule-based approach).

10.8 A whole literature exists on cybersecurity, which we will not go into here. We do recommend that during testing a dedicated (possibly external) “white hat” hacker team seriously attempts to attack the new AI system. We also note that adversarial robustness is a rapidly emerging research field in ML, where researchers attempt to make their AI systems robust even in the face of maximally bad inputs that an adversary might provide.

10.9 **Safety** – The safety of the operators and affected people should always be a priority. Safety is about ensuring that the system will indeed do what it is supposed to do, without harming people, resources or the environment. It includes minimizing unexpected and unintended consequences and errors in the operation of the system and is enhanced by higher accuracy and robustness. However, since errors typically cannot be prevented entirely, careful consideration is needed for how the AI is used, how it can fail, and what it is allowed to control. AI systems should be developed and deployed with potential errors in mind: what is the impact of an error in the situation where the AI is used, (how) can this impact be reduced, and is this acceptable? Processes to clarify and assess potential risks associated with the use of AI services should be put in place. Moreover, it should be noted that the behavior of AI systems can change over time as the system learns and/or the world around it changes.

10.10 An explicit and well-formed development and evaluation process is necessary to ensure performance, robustness, security and safety. Verification and validation play a central role here. Verification is the process of checking that development at each step happens in accordance to the specification and that there are no defects. Formal verification attempts to logically prove this using model-checking tools, but is not always feasible, especially for ML systems (although this is an active area of research). Testing (including unit testing, integration testing, system testing and stress testing) should always be done to gain confidence that the system is also working as intended in practice. Validation involves checking that the system actually meets the desires of its owners. It involves gaining an understanding of the functional and non-functional requirements, as well as evaluating the impacts of (projected and measured) benefits and potential risks. An important question here is whether the system actually saves time, helps catch criminals, improves service to civilians, or accomplishes the ultimate goal that its owners envisioned.

PART III – Conclusions and Recommendations

11. Strategies

11.1 In the preceding chapters we have described specific points of attention for ethical principles we identified as specifically relevant in the police domain. In this chapter we will consider more general strategies that apply across these areas and the entire police organization.

11.2 **Design for values** – Making ethical values a central point in the development of AI applications can help ensure that they are taken into account from the beginning. The design phase still provides degrees of freedom, while it may be difficult to tackle issues in later phases (36). It is important to be aware that ethics and AI design interact in several ways:

- Ethics *by* design: The integration of ethical principles – such as ethical reasoning facilities, fairness and transparency – into the behavior of the technology;
- Ethics *in* design: the regulatory and engineering methods that support the analysis and evaluation of the ethical implications of AI systems as these integrate or replace traditional social structures;
- Ethics *for* design: the codes of conduct, standards and certification processes that ensure the integrity of developers and users as they research, design, construct, employ and manage artificial intelligent systems.

11.3 **Regulation** – It almost goes without saying that all people and organizations should obey the law, and that this holds for law enforcement in particular. While laws often lag behind in that they are not optimally tuned to new technological developments, it is a misconception that there are no laws that regulate AI: AI is regulated by existing laws, just like any technology is, and different laws affect AI's use in different, sometimes unexpected ways. Our interviews indicate that laws are greatly valued, but that some changes may be desirable to further enable the responsible use of AI. In our small survey among data scientists (see Appendix 3), 6 out of 16 respondents felt most major issues (e.g. privacy, non-discrimination and transparency) are sufficiently covered, while 10 thought some unaddressed issues still need to be encoded and 1 thought some issues should be deregulated (respondents could pick multiple answers).

11.4 But legislation is not the only method of regulation. Organizations can put in place their own rules, standards, guidelines, incentives and codes of conduct. International organizations like the International Organization for Standardization (ISO) (37), the IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems and the European Commission's High-Level Expert Group on AI (11) are developing standards and guidelines that could be adopted or adapted by other organizations like the National Police.

11.5 As we have mentioned before, it is important to align incentives in a way that empowers police workers, and don't hamper their work. For instance, programmers should not be held fully accountable for damage caused by a bug in their code, and people should generally be afforded the freedom to contest outcomes of AI decision support systems.

11.6 As a specific consideration, the police Code of Ethics could be supplemented with a specific Code of AI Ethics that states how AI should (not) be developed and used. Our survey showed that only 1 respondent disagreed; 16 out of 20 respondents agreed that this might be a good idea, among them 4 who strongly agreed.

11.7 **Standardization** – AI projects at the police currently tend to be started bottom-up: individual employees come up with an idea for how AI might be used to solve a (local) problem, and this is then approved by their superintendents. The police culture is eager to quickly and creatively solve practical problems. This can result in fast solutions, but also in duplicated efforts and implementations and contracts that are not optimally sustainable.

11.8 Without losing the flexibility that the current approach affords, some central coordination and standardization could greatly aid the efficiency, effectiveness and responsible use of AI. Central coordination within the police organization can aid in making decisions that benefit the corps as a whole, and to avoid duplicating similar efforts across multiple regions. Centralization of expertise is already happening to some degree with the National Police Lab AI[13].

11.9 Standardization across services in law enforcement can guide new projects in the right direction and help with their efficient implementation. Templates for contracts, standardized protocols, project and product evaluation suites, and corps-wide technology platforms can ensure that the police don't get stuck in bad legal partnerships, systems are able to interoperate, and that it's easier to have people from different projects work together. It is also an opportunity to standardize approaches to accountability, transparency, privacy, fairness, autonomy and efficacy.

11.10 **Awareness** – Awareness and education into both AI and ethical challenges can aid in the adoption, trust and responsible use of AI. While awareness of these issues has been growing in the past few years among AI practitioners, it is still common that they are primarily concerned with measures of performance, and unaware of ethical impacts and e.g. biased decision-making. A first step to remedy this is to educate all involved parties on these issues, and to introduce them to potential methods for addressing these concerns.

11.11 Many misconceptions also exist among people that lack a background in AI, both within the police organization and in the wider society. A common view that may be fed by science fiction narratives, is that AI is much more powerful and objective than it currently is. This can lead to great fears as well as mistaken belief in AI as a magic bullet that can solve everything.

11.12 Awareness about the abilities and limitations of AI can furthermore increase uptake of AI within the police organization, with all of the associated benefits. It is also important to emphasize that the goal is not to replace employees but rather to aid police workers in doing their jobs better, taking over boring and repetitive work, and letting them focus on more significant tasks or just perform their job better. Knowing more about the technology can also inspire people to do their part in it: e.g. if a data entry clerk knows the effects of filing police reports in a structured way, they may be more inclined to do so. Awareness of limitations is also important so that police workers realize that they are still needed to provide oversight or

[13] The National Police Lab AI is a collaborative initiative of the Dutch Police, Utrecht University and the University of Amsterdam and aims to develop state-of-the-art AI techniques for improvement of the safety in the Netherlands in a socially, legally and ethically responsible way.

actively control the AI, and so they can scrutinize these systems from the perspective of their different roles and responsibilities in the police organization.

11.13 Finally, efforts may be taken to educate the public about the police's use of AI, and what that means for civilians. Showing that the police is has modern capabilities to perform their duties can increase trust, but trust can also be eroded by misconceptions about the danger and potential misuse of AI. Removing such misconceptions can further enhance trust.

11.14 **Dialogue** – Yet it is not enough to spread awareness: the police should also listen to civilians, to experts and to other organizations. A social dialogue can help the police stay up-to-date and attuned to societal attitudes on what is and is not considered proper and ethical use of AI. Democratic principles should be upheld where affected parties deserve a say in the conduct of the police. On the other hand, it's also important for the police to engage in public discussions to promote their own views and interests in this area.

11.15 The police should also seek out expert partners to ensure that they can stay on top of technological developments. Collaborations with e.g. universities are important for innovation and external validation of police practices when it comes to complex, technical subject matter. The current project is a good example of this, where the police can benefit from a fresh and independent set of eyes, with expertise in AI, law and ethics.

12. Conclusions

12.1 It is impossible to anticipate all effects of the use of new technologies in society, and this also holds for applications in the police domain. It therefore helps if the introduction of new technologies is treated as a social experiment: a process that must be continuously evaluated (38). Like in the case of scientific experiments, those carrying out the experiment have a moral responsibility. This responsibility requires continuous ethical reflection and evaluation around the application of AI by the police, with emphasis already in the very first pilot phase.

12.2 All interviewees acknowledged that AI is shaping the way they will perform their jobs as well as the police-to-civilian interactions. They are generally enthusiastic about AI; however, they also see certain risks and acknowledge the need to handle them. Nevertheless, all of the interviewees highlighted that threats are not posed by the technology itself, but by the way individuals and organizations are going to use it: there is no such thing as responsible or ethical AI, only responsible or ethical *use* of AI. Therefore, they all acknowledged the need for having a *robust ethical framework, where such threats are minimized while cultivating benefits that AI can provide.*

12.3 Interviews indicated a need for more cooperation and a holistic approach to the development and use of AI across the "*ketensamenwerking*". There is a need for frameworks and guidelines for the development and use of AI within the police organization that will tackle all ethical concerns, risks, and vulnerabilities present in the police domain. These need to address the actions and responsibilities of all stakeholders in law enforcement; not just the police, but also the OM, local government and the Ministry of Justice and Security, which sometimes makes high-level decisions on the business processes of the police organization. This ensures that a broad picture of the criminal situation in the nation, as well as the general societal impact, is taken into account while designing the system. In this white paper we have identified some of the ethical concerns such frameworks will need to address, and some high-level recommendations for dealing with them. Specifically, responsible use of AI requires accountability, transparency, privacy, fairness and inclusivity, human autonomy and agency and task efficacy. Aside from specific recommendations regarding each concern, we recommended a number of general strategies for facilitating the responsible use and development of AI: value-sensitive design, regulation, standardization, increasing awareness and dialogue. Future research will need to flesh out the details and require larger projects that involve all stakeholders to build concrete frameworks and guidelines for the ethical use and development AI within the entire law enforcement ecosystem.

13. Recommendations

We give a list of recommendations for the responsible use of AI by the Dutch Police, to ensure alignment with ethical principles applicable in the Netherlands and the EU:

Organization:

Recommendation 1: Create an AI review board within the organization and consider appointing an "AI Ombudsperson" to ensure independent critical evaluation of the use of AI within the organization.

Recommendation 2: Update the "Code of Ethics" in the organization to include considerations particularly important for AI scientists and/or develop clear ethics guidelines for AI scientists working in the organization.

Recommendation 3: Support and incentivize the inclusion of ethical, legal and social considerations in AI research projects.

Recommendation 4: Train AI scientists continually to raise awareness about the ethical considerations and keep them up-to-date on the recent developments in AI and insights about their ethical impact.

Recommendation 5: Develop the redress process for a wrong or grievance caused by AI systems (e.g. an official apology, compensation, etc.).

Recommendation 6: Put clear and fair processes in place for assessing accountability and responsibility for the results of an AI system. Consider taking responsibility as an organization rather than burdening employees with large amounts of liability.

Responsible development and use of AI systems:

Recommendation 7: Install evaluation procedures for the development and use of AI systems that include ethical evaluation (e.g. AI impact Assessment, Privacy Impact Assessment).

Recommendation 8: Develop auditing mechanisms for AI systems to identify unintended effects such as bias.

Recommendation 9: Develop and deploy AI systems taking into consideration that errors will occur. Assess the error tolerance and acceptability in the envisioned task domain, and put in place measures to prevent, detect and mitigate the (possibly erratic and unpredictable) errors. Ensure the safety of operators and people affected.

Recommendation 10: Ensure that used AI systems are sufficiently transparent to enable accountability, usage in courtrooms and the enhancement of trust from the public. This involves using interpretable AI models or investing in their explainability. Taking privacy and security concerns into consideration, find the adequate degree of openness based on the goal the transparency is supposed to serve, and for whom. (see 6.2-6.4)

Recommendation 11: Respect the privacy of individuals. Don’t gather more data than needed, store it securely, and realize that anonymization is an imperfect protection.

Recommendation 12: Ensure that users of AI retain a sense of human agency and feel empowered by the system rather than marginalized.

Recommendation 13: Ensure that humans retain meaningful human control over AI systems. There must always be one or more humans who are responsible, so they must have the knowledge and ability to intervene in the operation of AI systems.

In relation to the societal context:

Recommendation 14: Contribute to a holistic approach for ethical use and development of AI within the Dutch Justice and Security domain that enhances cooperation with the ministry as well as other interlinked institutions (Interpol, UN, municipalities).

Recommendation 15: Support and incentivize research about public perception and understanding of AI and its applications.

Recommendation 16: Continue to join forces with external parties to further develop research and insights in AI and ethics.

References

1. Dignum, Virginia. Responsible Artificial Intelligence: Designing AI for Human Values. September 25, 2017, No. 1, pp. 1-8.

2. Den Hegst, Marielle, Ten Brink, Tjeerd and Ter Mors, Jan. Informatiegestuurd politiewerk in de praktijk. [ed.] Mariëlle den Hengst, Tjeerd ten Brink and Jan ter Mors. Deventer : Vakmedianet, 2017.

3. *The Dutch Police Law (Politiewet).* 2012.

4. Boumans, Joris. *Technologische Evoluties in Wetshandhaving en Legitimiteit: Tussen Optimisme en Onbehagen.* Tilburg : Tilburg University, 2018. Master Thesis.

5. van der Vijver, Kees. Legitimiteit, gezag en politie. Een verkenning van de hedendaagse dynamiek. [ed.] C. D. van der Vijver and F. Vlek. Den Haag : Elsevier, 2006, pp. 15-133.

6. Centraal Bureau voor de Statistiek. Meer vertrouwen in elkaar en instituties. [Online] 5 28, 2018. https://www.cbs.nl/nl-nl/nieuws/2018/22/meer-vertrouwen-in-elkaar-en-instituties.

7. Amodei, Dario, et al. Concrete Problems in AI Safety. 2016, Vol. arXiv:1606.06565.

8. Nguyen, A, Yosinski, J and Clune, J. Deep Neural Networks are Easily Fooled: High Confidence Predictions for Unrecognizable Images. IEEE, 2015.

9. Oosterloo, Serena and van Schie, Gerwin. The Politics and Biases of the “Crime Anticipation System” of the Dutch Police. [ed.] Jo Bates, et al. Sheffield : CEUR Workshop Proceedings, 2018, pp. 30-41.

10. La Fors-Owczynik, K. and Valkenburg, G. Risk Identities: Constructing Actionable Problems in Dutch Youth. [ed.] I. van der Ploeg and J. Pridmore. *Digitizing Identities. Doing Identity in a Networked World.* Rou.tledge/Taylor & Francis Group, 2016, pp. 103-124.

11. AI HLEG. *Draft Ethics Guidelines for Trustworthy AI.* High-Level Expert Group On Artificial Intelligence, The European Commission. Brussels : European Commission, 2018. Working Document for Stakeholder's Consultation.

12. Luciano Floridi, Josh Cowls, Monica Beltrametti, Raja Chatila, Patrice Chazerand, Virginia Dignum, Christoph Luetge, Robert Madelin, Ugo Pagallo, Francesca Rossi, Burkhard Schafer, Peggy Valcke, and Effy Vayena. AI4People—An Ethical Framework for a Good AI Society: Opportunities, Risks, Principles, and Recommendations. November 26, 2018, Vol. 28, 4, pp. 689–707.

13. McCrudden, C. Human Dignity and Judicial Interpretation of Human Rights. 2008, Vol. 19, 4.

14. Dubnick, Melvin J. Seeking Salvation for Accountability. *Presented at Annual Meeting of the American Political Science Association.* 2002.

15. Thompson, D. F. Moral responsibility and public officials. 1980, Vol. 74, pp. 905–916.

16. Hudson, Matthew. Artificial intelligence faces reproducibility crisis. 2 16, 2018, Vol. 359, 6377, pp. 725-726.

17. Gunning, David. *Explainable Artificial Intelligence (XAI).* Defense Advanced Research Projects Agency, 2017.

18. Samek, Wojciech, Wiegand, Thomas and Müller, Klaus-Robert. Explainable artificial intelligence: Understanding, visualizing and interpreting deep learning models. 2017, Vol. arXiv/1708.08296.

19. Ribeiro, T. M., Singh, S. & Guestri, C. "Why Should I Trust You?" Explaining the Predictions of Any Classifier. 2016.

20. Wachter, S., Mittelstadt, B., Russel, C. Counterfactual Explanations Without Opening the Black Box: Automated Decisions and The GDPR. Harvard Journal of Law & Technology, 2017. Vol. 31, 2.

21. European Data Protection Supervisor. *Opinion 4/2015 Towards a New Digital Ethics Data, Dignity and Technology*. 2015.

22. Gorman, Cilian. Is Society More reasonable than You? The Reasonable Expectation of Privacy as a Criterion for Privacy Proection. *Master's Thesis.* University of Tilburg, 2012.

23. European Commission. Directive on the processing of personal data for authorities responsible for preventing, investigating, detecting and prosecuting crimes. 2016.

24. *General Data Protection Regulation (GDPR).*

25. Police Data Act (Wet politiegegevens 2019).

26. Salmeron-Majadas, Sergio, et al. A Machine Learning Approach to Leverage Individual Keyboard and Mouse Interaction Behavior From Multiple Users in Real-World Learning Scenarios. *IEEE Access.* 2018. Vol. 6. doi: 10.1109/ACCESS.2018.2854966.

27. Kosinski, Michal, Stillwell, David and Graepel, Thore. Private traits and attributes are predictable from digital records of human behavior. PNAS, 2013. Vol. 110, 15.

28. Charter of Fundamental Rights of the European Union.

29. Abraham, Manja, et al. *Pilots ProKid Signaleringsinstrument 12- geëvalueerd.* Amsterdam : Ministry of Security and Justice, 2011.

30. Caliskan, Aylin, Bryson, Joanna J. and Narayanan, Arvind. Semantics derived automatically from language corpora contain human-like biases. American Association for the Advancement of Science, 2017, Vol. 356.

31. Bellamy, Rachel KE, et al. AI Fairness 360: An Extensible Toolkit for Detecting, Understanding, and Mitigating Unwanted Algorithmic Bias. 2018, Vol. abs/1810.01943.

32. Verma, Sahil and Rubin, Julia. Fairness Definitions Explained. Stockholm, Sweden : ACM, 2018.

33. Willems, Dick, et al. Predictive policing. [ed.] Mariëlle den Hengst, Tjeerd ten Brink and Jan ter Mors. *Informatiegestuurd politiewerk in de praktijk.* Deventer : Vakmedianet, 2017.

34. Santoni de Sio, Filippo and van den Hoven, Jeroen. Meaningful Human Control over Autonomous Systems: A Philosophical Account. 2018, Vol. 5.

35. Society of Automotive Engineers. *SAE J 3016-2018*. 2018.

36. Aldewereld, Huib, Dignum, Virginia and Tan, Yao-Hua. Design for values in software development. [ed.] Jeroen van den Hoven, Pieter E. Vermaas and Ibo van de Poel. *Handbook of Ethics, Values, and Technological Design*. Dordrecht : Springer, 2015, pp. 831-845.

37. ISO/IEC WD 22989. *Artificial intelligence - Concepts and terminology*. under development.

38. Poel, Ibo van de. An Ethical Framework for Evaluating Experimental Technology. June 2016, Vol. 22, 3, pp. 667–686.

39. Walker, Samuel. *The New World of Police Accountability*. SAGE Publications, Inc., 2005.

40. Dubnick, Melvin J. Clarifying accountability: An ethical theory framework. [ed.] Noel Preston and Charles Sampford. *Public Sector Ethics: Finding and Implementing Values*. Routledge, 1999.

41. Dubnick, Melvin J. Seeking Salvation for Accountability. *Presented at Annual Meeting of the American Political Science Association*. 2002.

42. The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems. *Ethically Aligned Design - A Vision for Prioritizing Human Well-being with Autonomous and Intelligent Systems*. 2018.

43. Cath, Corinne, et al. Artificial Intelligence and the 'Good Society': the US, EU, and UK approach. April 2018, Vol. 24, 2, pp. 505-528.

44. Dignum, Virginia. Responsible Autonomy. 2017, pp. 4698-4704.

Appendix 1 – Interview Setup

The interviews lasted 60 to 90 minutes. The interviews were held in Dutch or in English, at the preference of the interviewee. With written consent of the interviewee, an audio recording of the interviews was made and written into an English interview report, which was then presented to the interviewee for comments and corrections. The audio recording was subsequently deleted. Texts attributed to the interviewees are anonymized and/or slightly paraphrased versions of the literal transcripts. The summaries of the interviews were put up for approval by the police organization before further processing for publication.

Personal Information	Name	Gender	Age
	Organization	Department	Function
	Which role best describes you?		
	What is our working experience/connection with the law enforcement domain?		
Technical Knowledge	What is your conception of Artificial Intelligence (AI)?		
	Do you distinguish between different kinds of AI?		
	Are you generally enthusiastic or apprehensive about AI?		
	How would you describe your own experience and expertise with AI?		
	In your opinion, what are the techniques/functionalities associated with AI that are most relevant (for your daily work and in general)?		
	What is, and can be, the role of AI systems in law enforcement?		
	Can you describe AI tools currently in use / planned to be used in (your part of) the police department / organization?		
	What are the main reasons to adopt (or not adopt) AI in your department / role?		
	What kind of societal objectives do they seek to achieve (organizational, public, political, etc.)?		
	Could the same objectives be achieved by using other means, procedures or mechanisms?		
Impacts and Responsible Use	In your opinion, what are the impacts of using such AI systems?		
	a) Who/what is poised to benefit most and why?		
	b) Who/what is poised to lose most and why?		
	In your opinion, what are the main concerns in deploying AI systems?		

	How do you think that trust can be best ensured in AI systems (e.g. technology, regulation, etc.)?
	What steps or measures do you take or should be taken to ensure responsible use of AI for policing?
	a) If you are involved in the development of the tools?
	b) If you are involved in the deployment?
	c) If you are involved in the use of the tools?
	What is most needed in your department/role to ensure proper implementation and use of AI?
	The police are, based on its special role within society, specifically accountable for its operations. In your view and experience, does the use of AI systems in policing change the ability of the police to account for its actions?
	What accountability mechanisms are in place to address this, and/or where are new mechanisms needed?
Concluding Remarks	

Appendix 2 – List of Interviewees

1	Timo Veldt	Software Engineer at *Ordina*
2	Bas Testerink	AI scientist at the *Police*
3	Tim den Uyl	Managing Director at *Sentient*
4	Michel de Ruiter	Senior Software Engineer at *Sentient*
5	Dennis de Kool	Researcher at *Risbo*
6	Ger Baron	Chief Technology Officer at *City of Amsterdam* (Municipality)
7	Jannine van den Berg	The Police Chief of the National Unit of the *Police*
8	Remco Boercma	Big Data Project Leader at *Ministry of Justice and Security*
9	Lodewijk van Zwieten	v/h High Tech Crime Unit at *Public Prosecution Service (Openbaar Ministerie)*
10	Irakli Beridze	Centre for Artificial Intelligence and Robotics at *United Nations Interregional Crime and Justice Research Institute (UNICRI)*

Appendix 3 – Survey

We conducted a survey among data scientists (21 participants) working for the National Police (internally or externally) to get both quantitative and qualitative responses from a group of experts in both data science and the law enforcement domain. For the purpose of the questionnaire, we defined AI in a very broad way that includes but is not limited to data science, machine learning, big data, pattern recognition, predictive analytics, autonomous systems, inference engines, etc. We were interested in the opinions on AI and ethics of people who work in the police domain and/or the AI / data science field.

You can choose as many answers as you wish and add another open answer if your opinion is not present in multiple choice answers. You can optionally provide your reasoning for your selection after each multiple-choice question.

Age	Gender

Occupation
a. AI Developer / Computer Scientist
b. Data Scientist
c. Software Engineer
d. Police Analyst
e. Department Head / Group Leader
f. Other:

Where Do you Work
a. Police employee
b. Externally hired by police
c. Public prosecution
d. Ministry of Justice and Security
e. International institute
f. Other government
g. Industry, providing services to police
h. Other:

1. What are the most important objectives for the utilization of AI and data science in your work domain within the police? (you can select up to 3 answers)	
a. To make innovative solutions to present problems;	
b. To raise internal efficiency of the organization;	
c. To improve the police's effectiveness at solving and prosecuting crimes;	
d. To enable proactive policing and prevent crime;	
e. To facilitate interactions with citizens and other institutions;	
f. To counter the increasing technological capabilities of criminals;	
g. Other:	

2. What are the most important objectives for the utilization of AI and data science in your work domain within the police? (you can select up to 3 answers)
a. To make innovative solutions to present problems;
b. To raise internal efficiency of the organization;
c. To improve the police's effectiveness at solving and prosecuting crimes;
d. To enable proactive policing and prevent crime;
e. To facilitate interactions with citizens and other institutions;
f. To counter the increasing technological capabilities of criminals;

g. Other:	

3. According to you, what are the most important things that need to be considered when introducing data science processes / AI applications within the police domain? (you can select up to 3 answers)	
a. How to integrate them into the organization (e.g. how it affects hiring);	
b. How they will affect existing workflows (e.g. how somebody's work changes);	
c. How to ensure and respect privacy;	
d. How to comply with laws and regulations;	
e. How to evaluate success;	
f. How to make sure they are fair or fairly applied;	
g. How to make them secure against adversarial actions;	
h. How to explain/understand the system's or process's output;	
i. How to convince others of their importance;	
j. Other:	
Elaborate	

4. Do you think that data scientists and AI professionals should have a "code of ethics" like e.g. lawyers and doctors do?	
a. Yes	
b. No	
Elaborate:	

5. What top 3 ethical principles and human values must be considered in the design, implementation and use of AI systems or data-powered processes that we use to automate or aid us in decision making? (you can select up to 3 answers)
a. Fairness;
b. Equality;
c. Equity;
d. Privacy;
e. Justice;
f. Safety/Security;
g. Health;
h. Happiness;
i. Freedom;
j. Variety;
k. Human Dignity
l. Sustainability
m. Democracy
n. Human autonomy
o. Accountability
p. Rule of law

q. None;	
r. Other:	
Elaborate:	
6. What are the best ways to foster trust in AI systems or data-driven processes? (you can select up to 3 answers)	
a. Accountability of decision makers;	
b. Explainability of output/decisions;	
c. Transparency about how the system was designed and implemented;	
d. Auditability by the general public;	
e. Certification mechanisms	
f. Self-regulation	
g. Laws and regulations;	
h. Demonstration of high performance;	
i. Other:	
Elaborate:	

7. Regulation can be an important tool to implement societal values. What are the main issues that should be encoded in a legal framework for data science? (See the issues and procedures mentioned in questions 5 and 6)	
Elaborate:	

over patrolledpatrolled8. AI systems and statistical models can make judgements that are biased or prejudiced. What are the most likely causes? (top 3)	
a. Labeled data reflects the human biases / prejudice of the labelers;	
b. The model is designed to mimic (historic) human decisions or behavior;	
c. There are feedback loops in how the data is gathered (e.g. more crime is noticed in over patrolled areas, so you patrol them more);	
d. The dataset is too small or too noisy;	
e. The system reflects the human biases / prejudice of its designers / programmers;	
f. The system reflects reality, but we don't want to acknowledge that, so we call it bias;	
i. Other:	
Elaborate:	

9. What are the best ways to deal with the risk of bias in our systems and analyses? (top 3)
a. By abandoning use of biased systems altogether;
b. By creating an algorithm that can adjust for the bias;
c. Through human oversight of produced decisions;
d. By careful checking of our data's accuracy (i.e. whether it reflects reality);
e. By increasing awareness of bias risk among AI experts, data managers, analysts, project managers, users or the general populace; By keeping datasets rich enough, i.e. including sensitive characteristics, to create insight in correlations with seemingly non-sensitive characteristics to better avoid indirect discrimination;

f. By removing discriminating variables from the data (e.g. omit gender/race);	
i. Other:	
Elaborate:	

10. How should the police act on a prediction by an AI system that someone will likely commit a crime in the future, based on observed characteristics and (so far) legal behavior?	
a. Don't do anything: it's unethical to bother someone who has done nothing wrong yet;	
b. It is okay to violate the person's privacy somewhat by keeping an eye on them and collecting data;	
c. It is okay to more explicitly monitor this person, e.g. through tapping their phone or giving them a tail;	
d. It is okay to ask others to keep an eye out or talk to this person (e.g. neighbors or their kids' teacher);	
e. It is okay to send the neighborhood police officer to have a talk with them;	
f. It is okay to stop this person in the street, and check if they have contraband or if e.g. the car they're driving belongs to them;	
g. It is okay to arrest this person in order to prevent the predicted crime;	
i. Other:	
Elaborate:	

11. Consider replacing a highly accurate black-box model with a more understandable one for making important decisions. Under what conditions would you prefer the understandable model?	
a. Almost always: black-box models are near useless for important decisions;	
b. Understandability weighs heavier in the trade-off with accuracy;	
c. Accuracy weighs heavier in the trade-off with understandability;	
d. Only if the understandable model's accuracy is just as good;	
i. Other:	
Elaborate:	
12. Concluding Remarks	