Tynke Schepers

AUTOMATED DECISION-MAKING SYSTEMS AND THE EPISTEMIC CONDITION OF MORAL RESPONSIBILITY

Public sector decision-making with decision aids

Master thesis for the MSc Philosophy of Science, Technology, and Society

University of Twente, Faculty of Behavioural, Management, Social Sciences
Enschede, the Netherlands
21th of July, 2023

Supervisor: dr. Dina Babushkina
Second reader: dr. Yashar Saghai

Table of contents

Summary

The introduction of digital tools has changed the way in which decisions have been made in the public sector tremendously. These systems can vary from rule-based systems, such as decision trees, to machine learning applications, where there are issues related to the transparency, traceability, and explainability of these systems. Indeed, with the introduction of machine learning systems, citizens' expectations have been that decisions made by the government would increasingly be tailored to the specific citizen, therefore going towards an individualised decision. Within the Dutch public sector, however, there are several additional requirements in place which ensure accountable decision-making: the General Principles of Good Governance. Two of these principles are particularly important when talking about making morally responsible decisions with machine learning decision aids, namely the motivation principle and the principle of legal certainty. The first states that the reasoning behind decisions needs to be explained to the applicant, while the second declares that citizens have the right to know what policies or legislation their decisions are based on.

Rule-based systems make it easy for the civil servants to satisfy both principles, yet opaque systems – where the decisions are not easily traceable – prove a bigger challenge. Explanations from opaque systems are always approximations; and instead of an explanation a justification is often offered to the applicant. Where an explanation is a theory or reason why a certain decision was made, a justification on the other hand has a normative aspect: with a justification, one aims to justify why the decision was correct. This does not work in order to satisfy the motivation principle, as citizens have the right to information on a particular case rather than a general approximation or justification. As this information is also not available to the civil servants who have to explain their own decisions, an interesting question regarding the moral responsibility for these decision arises. In most cases, the civil servants have the powers and capacities to work with these systems and make decisions in these cases (that is after all their job). Because they've made the decision, they are also causally connected to it. Additionally, in most cases, the civil servant is aware of the moral significance of their actions (and thus whether they might be committing wrongdoing). It is the epistemic condition of moral responsibility, however, which is causing problems.

When the civil servant has to work with the output from an automated decision-making system, most civil servants are aware of how the system in general works. Yet they cannot retrace the steps for a particular situation, and ex post explanations do not remedy this situation. This has consequences for the awareness of the action itself, the awareness of the moral significance of the action, the awareness of the possible consequences, and the awareness of alternative options available to the civil servant. Furthermore, because the civil servant is unable to fully understand the system, they are also not capable of explaining the system and the consequent decision to the applicant. The applicant

is therefore also impacted in their own right, as they cannot make use of the information to make decisions for their own lives.

There are several conceptualisations of legal mechanisms which propose solutions for this issue, such as the right to an explanation, the right to a justification, and the right to contest. The right to contest would not function without having an explanation, as appealing a decision without knowing exactly what it was based on is impossible to do. The right to a justification also falls short, as this comes closer to an ex-post explanation (meaning an approximation) than a true explanation of the facts. Indeed, only the right to an explanation would solve this issue. However, due to the difficulties in creating explainable machine learning or artificial intelligence systems, this is not a solution either. It is therefore crucial to realise that while the use of automated decision-making systems might make work faster, easier, and more individual, without legal protections in place for both the citizen and the civil servant, it is currently not possible to create a situation where these decisions can be taken while respecting the epistemic condition of moral responsibility.

Preface

Before you lies my thesis on the use of automated decision-making systems and whether this undermines epistemic condition of moral responsibility. This project was the final step in completing the master's degree Philosophy of Science, Technology, and Society at the University of Twente. I have been engaged in researching and writing this thesis for my final semester in this master's programme.

I would like to thank my advisors – dr. Dina Babushkina and dr. Yashar Saghai – for their valuable feedback and support during this project. Your flexibility and creativity with regards to my research focus was wonderful.

Furthermore, I would like to thank my family and friends for their support and inquisitive nature. There was always another question to be asked about my project, and this often encouraged me to look at the subject matter from a different perspective. To my fellow PSTS students, I also say thank you – going through this process together made this a success. It was incredibly helpful to discuss and debate with you (even on different subjects altogether) and I loved hearing about your projects as well.

I hope you enjoy reading this thesis.

Tynke Schepers

Molenschot, 19th of July, 2023

1. Introduction

Towards the end of June 2023, it became known that the Dutch government organisation DUO – 'Dienst Uitvoering Onderwijs' or the Education Executive Agency – used a biased algorithm to find students possibly committing fraud with their scholarship (Schravesande, 2023). Checking whether or not students commit fraud with their financial aid or scholarships provided by the government is something this agency has done for years, by randomly visiting the homes of students to determine whether or not they actually live there, as students are eligible for more money if they are living in student accommodation. Annually, the estimated total cost of the fraud can get up to a couple of million euros. In theory every student can be visited by the agency to check their actual address, yet this has not been the case in practice (Van Bekkum, 2023). Indeed, most of these visits were paid to students with a migration background. The explanation for this: the 'risk profile' used by DUO, which they used to determine which students to visit.

The risk profile has been established in 2011, when the Minister for Education promised to be 'tough' for those students committing fraud with the system (Schravesande, 2023; Van Bekkum, 2023). While several variables such as age, type of education, living situation, and distance from the student accommodation to the parental address and educational institution were included, variables such as ethnicity or migration background were not. Yet this does not mean that in reality this was not an aspect, even though it was implicit – students with a migration background often go to a lower level of education due to language barriers, and that type of education was marked as a higher risk (Heilbron et al., 2023). Students with a migration background often live closer to their parental home than students with a Dutch background, which again equals a higher risk score. When students were said to be a high risk for fraud, it was moreover up to the student to prove that they were not committing fraud, which several court cases have proven is difficult to do.

The algorithm – which gave DUO recommendations on who to check up on, and who was not a high-risk case for committing fraud and could thus be left alone – is no longer in use (Oost, 2023). The current Minister for Education – Robbert Dijkgraaf – has decided that the organisation should stop using the algorithm immediately, and that instead they should make use of a random sample. Indeed, several experts have already spoken up and asked why an algorithm was used in the first place, as a simple model with a couple of variables could also have been used in this case (Schravesande, 2023).

Although the current challenges with the DUO algorithm are the most recent in the Dutch context of public decision-making, it is not the most prolific from the last couple of years. Indeed, the 'toeslagenaffaire' or Dutch childcare benefits scandal is the one that accelerated the discussion on the use of or aid of algorithms for decision-making in the Dutch context (Redactie Volkskrant, 2021).

Between 2005 and 2019, the algorithm was used to help determine which parents were more likely than others to commit fraud with childcare benefits. This led to an estimated 26.000 parents being wrongly accused of making fraudulent claims, and they were required to pay back the benefits they had received (Rutten, 2022). This could go as high as several tens of thousands of euros, which drove families into severe financial hardship. The scandal was brought to public attention in 2018, and led to the fall of the third Rutte cabinet.

In order to get to these risk scores, the government made use of the so-called "Fraude Signaleringsvoorziening", a machine learning algorithm that, like in the case of the DUO scandal, was able to connect different datapoints and make predictions on who could be committing fraud. The connections made between the datapoints were not always valid, which has led to discrimination on a couple of different bases (Amnesty International, 2021). One of these was discrimination based on ethnicity. If a person had a double nationality, then this was almost automatically a reason to investigate this person further, or to get this person placed on a list of (possible) fraudulent applicants. When looking into the entire scandal and how this had gotten to this point, researchers found several dozens of examples where the risk of fraud was based on nationality or appearances (Rutten, 2022). Another point that was added to the database was medical data, or whether someone had a criminal history. While an algorithm was involved in the determination of who could be committing fraud (and thus who should be closely investigated or followed), there were also cases where the information was added by the civil servants themselves, and where they placed citizens on the list of (possible) frauds based on a single characteristic.

In both of the example cases mentioned above – DUO and the 'toeslagenaffaire' – civil servants involved in decision-making on financial aid for citizens were asked to keep a closer look at those who might possibly be committing fraud. Indeed, with the use of these machine learning applications, it was said that it would become easier for them to find those who were committing fraud, as these applications could integrate different datapoints and come to 'smart' and individual decisions, something citizens had been asking for (Diakopoulos, 2016; Frissen, 2023).

1.1. Research question

What was meant as a system that would aid civil servants and citizens alike turned into a scandal, as it turned out that the applications were discriminatory and that people's lives were heavily impacted. Civil servants used the information coming out of these systems in their decision-making, although it is at times unclear where the information comes from, and what the outcome actually indicates. Although decisions are often made by a group of civil servants, or by civil servants across different departments, I am not taking distributed moral responsibility into consideration in this thesis as this would turn into a much larger project. It is therefore the individual civil servant who makes the decision in this thesis, and who is additionally bound by several principles with regards to transparency and explanations for decision-making in the public sector; and the use of these systems could potentially undermine these principles. For the citizen or applicant, who has to make decisions for their own lives based on the decision of the civil servant, these explanations are crucial. Apart from having the right to know what a decision was based on, they also have the right to know what kind of information was used to get to a certain decision.

In this thesis, it is this question of the use of automated decision-making systems in the public sector and what this means for the civil servant (or agent) working with the system and the applicant (or moral patient) dependent on the system that I aim to answer. After all, an applicant has a moral right to an explanation when it concerns decisions they are subject to, which in turn creates a moral obligation for the civil servant to provide this information. Taking into account that most machine learning systems are seen as opaque – transparency is after all difficult to guarantee when working with them, and retracing their steps is basically impossible due to the large number of connections they make – how can we then ensure that automated decision-making systems do not undermine the epistemic condition of moral responsibility?

This can be further divided into several subquestions:

- The first subquestion to take into consideration would be what role automated decision-making systems play in current decision-making processes in the public sector, and what requirements exist for these processes.
- The second subquestion focuses on moral responsibility and the epistemic condition of moral responsibility, and how this can be conceptualised in the case of public sector decision-making.
- The third subquestion looks at possible legal solutions aimed at honouring the moral right to an explanation, and whether these would be sufficient to avoid undermining the epistemic condition of moral responsibility.

1.2. Framework and approach

The main approach that I will be working with in my thesis is the ethics through epistemology approach, as most of the ethical questions concerning the use of artificial intelligence (and, in an extension of that, automated decision-making systems) are based on or connected to epistemological questions. Additionally, I will make use of Van de Poel's (2015) framework of moral responsibility, specifically focusing on an attributive account of moral responsibility. However, as the research question not only focuses on moral responsibility but also touches aspects of legal philosophy, legal studies, and ethics of technology, literature on these topics will be included in the text where appropriate. By using literature on the intersection of legal philosophy, ethics, and epistemology, I will be able to fill a part of the gap that currently exists in the literature on decision-making with technologies within the public sector. Apart from the literature review standard for philosophy on decision-making, moral responsibility, and the epistemic condition of moral responsibility, I will also make use of conceptual analysis to take a closer look at what an explanation and justification actually are, and what the current explanations from machine learning systems can be classified as.

Within this project, I have taken an ethical perspective to look at current legal issues. This combination of both legal studies and ethics aims to give not only new insights into the current debate on ethics of AI, but also take a closer look at current legal solutions for working with automated decision-making systems. Where I therefore refer to 'rights', in this thesis, I am focusing on moral rights except where I have explicitly stated that this is not the case. Furthermore, I am equating the civil servant with the philosophical concept of the moral agent, and the applicant with the philosophical concept of the moral patient throughout this thesis.

1.3. Structure of the thesis

In order to answer the main question of this thesis, I will first take a closer look at the first subquestion on decision-making processes in the public sector and the role automated decision-making systems play in these processes in respectively chapters two and three. Here, I will also go into several issues currently surrounding these systems related to transparency, traceability, and explainability, and give an overview of the different explanations one can get for the functioning of these systems. In chapter four I will focus on the second subquestion, on moral responsibility for public sector decision-making processes. I will give a general overview of the concept first, and then go into detail on the epistemic condition of moral responsibility. The third subquestion, on legal mechanisms, will be answered in chapter 5 with a focus on three different rights: namely the right to an explanation, the right to a justification, and the right to contest. In chapter 6 I will go into detail on what this means for both the civil servant working with these systems and the applicant dependent on this decision. I will end by concluding that in their current form, automated decision-making systems cannot be used by the civil servant without undermining the epistemic condition of moral responsibility, without (legal) protections for the applicant and the civil servant.

2. Decision-making in the public sector

When a person makes a decision or acts in a certain manner, we often state that this person is responsible for this decision or action. Within our every day conversations and exchanges, the responsibility we are talking about often relates to legal or causal responsibility. In cases of wrongdoing, it is however not only the legal or causal responsibility that needs to be taken into account, but also the moral responsibility that might or might not be attributed to the agent or decision-maker. Before being able to establish who is responsible for a certain decision though, it is necessary to look at the decision-making processes specific for the public sector, as there are certain additional requirements such as the General Principles of Good Governance for this sector which have to be taken into account (Rijkswaterstaat, 2019). I will finish this section by taking a closer look at how these different principles require the government to be transparent about these decisions, and how automated decision-making systems can introduce difficulties here.

2.1. Decision-making processes in the public sector

Decision-making is often seen as an activity aimed at problem-solving, specifically at finding a solution to a certain issue that is optimal to resolve a situation, or if that is not possible to attain, a solution that is at least satisfactory for those involved in the process or those subject to the decision (Brockmann & Anthony, 2002). The decision-making process has both rational and irrational aspects, and can be based on tacit and/or explicit knowledge, where tacit knowledge is often used to fill the gaps found within the explicit knowledge.

There are several aspects of decision-making which can make the process complex, such as the variety of different options available, the lack or abundance of information, and the possible consequences of the decision (Van Wart, 1996). This used to be a process for which human beings were held responsible, as they were the ones taking all these different options into account, yet with the introduction of (digital) technologies aimed at supporting decision-making much has changed. One example of these changes concerns the abundance of information available to decision-makers, both for individuals and organisations, for the public and the private sector, and for decisions with different levels of importance (Levy, Chasalow & Riley, 2021; Oswald, 2018). The introduction of digital technologies has also led to criticism; for example in a sketch shown on British television, where the phrase "the computer says no" was used to criticise public sector organisations (Hyde, 2013). A citizen requested information, which was then typed into the computer by the civil servant. The answer to the question, according to the computer, was 'no', and this was how the inquiry was answered. Any attempts to reason with the civil servant for a different outcome were in vain, as the computer had said no, and therefore it was not possible to change the answer the civil servant had given.

Although the sketch where “the computer says no” originally came from Britain, it has been replicated in several languages and cultural contexts to demonstrate the issues of digitalisation within the public sector. Indeed, the digitalisation of public administration has become widespread, and currently digital systems are playing an active role in decision-making (Wihlborg, Larsson & Hedström, 2016). At the same time, decisions are no longer taken within a single pillar – when something changes in the financial aid someone receives from the government, this can also have consequences for their legal situation, educational opportunities, etc. It can be argued, therefore, that individual decisions taking all of these aspects into account are needed, and that digital technologies can be used to do so even if this does make the decision process less transparent because of the increase in factors to take into account (Diakopoulos, 2016; Frissen, 2023). Another argument for the introduction of digital technologies in the public sector concerns the efficiency of these systems and the lower costs. Through automated decision-making the decisions would not only be made faster, but would also be more impartial and would increase equality (Levy et al., 2021). For simple requests this can indeed be the case, as the standard aids for decision-making such as a decision tree could be automated quite easily. For more complex tasks, particularly when talking about complex personalised services for citizens, these standard aids would not be suitable. Yet automating these more complex tasks would result in more financial and temporal gain than automating the simple tasks would.

With the recent technological developments in machine learning, these more complex situations and decisions have to a certain degree been automated as well. Using the COMPAS system in the United States as an example, recidivism risk scores given by the system are used to determine what kind of punishment someone should receive (Oswald et al., 2018). While there is still a human in the loop in this situation – as is the case in most if not all decision-making processes in public administration, especially in the European Union where the General Data Protection Regulation (GDPR) requires a human in the loop[1] – there is also a tendency to trust the system more than other factors which might influence the process. Indeed, even though the system is opaque and it is not possible for those working with the system to retrace the steps to understand how the score was calculated,

[1] Some scholars argue that when a human is indeed in the loop, it is no longer possible to speak of an automated decision-making system, but instead the system is used as a decision support system (see for example Malgieri, 2021). I do continue to use automated decision-making systems here, because if a system ‘makes’ a decision, which the human being later only ratifies without taking other factors into account, or without checking what the basis for the decision is, it can be argued that this is an automated decision-making system still.

many still see the input from the system as more trustworthy than other input (Marcus & Davis, 2019).

Because of the different standards of accountability for the public sector compared to the private sector, the use of algorithms in the public sector is arranged differently (Diakopoulos, 2016). A state's power is regulated through legislation and regulation, and a government is legitimate only to the extent that it is accountable to the citizenry. Yet the use of algorithms is largely unregulated, and as the use of these algorithms can have a huge impact on citizens' lives, they should be accountable to the citizenry (Levy et al., 2021). Holding a government responsible and accountable for the decisions they make – whether on a large scale which impacts groups of people, or on a smaller scale where it might affect individuals – means that the algorithms they use should be transparent and available to the public, just as 'regular' legislation is. Simple decision trees based on legislation are therefore not a problem for accountability and responsibility, as it is possible to trace back each step, yet the machine learning applications can be a challenge (Belle & Papatonis, 2022).

An often heard response when talking about the algorithms and machine learning systems used by governments is that the software is based on intellectual property, and that the information on the algorithm therefore cannot be disclosed (Diakopoulos, 2016). However, there are several models for transparency that can audit and disclose information on the system that is of interest to the public, without conflicting with intellectual property or trade secrets. Additionally, if citizens do not know what laws they are subject to and what they can expect from the government, then the transparency requirement is lost. Citizens should therefore be informed about the system, and what information is used by the system. Fears of manipulation or gaming would even be unfounded in that regard, as allowing people to know what certain decisions are based on might positively influence their behaviour. Of course, knowing what the safety criteria are for smoke alarms and how to ensure that home owners are not fined is a different matter than assigning fraud scores for people applying for financial aid. Yet it is important to ensure that the information on the parameters and the numerical values attached to those is made public, as people are subject to the automated decision-making system (Wirtz, Weyerer & Geyer, 2019). Indeed, in many legislations, citizens have a right to demand information and this has been codified into law. People can appeal to these pieces of legislation to gain access to the information used for the decision-making process, and the process itself.

2.2. Additional requirements for decision-making in the public sector

Both written and unwritten rules play a role within decision-making in the public sector, and in the Netherlands these have been codified into the 'Algemene beginselen van behoorlijk bestuur' (in English: General Principles of Good Governance) (Rijkswaterstaat, 2019). Any public office in contact with citizens and/or businesses needs to follow this code, additionally to other formal legislation that might apply to the particular case. These principles can be divided into three different categories based on where these principles apply, focussing on 1) the preparation process and the processes by which civil servants make decisions, 2) the motivation and design of decisions, and 3) the content of these decisions. Each of these categories includes a couple of the General Principles of Good Governance which have an impact on the way decisions are made, and influence how governmental obligations with regard to the provision of information for citizens are fulfilled. Another division of these principles focuses on whether they apply to the content of the decision (so called 'material principles') or whether they apply to the procedure used to get to a decision (so called 'formal principles') (Jaspers, 2018).

2.2.1. The preparation process and the processes with which civil servants make decisions

Within this first category, three of the General Principles of Good Governance are grouped together. These are the principle of careful preparation, the fair-play principle, and the ban of a *détournement de procédure*. I will explain each briefly.

2.2.1.1. The principle of careful preparation ('zorgvuldige voorbereiding' or 'zorgvuldigheidsbeginsel')

When preparing a decision, the public organisation in question has to identify all *relevant* factors and circumstances, which should be taken into account in the decision-making process (Rijkswaterstaat, 2019). Part of this principle has been formalised in Dutch law, and can be found in article 3.2 of the General Administrative Law Act. When any kind of application is submitted, the public organisation or authority needs to check whether they know all relevant information to make a decision (Jaspers, 2020). If this is not the case, it is up to the applicant to provide the missing documents, after being notified by the authority. Once all necessary information is available, the different interests that are directly involved in the decision are weighed – this is included in article 3.4 of the General Administrative Law Act. In case it is not possible to make a decision, civil servants are tasked to gather more information or clarify information before going further, until a decision can be made.

2.2.1.2. The Fair-Play principle ('Fair-Play beginsel')

The authority making the decision does so without any kind of bias or partiality, as stated in article 2.4 of the General Administrative Law Act (Rijkswaterstaat, 2019). The authority in question is not allowed to deprive citizens of their opportunities to defend their own interests. Any kind of partiality is to be avoided. Additionally, the authority is not allowed to delay or complicate a decision in which a citizen or citizens have an interest.

2.2.1.3. The prohibition of a "détournement de procédure" ('Verbod van détournement de procedure')

The prohibition of a ***détournement de procédure*** means that the relevant public authority is obliged to choose for the procedure that offers the most legal protection to the citizen (Jaspers, 2020). It can be literally translated as a prohibition of the distortion of procedures. A citizen has legal protection when they have the possibility to use a legal mechanism when they do not agree with the decision made by the government. Different legal procedures offer different opportunities for the citizen to make use of, and the government has to take this into account when choosing a procedure (especially when it is not specifically laid down in the law which procedure has to be used) (Rijkswaterstaat, 2019).

2.2.2. The motivation and design of decisions

Within this category, there are two General Principles of Good Governance especially important. These are the motivation principle and the principle of legal certainty (Rijkswaterstaat, 2019).

2.2.2.1. The motivation principle ('motiveringsbeginsel' or 'draagkrachtige en kenbare motivering')

The motivation principle states that the government has to substantiate its decisions in an understandable, sound, and complete manner (Van Goud, 2016). This is based on article 3.46 of the General Administrative Law Act. If a justification for a certain decision is considered to be unsatisfactory, then this can still be supplemented or improved during the objection phase, if a citizen decides to appeal the decision. Decisions that are not based on proper reasoning and motivation even after the government has received the opportunity to add further information can be (and often are) annulled by the administrative court on the basis of a lack of motivation. If, after a decision has been made and a citizen has handed in their appeal, it is shown that a decision has been made based on sound reasoning, the legal consequences of the decision can be upheld.

2.2.2.2. The principle of legal certainty ('rechtszekerheidsbeginsel')

The principle of legal certainty – which encompasses the principle of legal consistency (Engstad, 2017) – entails that citizens must be able to rely on the government or relevant authority to act in a consistent manner, so that citizens have certainty in what the government is allowed to do (Rijkswaterstaat, 2019). Rules set by the government must be complied with and decisions taken by the government must be formulated in such a way that citizens should be able to know at all times what is expected of them in terms of obligations, but also what rights they are accorded. Ambiguous decisions which are open to multiple interpretations are not allowed according to this principle, and the same generally applies to decisions that have a retroactive effect.

2.2.3. The content of decisions

In the third category, several of the General Principles of Good Governance play a role when talking about the content of a certain decision. These principles are the principle of legal certainty (which I have discussed in the previous section, and as such I will not go into detail here), the principle of legitimate expectations, the principle of equality, the prohibition of *détournement de pouvoir*, the principle of diligence, and the proportionality principle (Rijkswaterstaat, 2019).

2.2.3.1. The principle of legitimate expectations ('rechtszekerheid- en vertrouwensbeginsel')

Citizens have certain rights and obligations, but this is also the case for the government itself. This means that the government (which includes municipalities, provinces, and other levels of government) have to comply with a number of rules, and the principle of legitimate expectations is incredibly important in this respect (Jaspers, 2019b). As a citizen, one should be able to rely on an administrative body. If a public authority makes an agreement or promise, then it is essential that these agreements are upheld, as this trust in the other party forms the basis of the relationship between the citizen and the government. When commitments are not kept or agreements are broken, then a citizen can appeal to the principle of legitimate expectations in an administrative court to rectify the situation and ensure that the agreement is upheld (Rijkswaterstaat, 2019).

2.2.3.2. The principle of equality ('gelijkheidsbeginsel')

According to the law everyone is considered equal – see the first article of the Dutch constitution – and this is also the basis for the principle of equality, one of the General Principles of Good Governance in the Dutch context (Rijkswaterstaat, 2019). This principle can be used to correct cases in which citizens are not treated equally, to ensure equal treatment in equal cases (Jaspers, 2019a). Of course, it is difficult to find cases where the situation is completely equal as certain important details for the situation can lead to differing decisions. This means that there often is a good reason for the government to make a different decision in cases that might seem relatively equal, but the government does have the obligation to explain their reasoning in these situations (see also the motivation principle).

2.2.3.3. The prohibition of détournement de pouvoir ('verbod van détournement de pouvoir')

In article 3.3 of the General Administrative Law Act, it is stated that an administrative body may only use its power for the purpose for which it was given this power by the legislator (Rijkswaterstaat, 2019; Jansen, 2018). This means that the administrative body can only act for the public interest (meaning that personal motivations are excluded) and for the public interest only for those specific purposes as envisioned by the legislator. In French, this has been called the prohibition of *détournement de pouvoir*, which can be translated as a prohibition of arbitrariness in English (Van Goud, 2016).

2.2.3.4. The principle of diligence ('materiële zorgvuldigheid')

Decisions made by the government have to be decisions that do the least damage (Rijkswaterstaat, 2019). It is sometimes inevitable that citizens sustain damage because of decisions made in the public interest, yet so long as the administrative body has taken all relevant factors and stakeholders into account when making the decisions, this is something that can be considered acceptable. It has to be noted here, though, that it is often impossible to support (possible) future developments with concrete calculations or other arguments (Burggraaf, 2021).

2.2.3.5. The proportionality principle ('evenredigheidsbeginsel')

The proportionality principle is set out in article 3.4 of the General Administrative Law Act, and states that the (adverse) consequences of a government decision may not be disproportionate to one or more interested parties in relation to the objectives served by the decision (Van Goud, 2016). As government decisions can have drastic consequences for citizens, it is important to include the possible consequences of the decision in the process, even though it is often difficult to fully predict what these consequences might be (Burggraaf, 2021). Nevertheless, if a government decision is disproportionately disadvantageous for one or more interested parties, the government is obliged to compensate this disadvantage (Rijkswaterstaat, 2019).

2.3. The general principles of good governance and transparency

Ensuring that decisions made by public authorities are transparent, just, and accountable is the main goal of the principles mentioned above, yet with the introduction of digital technologies the motivation principle and the principle of legal certainty are particularly important.

The motivation principle requires that the public authorities motivate their decision and show the reasoning behind these decisions (Rijkswaterstaat, 2019). Citizens can then understand why particular decisions have been made, and that gives them the ability to contest these decisions and hold those who made the decision responsible. When decisions are made based on policies or rules established in advance, this is often not difficult to do even though individual situations can differ greatly from each other. Based on the differences between situations – even though this difference might be very small – public authorities can motivate their decision, and going back to the rules explain why certain decisions were made and based on what kind of information this was done. The principle of legal certainty demands that citizens are knowledgeable about their own rights and duties, as well as those of the government (Rijkswaterstaat, 2019). This means that the rules established in advance give an indication, if not precise information, as to how a certain decision is to be made and based on what information and parameters.

With the introduction of decision-support systems or automated decision-making systems, however, complying with the motivation principle and the principle of legal certainty of the General Principles of Good Governance can become a challenge. Decision support systems and automated decision-making systems (both those that are rule-based and those that are working with machine learning or artificial intelligence) are increasingly used by the government to make complex decisions based on large quantities of data (Diakopoulos, 2016). These systems are used for making decisions for taxation, fraud detection (as is the case in the main example used in this thesis – the 'toeslagenaffaire'), financial aid, permits, security issues, asylum seekers, etc. Before going more specifically into the different issues and challenges decision-makers in the public realm face, I will go into the different decision-support systems that can be used.

3. Decision-support systems

Civil servants have long used decision-support systems to aid with decision-making, though these systems have most often been rule-based systems such as decision-trees. Recently, with the introduction of machine learning applications, different decision aids have been introduced within the public sector (Diakopoulos, 2016). Here, I will take a look at both rule-based systems, and opaque systems, following Belle & Papatonis' (2021) categorisation.[2]

3.1. Rule-based decision-making systems

Using algorithms to help with decision-making is not something that has only recently been introduced with automated decision-making systems or decision support systems; indeed, there are several different models that have been used for these purposes that do not rely on machine learning or deep learning at all. To better understand why it is particularly this deep learning or machine learning aspects of automated decision-making systems or decision support systems that are problematic, I will first demonstrate why this is not (always) the case with the models that do not rely on machine learning methods.

Belle and Papatonis (2021) make a distinction in their article on machine learning and explainability, where they focus on 'transparent' versus 'opaque' models. In total, they identify nine different models that can be used to help out with decision-making, and they go through each in detail. The six 'transparent' models, which are those who do not rely on machine learning, are based on mathematical formulas which are accessible for those creating the model and in certain cases also for those working with the models. These mathematical formulas are determined in advance, which means that these models are not asked to create classifications or categories themselves, but rather that they have received all information necessary beforehand. They are asked to simply place the data in the correct box. In these systems, transparency is built into the systems itself. Based on the rules established in advance, it is therefore always possible to explain why the recommendation for a decision is as it is – the rules are already in place. It is important to keep in mind here though that for those not as familiar with these types of modelling, the 'transparent' models might not be transparent at all, as for them these models can also be seen as a black box.

The six transparent models that Belle and Papatonis (2021) call 'transparent' are models that rely on linear or logical regression, decision trees, k-nearest neighbours, rule-based learners, general additive

[2] Belle and Papatonis (2021) are working in the field of computer science, and therefore base their categorisation of decision-support systems on what in the field of computer science is considered to be transparent or opaque. For those not as familiar with the field, the systems that Belle & Papatonis consider to be transparent might not be as clear at all.

models, and Bayesian models. In the table in appendix 1 I have briefly summarised how these models work, and for what purposes they can be used.

The main difference between these systems and the 'opaque' systems – using the terminology from Belle and Papatonis (2021) – is that in the opaque systems, the rules are *not* established in advance. The three models that they identify as opaque are random forest models, the support vector machines, and multi-layer neural networks. As these three models are all working with some type of machine learning, the rules on categorisation or classification are not given in advance. Rather, the algorithm or system is asked to come up with the categorisation itself, based on patterns it finds in the data. In the following section, opaque systems are the main focus.

3.2. Machine learning systems or 'opaque' systems

While the aids for decision-making discussed earlier are more appealing when it comes to transparency as the composition of these aids are known in advance, and people can access the model easily, they are not always the most efficient models. Especially when it comes to predictive accuracy on standard vision datasets, for example, the opaque models can offer insights that the transparent models cannot (Belle & Papatonis, 2021). Here, I will look into three different models – random forests, support vector machines, and multi-layer neural networks that can be used for decision-making – that have a higher accuracy utilising complex decision boundaries at the expense of transparency and explainability.

The first of these models concerns the random forests. These models were initially meant to improve the accuracy of a single decision tree, as these can suffer from overfitting (meaning that the model contains more parameters than can be justified by the data used (Everitt & Skrondal, 2010)) and poor generalisation. By combining these individual decision trees into one model – a random forest – these issues can be addressed to get to a more accurate generalisation (Belle & Papatonis, 2021). To make a random forest, each individual decision tree is trained on a different part of the training dataset which means that they capture different characteristics of the distribution of the data which will lead to an aggregated prediction. Combining these single trees leads to a large and more accurate model, though this can come with a loss of interpretability considering the complexity of the finished model. To gain an understanding of the entire model, one then has to look to ex post explanations. Random forests are for example used in healthcare, where they can be used to predict the disease risk of individuals, based on their medical diagnosis history (Khalilia et al., 2011).

Random forest models do have several advantages over the other opaque types of models. Cutler et al. (2012) state that random forests can model interactions, can handle both regression and (multiclass) classification, can handle missing values in the predictor values, scale well for larger sample sizes, and can work well with irrelevant predictor variables.[3] This gives them an advantage over support vector machines and multi-layer neural networks, as they have difficulties with the aforementioned tasks. The random forest also has several disadvantages; the model is not as accurate as the other two opaque systems, is known to be unstable when the data is perturbed slightly (this can change the tree substantially, which can then have a huge impact on the forest as a whole), and is less useful for capturing relationships that involve linear combinations of predictor parameters.

The second type of model that Belle and Papatonis (2021) categorise as an opaque model is a support vector machine (SVM). These are models based on geometrical approaches; initially used for linear classification but later also for non-linear cases. This made this type of model also suitable to be used for real-life cases. In a binary classification setting, an SVM will categorise data by separating it hyperplane with the maximum margin, meaning that the distance between the different datapoints of each category is as large as possible. SVMs can be used for regression or clustering problems. These models are quite successful in categorisation and classification, but due to their high dimensionality and the potential data transformations that happen within the model, they are also very complex and opaque (Belle & Papatonis, 2021). An SVM can for example help in the detection of economic crimes, and would be trained on the already existing data from different companies (Krysovatyy et al., 2021).

SVMs work particularly well when the data is clearly separated (by a relatively large margin), and in more high dimensional spaces (Dhiraj, 2019). Yet the SVM algorithm is often not suitable for larger datasets that neural networks and random forests are able to deal with, or when the target categorisations overlap. In cases where the number of parameters for each data point exceeds the number of training data samples, the SVM will also not work optimally. Additionally, there is no probabilistic explanation offered by the SVM about the classification or categorisation.

Multi-layer Neural Networks (NNs) are the third type of model that Belle and Papatonis (2021) categorise as an opaque model. This type of model is used extensively for many different purposes; amongst which recommendation systems (van den Oord et al., 2013). Since it is not clear how the different levels interact with one another or what kind of high-level features the model picked up, interpreting the model is difficult to do. Since the theoretical and mathematical understanding of the different properties of these models have not yet been sufficiently developed, they are often also called ‘black box models’. Neural networks consist of several layers of nodes, which connect the input features to the target variable. Each node (on each level) collects and aggregates the output of the layer before, and sends on this new variable to the next level through what is called an ‘activation

[3] For a full overview, see Cutler et al. (2013) page 7.

function'. This process is then continued until the final layer of the model is reached. When the number of layers in this model increases, so does the complexity of the model. If the number of layers remains relatively low, it would be possible to consider this Neural Network as a model that can be simulated, rather than an opaque model. However, because these simple models are not considered to be of much practical use nowadays, most neural networks fit within the category of opaque models. Although neural networks are used in many different sectors – mainly in financial services – they are not yet commonly seen within the public sector (Kosmas et al., 2023), though some have been introduced in the field of cybersecurity (Anaeko, 2019).

The main advantage of a neural network is that it is capable of processing unorganised data, by structuring this into similar patterns (Rawat, 2022). Unlike the other two types of models, neural networks are adaptive in nature, which means that the neural network can change the structure according to the differing purposes. Because they can adapt to different circumstances, neural networks are often left untrained in order to find the structures themselves. This can be both an advantage and a disadvantage, as the neural network can find similarities and patterns in data faster and better than human beings are currently able to do, though some of these patterns might be irrelevant, or in reality non-existing (Babushkina & Votsis, 2022). As this can lead to incomplete results, this can also be seen as a disadvantage of the model. Furthermore, the neural network is also affected by the data that is made available to them. Lastly, due to the adaptive nature of the model, the people training the system and working with it often have little control over the model.

3.3. Issues with machine learning systems

When talking about the output of all of these nine systems as defined by Belle & Papatonis (2021) – meaning transparent and opaque combined – there is often a question about how the output of these systems should be understood. Indeed, for the transparent systems, this question is easier to answer as the more easily understandable mathematical formula or steps to take to get to a decision are established in advance, which means that to provide an explanation as to how one got to a certain decision, one only has to follow this formula or take the appropriate steps. Like legislation in a way, if one is aware in advance of what consequence a certain action or calculation has, the model is considered to be transparent and offering an explanation is possible because these parameters were known in advance.[4]

However, for the opaque systems, due to their complexity, increasing transparency or explainability is not as easy as ensuring that everyone is familiar with the mathematical formula. Indeed, as I have set out above, transparency is a necessary component of an explainable system. I will briefly focus on the

[4] Legislation and regulation is here understood as being created in a top-down manner.

aspects of transparency and traceability – both which are said to be essential for responsible and explainable decision-making (see for example Levy et al., 2021). Then I will look at explainability in more detail – what are necessary components, and what kind of explanations we can get from an automated decision making system or decision support system.

3.3.1. Transparency and traceability

Without transparency about the way a certain system works, it is difficult to trace or even explain how a system got to a certain decision (Ivanov, 2022). The term "algorithmic transparency", a favourite amongst governments and international organisations who state that it is necessary to be open about the algorithms used in decision-making, also refers to transparency about the algorithms themselves (Watson & Nations, 2019). If an automated decision-making system is sufficiently transparent, then it would be possible to trace the (recommendation for) decisions made by the algorithm (Ivanov, 2022). This in turn helps determine who or which system had decided what, on what grounds the decision was made, and who could be held (morally) responsible for the decision made with the aid of the system.

The degree to which transparency is necessary for particular systems is a topic of debate; for example, Ivanov (2022) states that the requirement for transparency is only necessary for systems working with complex decisions. Under complex decisions, Ivanov groups those decisions that "could have a profound impact on people's lives". Systems that are used for repetitive and relatively simple decisions – the example Ivanov (2022) gives here concerns a system which decides whether or not a product fulfils the criteria to be sold – would not require the same level of transparency as those complex decisions would. Günther and Kasirzadeh (2022) make a similar argument, and state that a categorisation of automated decision-making systems is necessary to determine which of these systems require a higher standard of transparency.

3.3.2. Explainability

If the transparency and traceability of a system are both considered to be sufficient, then a system could be considered explainable as well. Explainability has been a topic of discussion for a longer period of time, especially in connection to systems based on machine learning (though not necessarily used for decision-making or offering recommendations). Since work on artificial intelligence and machine learning began, it has been argued that the systems should explain how they got to a result as well (Xu et al., 2019). Where a rule-based system can 'explain' why a card payment was declined, a machine learning system does not have this built-in explanatory mechanism. Indeed, often the higher the accuracy of the system (meaning the predictability of the system and thus the recommended decision), the lower the expainability of the system, as the amount of layers within the algorithm and thus the amount of connections made increases.

Before going into the explanations offered by these automated decision-making systems, looking at what an explanation actually is, is helpful to better understand what explanations offered by automated decision-making systems should consist of. The Oxford Dictionary defines 'explanation' as 'a statement, fact, or situation that tells you why something happened; a reason given for something'. Important to take into account here is that there are different types of reasons and explanations, and not all of them might be sufficient for the applicant to fully understand the decision. Yet without new mechanisms built alongside the system, it is not possible for the neural network, the developer, or current explanatory mechanisms to explain the result or output of the system (Xu et al., 2019). Different architecture of deep neural networks such as convoluted neural networks, recurrent neural networks, or deep feed forward networks are designed for different problem classes and input data, yet all of these lack the explanatory mechanisms which would allow the result to be explainable, as the process by which these systems come to a certain decision can be difficult if not impossible to interpret.

In the last decade, explainability has gained interest from researchers, who have the aim to make the processes of these neural networks transparent. Within these projects, there are two main strands on how to approach the issue of transparency and explainability: some focus on transparent design of the algorithm, and others focus on the ex post explanation mechanisms that might offer some insight into these systems (see figure 1, from Xu et al., 2019). The strand which focuses on the transparency of the system tries to understand the structure of the model (for example the construction of the decision tree or random forest), the structure of the single components of the model (a single parameter used in logical regression), and the training algorithms. The group focusing on ex post explanations on the other hand aims to find why a certain result is the outcome, and does so from the perspective of the users of the system. The aim of an ex post explanation is to give analytic statements, give visualisations, and give explanations by example.

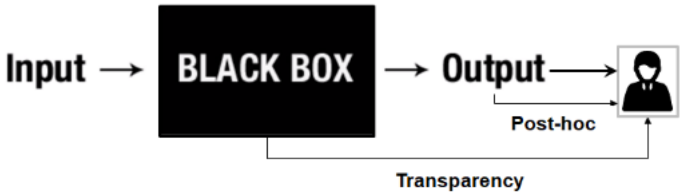

Figure 1: Ex post explanations. (Xu et al., 2019)

3.4. Explanations from decision-support systems

For the opaque models I set out in the previous section, there are several different kinds of ex post or post-hoc explanations (Arrieta et al., 2020; Kim & Routledge, 2018). The terms 'post-hoc' and 'ex post' are often used interchangeably to describe explanations or analyses that occur after the event has taken place or a decision has been made. These explanations are extracted from the model using post-hoc techniques, which is why they have also been called post-hoc explanations (Hamon et al., 2021). In the rest of this thesis, I will be using ex post to refer to explanations given after the decision or recommendation has been generated. Using the example of someone applying for financial aid for child care, the government is using several parameters to determine whether or not someone is eligible for this type of assistance, and even to determine what amount of money they should receive (Klein, 2019). In order to determine this, they are making use of an equation with a number of parameters. While this equation might be easy to describe, it can be very large and the parameters used can be a huge number. As many of the automated decision-making systems or decision support systems are based on large datasets, with as many as a thousand possible parameters or coefficients, it is not possible to fully specify the entire model (Kim & Routledge, 2018).

An ex post explanation is a generic explanation, which might explain the top five, perhaps the top ten, of the different parameters that make up the recommendation (Kim & Routledge, 2018). In the case of applying for financial aid, these parameters could be the amount of money someone makes, the number of hours they work, how high the rent of their flat is, etc. This can provide a meaningful explanation of the algorithm, if these parameters are all that are used as basis for the decision. In more complicated settings, indicating which parameters were used and how important these were thought to be would not be as useful. If the amount of money is further detailed into the annual income, the monthly income, the bonuses someone might receive, and the other sources of income they might have, then not only these parameters would require an explanation (and ranking of importance), but also the different categories themselves. Additionally, if these parameters lead to the discrimination of a particular group, it is not sufficient to merely report the numerical importance attached to these parameters. To sum up; if (some of) the information of these parameters is disclosed, this does not mean that the specific outcome is addressed.

3.4.1. Different ex post explanations

The following is a selection of ex post explanations which can be offered for a non-transparent model (using Belle & Papatonis (2021) categorisation), including automated decision-making systems. These can be used to explain what decision has been made, though here I mostly use them to offer insight into how these explanations explain the outcomes of a model.

- Text explanation: this type of an ex post explanation produces an explainable representation using symbols, such as a natural language text, but can also include propositional symbols which explain the behaviour of the model by defining abstract concepts that represent high level processes (Belle & Papatonis, 2021).
- Visual explanation: here visualisations are used to help the user understand the model (Belle & Papatonis, 2021). Although there are some difficulties here – such as our capacity to grasp more than three dimensions when set out in a visual way – these approaches can be helpful when talking about the decision boundaries or the way certain features of the model interact with one another. Visualisations are often used as a complementary technique, especially when working with a lay audience.
- Local explanations: this type of ex post explanation does not focus on the model as a whole, but rather looks at a specific area of interest (Belle & Papatonis, 2021). This means it is not the model's overall behaviour that is represented, but rather the specific part of the model that the user wants to have explained.
- Explanations by example: this type of explanation takes a representative situation or data entry from the *training dataset* to show how the model operates (Belle & Papatonis, 2021). When comparing these different types of ex post explanations to explanations offered by humans – when they are asked how they would behave in a certain situation, barring any established rules in advance such as legislation – explanations by example are the most similar. Specific examples are used to make sense of a more general process. This does mean that for the example to make sense, the training dataset has to be in a form that is understandable for human beings, since simply offering numbers with several hundred variables would not be considered understandable by many. Another difficulty here is that the training dataset would have to be similar to the data entered later, and that the machine would have to focus on similar patterns as found within the training data.
- Explanations by simplification: for this type of explanations, certain techniques are used that approximate an opaque model by the use of a simpler one, easier to interpret for human beings (Belle & Papatonis, 2021). Of course, there is a reason why the choice was made for an opaque and more complex model, so the challenge in providing this type of explanation lies in the fact that the simple model has to be flexible enough to approximate the complex model (or parts of the complex model) accurately. In most cases, this is tested by comparing the accuracy of the two models by classification problems.

- Feature relevance explanations: this type of ex post explanations attempts to explain a model's decision by quantifying the influence of each input variable (Belle & Papatonis, 2021). This means that the input information is ranked in importance scores; where a higher score means that the variable was more important for the model, and thus for the outcome. In itself, this type of ranking does not constitute an overall explanation, but this can help for gaining insight into the functioning of the model.

These can also be categorised into 'model-agnostic' explanations – meaning a method of explaining a type of model in general – and 'model-specific' explanations – meaning that they can only be used for a specific model, and cannot be used for another. Under model-agnostic, the 'explanation by simplification', 'future relevance explanation', 'local explanations', and 'visual explanations' are included. The 'explanation by simplification' and 'feature relevance explanation' can also be used as model-specific explanations.

Again, it is important to keep in mind that ex post explanations are always an approximation of the model, based on the outcome, and that these explanations do not dive into the specific workings of the models themselves (Belle & Papatonis, 2021). For example, using an 'explanation by example' might give insight into the way a certain decision is established, but it does not give the specific steps taken to reach a certain decision (or risk score, or probability, etc.). This is precisely why the explainability of a model is important, not only for establishing whether or not someone can be held responsible for a decision, but also for the right the applicant or moral patient has to an explanation, in order to make decisions about their life.

4. Responsibility for decision-making in the public sector

Within the Dutch democratic system, civil servants cannot be held legally responsible for their actions and decisions if they are made along the lines of policies set by the government. Instead, ministers are held responsible for these policies and are meant to account for the procedures which have gone wrong – either through policies which were not having the effect they were intended to have, mistakes made by civil servants which led to serious failure, or wrongful conduct of civil servants (De Jonge, 2017). This has also been called 'system responsibility'. While a minister can be sent away from their post for the failure, civil servants do not face these kinds of consequences in this particular case. Although this deals with the political responsibility for mistakes – indeed, in the case of the 'toeslagenaffaire' this lead to the fall of the cabinet and consequently new elections (Rutten, 2022) – it leaves questions about moral responsibility open, especially when talking about those who actually made the decisions, and not only who are assigned the accountability. Leaving aside current legal mechanisms to ascribe responsibility to the ministers, I will look at the issue of moral responsibility of civil servants, and see whether they can be held morally responsible for the decisions they make with the aid of decision-support systems or automated decision-making systems. As I mentioned earlier, I will look at individual civil servants rather than a group, and leave the question of distributed moral responsibility within the public sector for a next project.

4.1. Moral responsibility for decision-making

In order to demonstrate why making decisions with automated decision-making systems is an epistemological issue, I will first set out the different aspects necessary to attribute moral responsibility. In the next section, I will connect this to the use of automated decision-making systems, and focus in further detail on the epistemological questions related to automated decision-making systems.

Stating that someone can be held morally responsible is different from political or legal responsibility, yet remains a fundamental part of our day-to-day lives and interpersonal relationships. Holding someone morally responsible for their actions means that there are certain powers and capacities that we attribute to a person, and judging whether or not this person has 'used' these powers and capacities in the right way based on the way they acted (Talbert, 2022). Although the capacities and powers necessary to act in a certain way can be different based on the context and specific situation, if a person is in possession of these powers and capacities they can be held morally responsible in a general way – as an agent who can be held morally responsible for certain exercises of agency. Human beings in general are seen as agents; meaning that they do have (or are thought to have) the powers and capacities necessary to be held responsible for their own conduct. Very young children who are not yet capable of acting independently, human beings with severe developmental problems, or with diseases such as dementia are generally considered as lacking these powers and capacities (Van de Poel et al., 2015). Before going into the different aspects of moral responsibility in more detail, it is important to stress that moral responsibility cannot simply be equated to causal responsibility – even though causal relations between the action and the (intent of the) agent are necessary to hold someone morally responsible.

Attributing moral responsibility to someone starts with the 'free will argument' (Van de Poel et al., 2015; Talbert, 2022) which has been used to determine whether or not someone was acting out of their own free will, or whether the action was coerced or forced in a certain way. This has also been conceptualised as a form of control; in the sense that the free will argument states that the agent should and could have been capable of acting in a different way if they wanted to (Douglas, Howard & Lacy, 2021). Some scholars argue that the question concerning free will is not something important to take into account when attributing moral responsibility to someone, as everything is already causally determined and someone does not have the free will or the power to act in a certain way (see for example Wiggins, 1973). Those who believe in causal determination state that all events are already set in stone, and that nothing a human being does can change this.

Yet there are also scholars who argue that the concept of a free will is compatible with causal determinism, and versions of this argument have been set out already in ancient Greece. The Stoics (Chryssipus in particular) believed that determinism did not mean that a human's actions are completely explained by external factors (Salles, 2005). Hobbes and Hume have made a distinction between a general way in which our actions might be causally determined, and the specific instances in which it is not possible to act in the way we choose because of specific constraints (Talbert, 2022). The main difference between those who do not believe that causal determinism is compatible with the concept of free will and those who do, is the fact that even though a person's action might be determined in advance, this does not mean that they do not *choose* to act in a certain way. Schlick (1966) made a similar argument, and stated that freedom means the opposite of compulsion – someone is free if they are not acting under compulsion.

An objection to this compatibilist view, somewhat based on Schlick's (1966) argument, is the fact that while someone might have the ability to act as they want, they might still be under serious compulsion (Talbert, 2022). If people are brainwashed, indoctrinated, or manipulated, and as a result of that have certain desires, are they then still acting out of their own free will? The agent might have the capacity to choose to do something different from what they were brainwashed to do, but because of the compulsion they choose not to. This leads to the question whether or not someone under compulsion actually has the choice or the ability to choose otherwise.

Another way to connect the concept of moral responsibility with causal determinism is to argue that moral responsibility does not require free will, or in other words the ability to do otherwise (Talbert, 2022). This connection is demonstrated by Frankfurt (1969), who states that an agent can be considered morally responsible even though they did not have a choice in the way they acted. In this argument, a person (Fred) is considering acting in a certain way, and another person (George) would like to see Fred actually perform this action. If necessary, George can force Fred to perform this action through some kind of intervention in Fred's decision-making process, for example by offering additional information or a new perspective, which would lead to Fred performing the action. Yet George doesn't have to do so, because Fred decides to act in the way George wants for his own reasons. George could and would have intervened if Fred had decided to act in a different way, thus still getting the result that George (and initially also Fred) wanted to get. This means that Fred could not have acted otherwise, but he can still be considered as morally responsible for his actions, considering he acted on a voluntary basis. Others question this line of argumentation, and disagree whether Fred would really be morally responsible for his actions (Talbert, 2022). While Fred might have acted on his own, if the intervention from George had taken place, he would not have acted based on his own reasoning. If someone, next to the power to act, also has the capacity or capability to act, then this means that they are able to act upon this free will (Talbert, 2022).

As I mentioned earlier, there needs to be a connection between the capacities and powers of the agent, and the actual action and possible consequences of this action. This relationship is often causal, and scholars have argued that this should be a central part of the question whether someone can be held morally responsible or not (Van de Poel et al., 2015). Yet a person's capacities and powers are not the same as someone's causal powers, so causality alone is not enough to ascribe moral responsibility to an agent (Talbert, 2022). If someone was not somehow causally connected to the act or possible consequences of the act, then it is difficult to hold them morally responsible. Indeed, even determining to which degree someone was causally responsible is difficult to do, and the problem of many hands[5] increases the complexity of the relationship even further (Van de Poel et al., 2015). This question also relates to the possible consequences of an action, as the agent can be considered as morally responsible for the action, but whether this also includes the consequences is another question (Talbert, 2022). Van de Poel et al. (2015) additionally argue that with the introduction of

[5] The problem of many hands refers to the difficulty to assign (moral) responsibility in cases where many different agents are involved in an action (Van de Poel et al., 2015). I will not go into this further here, as I focus on the individual civil servant.

technology, determining whether or not someone was causally responsible for an action has become even more difficult.

Based on the free will of an agent or the control an agent has over their actions, and the capabilities an agent has to be able to act in this way, it is possible to judge whether or not these actions are blameworthy or praiseworthy (though note that these are not necessarily opposites of each other) (Talbert, 2022). In the case of blameworthiness, it is usually the case that the agent has caused some kind of harm, has done wrong, or that a certain norm has been violated (Van de Poel et al., 2015). This also seems to require that a person is able to recognise and respond to moral considerations, in other words, that the person has moral competence (Levy, 2003). If this moral competence is weakened – for example because of a different type of upbringing, other environmental factors, or due to psychological issues – then it can be argued that this also undermines the moral responsibility ascribed to a person (Fricker, 2010). If someone does not know right from wrong, it can be considered as unreasonable to expect them to act in a way that takes right and wrong into account, which in turn makes it more difficult to hold these people morally responsible for their actions. Another reason for not holding people with an impaired moral competence accountable lies in the fact that they are not able to recognise the moral significance of their actions (Levy, 2007). A *failure* to do something – like failing to be kind to others, or giving up your seat in the bus for someone with crutches because you do not (actively) notice them – is not the same as actively doing something – like actively being rude, or actively deciding to stay seated on the bus.

While issues with moral competency in general focus on people with specific impairments, it does lead to the question whether or not an agent needs to know that something is wrong – either the action itself, or possible consequences of the action – in order to be morally responsible for it. This has also been called the epistemic condition of moral responsibility. This concerns the epistemic or cognitive state of the agent, and asks the question whether or not someone was *aware* of what they were doing – of the action itself, of the (possible) consequences of the action, and of the moral significance of both the action and the (possible) consequences (Rudy-Hiller, 2022). Awareness in this case can be seen in two different ways: this can either concern the content of the awareness (meaning the information that the person performing the action or making the decision needs to have) or the kind of awareness (meaning the mental state of the agent when the decision was made, or the action performed). Assuming that the person was acting out of their own free will, and had the capacity to do so, the epistemic condition focuses on what the person should have been aware of in order to be considered blame- or praiseworthy. There are four epistemic aspects with regards to the awareness that need to be addressed: whether the agent was aware of the action, aware of the (possible or probable) consequences of the action, aware of the moral significance of the action, and awareness of possible alternative ways to act (Rudy-Hiller, 2022).

Before going into the question whether or not a civil servant can be held morally responsible for the decisions they make with decision support systems or automated decision-making systems, it is important to make clear that moral responsibility does not necessarily refer to the duties and obligations of a person – a lawyer can have certain duties and obligations towards a client that do not fall within the scope of moral responsibility (Talbert, 2022). Moral responsibility focuses on the question whether a person has the right or correct relation to their own actions (and the consequences of these actions) in order to be held responsible or accountable for them.

4.2. Moral responsibility and automated decision-making systems

Making responsible decisions with the use of automated decision-making systems and decision support systems should fit within the concept of moral responsibility as described in the section above. I will go through each of the different aspects, and indicate where the digital tools might be a hindrance to attributing moral responsibility to the civil servant.

Determining whether or not the conditions of moral responsibility are undermined when civil servants are making use of automated decision-making systems starts with a closer look at the different powers and capacities the civil servant has. In the case of civil servants, both the powers and the capacities are present. It is after all in their job description to make decisions on certain cases such as whether or not someone is eligible for financial aid. It could be argued that these actions are coerced or forced in a way, as civil servants do have to follow the law and base their decisions on the regulations that are at play. Yet when looking at free will as a matter of control, it could also be said that these civil servants do have a choice in whether or not they are making the choice in general, or whether they are more suitable for another type of job within the public sector.[6] Civil servants can be said to have the powers and capacities necessary to be held morally responsible for their decisions. Using digital systems such as automated decision-making systems and decision support systems does not mean that the civil servant is no longer in control, or no longer has the capacity to exercise this control. In fact, with the introduction of article 22 of the GDPR (European Union, 2016), it has been

[6] Public Civil Service Law establishes a firmly grounded duty to obey – which in turn is subject to a number of exceptions, including the right to disobey or refuse to carry out unethical tasks (Chauvet, 2015). Although the discussion on this legal and moral right (or duty) is fascinating, I have to leave the discussion to the side in this project.

made obligatory to have a human being make the actual decision, rather than leave this to the system itself.[7]

The next step is to determine whether or not the powers and capacities of the civil servant are connected with the actual action and the possible consequences of this action – meaning to take a look at the causal relationship. While it can easily be said that when someone decides on a course of action, the actual action and the possible consequences of the action are causally related, this does not always have to be the case. If someone applies for financial aid and the civil servant decides that they are eligible, the possible consequences can vary from the applicant using the money to pay their rent to using the money for something else entirely. After all, it is the applicant who in turn decides for what the money is actually used, which is outside of the purview of the civil servant. In this causal relationship little changes with the introduction of automated decision-making systems or decision support systems, mainly also because of the legal stipulation that a human being should be the one to actually make the decision, rather than such a digital system.

Whether or not one can talk about wrongdoing remains an issue to be determined on an individual basis, when talking about specific decisions. If a decision was made which can be considered as morally wrong, then this does not change much with the introduction of automated decision-making systems or decision support systems. Asking the question overall, meaning whether it is morally acceptable to make decisions on which people's livelihoods depend with the use of such a system, is another issue entirely. Here again the GDPR plays a large role, as humans have to be present and involved in the decision-making process.

The epistemic condition of moral responsibility makes the issue of attributing moral responsibility to someone more complicated. When working with a simpler decision-making aid such as a decision tree, or other rule-based systems, it can be stated that there are sufficient conditions for the civil servant to be aware of what they are doing as they can check and recheck the algorithm or decision aid at any point. After all, the rules are established in advance through law, regulations, or policies, and merely translated into decision-rules used by these systems. When working with more complex models though it can already be difficult to keep track of all the different parameters playing a role in the process. Working with opaque models complicates the issue even further, as it is much more difficult to fully understand and be capable of explaining a certain decision when the system is identifying patterns, making connections between datapoints, and determining what the best course of action would be. Because these systems are often not transparent nor the decision traceable,

[7] While the argument on artificial agency is still a fascinating one to have, in practice it is currently not allowed in the EU to work with only these systems. A human being is still the final one to look at the information, and make the final decision. To read more on artificial agency, see Shapiro (2005).

determining which connections have been made and what numerical value has been assigned to each of the parameters becomes an approximation. This is not the same as a full explanation. One way to determine whether a system can be sufficiently understood in order to say that the civil servant was aware of what they were doing when using the system for decision-making purposes, is to see whether ex post explanations satisfy the epistemic condition of moral responsibility. Before doing so, however, I will first give a more detailed overview of what constitutes the epistemic condition of moral responsibility and what specific elements have to be kept in mind for evaluating the different explanations.

4.3. The epistemic condition of moral responsibility

The main element of concern when it comes to moral responsibility and decision aids is the epistemic condition of moral responsibility. Here, I will dive deeper into the different parts of the epistemic condition. To briefly recapitulate; the epistemic condition asks the question whether someone was *aware* of what they were doing – whether this concerns the action itself, the moral significance of the action, and even the consequences of the action (Rudy-Hiller, 2022). In this section, I will set out in more detail what the epistemic condition of moral responsibility entails, and how different interpretations of the epistemic condition have shaped the debate. I will focus on two parts of awareness here, as it can be seen in two different ways: it can concern the content of the awareness (in other words, the knowledge an agent needs to have or the things an agent needs to be aware of) or it can concern the kind of awareness required (meaning the mental state of the agent when the action was performed or the decision was made) (Rudy-Hiller, 2022). After this overview of the literature, I will apply this to the use of automated decision-making systems in the public sector.

4.3.1. Content of the awareness

When taking a closer look at the content of the awareness, the philosophical literature on the subject has been subdivided into four main segments: the awareness of the action, the awareness of the moral significance of the action, the awareness of the consequences of the action, and the awareness of possible alternative ways to act.

4.3.1.1. Awareness of the action

The first aspect I will focus on is the awareness of the action, which in other words states that in order to be held morally responsible for an action, the agent needs to be aware of what they are doing (Rudy-Hiller, 2022). If a person is ignorant of what their action means for either themselves or other people, it could be stated that they are not morally responsible for the action. Of course, this then introduces the question whether the agent *could* or *should* have known what the action is they are performing, and what the action would lead to (Van de Poel et al., 2015). This is the normative aspect of the epistemic condition of moral responsibility. For an agent to be directly blameworthy or praiseworthy of an action or decision, they need to be aware of what they are doing. I will go further into this aspect of the epistemic condition in the section on capacitarianism.

4.3.1.2. Awareness of the moral significance

The second aspect related to the content of the awareness is the moral significance of the action (Rudy-Hiller, 2022). In order to fulfil this, the agent should have a belief that the action is morally wrong, or that certain aspects of the action make the action morally wrong. That the entire action is morally wrong is a *de dicto* awareness of moral significance. This involves the awareness of the action's wrongness, connected to the normative statement that these kinds of actions are wrong. A second type of belief, about the specific aspects of the action that are morally significant and can be considered wrong, is called a *de re* awareness of the moral significance of the action. Some argue that *de dicto* knowledge is necessary for moral knowledge and thus moral responsibility (see for example Sliwa 2017), while others focusing on the *de re* awareness deny the necessity of moral knowledge for moral responsibility. Quality-of-will theorists have focused a lot on this issue, and I will explain in more detail later what exactly the different positions are, as this also relates to the different kinds of awareness.

4.3.1.3. Awareness of the consequences

The awareness of consequences is the third aspect, and states that an agent is not only responsible for the action itself, but also for the (possible) consequences that the action may have (Rudy-Hiller, 2022). This states that the agent should have some kind of belief about the consequences of his action. Some, like Zimmerman (1997), argue that this belief must be specific, meaning that if someone were to push a button on the wall which electrocutes someone else, then the person pushing the button must have the *specific* belief that that is the consequence of pushing the button. Others state that a more general belief is enough to be held responsible; meaning that if the person pushing the button had known that it would hurt someone else (though not specifically electrocute them), this would be enough to hold them morally responsible.

4.3.1.4. Awareness of alternative actions

The fourth and last aspect concerns the awareness of alternative actions (Rudy-Hiller, 2022). Some philosophers (such as Levy, 2011) write that an agent cannot be considered blameworthy for a wrong action if they do not believe that there was an alternative (and morally permissible) course of action open to them. However, not everyone believes this to be a necessary requirement. Going back to the Frankfurt cases discussed in the section on free will and moral responsibility, Sartario (2017) states that an agent could be considered blameworthy even if they believe that there were no other options, so long as this does not interfere with their own choice about whether or not to act.

4.3.2. Different kinds of awareness

In addition to the question about the content of awareness, there are also different aspects related to the kind of awareness. These focus on the mental state of the agent, and ask the question what mental state they should possess in order to be aware of their actions. Rosen (2008) presents the case of a man trying to poison a woman: he has bought arsenic from a local apothecary, and poisons the woman's tea with it in the belief that if she drinks it, she will die. Now, if the apothecary sold the man sugar instead of arsenic, but did not tell him – in other words, the man believes that what he bought is arsenic – he can still be held morally responsible for the attempted poisoning, even if his belief cannot be called knowledge. Others state that it is not a true belief which satisfies the epistemic condition of moral responsibility, but rather having a reasonable or justified belief (Timpe, 2011).

Assuming that awareness involves some type of belief, as set out in the paragraph above, means that it is necessary to consider how these beliefs are entertained by the agent, and in which way they have to be entertained in order to have the relevant awareness for moral responsibility (Rudy-Hiller, 2022). There are two main responses to this question, namely those who state that the belief has to be entertained *occurently* (for example Rosen, 2004; Levy, 2011), and those who argue that the beliefs can also be *dispositionally* entertained (Timpe, 2011; Levy, 2013). Entertaining a belief occurently means that an agent only satisfies the epistemic condition of moral responsibility if, while performing the action, they consciously believe that their action is right (or wrong) and if they have also taken some of the possible consequences into account (Zimmerman, 1997). If the agent hasn't done so – meaning that they are completely ignorant of all these contemplations concerning both the action and the possible consequences – then the agent would have an excuse for their wrongdoing, according to those who believe a belief should be entertained occurently.

Those who argue that a belief should be entertained dispositionally state that the 'occurist' position is too strict; because if someone forgets an essential piece of information to a particular situation – for example that someone is allergic to a certain type of food – one can intuitively be seen as morally responsible for the situation, even if they did not have the belief that they were doing wrong (Amaya & Doris, 2015). Taking this one step further, it could be argued that people can avoid moral responsibility by merely not thinking about the moral status of one's action. Information that is not consciously thought about (or is 'dormant') can then also be considered as sufficient awareness for attributing moral responsibility.

While the idea that some kind of knowledge or awareness of the action is necessary to hold someone morally responsible, this can also threaten to undermine the issue of responsibility (because what to do if the agent forgot the information?). Zimmerman (1997) originally started this line of argument, although several others (see for example Rosen, 2004 and Levy, 2011) have also taken this up and developed versions of it further. Rudy-Hiller (2022) calls those in favour of this argument the 'volitionists', and the strand of argumentation or the position they take 'revisionism'. This position takes a closer look at the issue of ignorance, and whether this ignorance can be considered culpable or blameworthy. When I introduced the epistemic condition of moral responsibility, I mentioned that one of the aspects of this condition concerned whether the agent *could* or *should* have known what the action is they are performing, and what the action would lead to. Volitionists look at the terms which would be necessary to determine whether an agent was culpable for their ignorance.

To answer this question, volitionists look towards the thesis of doxastic involuntarism: we lack direct control over our own beliefs; we do not decide ourselves what to believe (Rosen, 2004). If an agent is to be considered morally responsible (or culpably ignorant) then this should be because 1) they did something that they did have direct control over, 2) it caused them to have or lack certain beliefs, 3) doing this thing is considered to be wrong, and 4) they are blameworthy for having done it.[8] Smith (1983) states that something that fulfils the first three criteria can be a 'benighting act'; where the consequence of a certain action can be traced back to a true belief about the acceptability or permissibility of that person's action. Actions such as not reading the manual about the buttons on heavy machinery or not asking someone what the purpose behind certain safety rules is, are in this situation benighting acts. An agent would have direct control over the decision to read the manual (or not) and them not doing so leads to having an incorrect belief about the workings of the machine. This can be considered as wrong, as reading the manual for machines in a workplace is considered important, especially if the agent is in charge. But acting in the wrong way does not immediately equate responsibility, which is where the fourth condition comes in and where the question whether the agent is blameworthy for having acted the way they did is considered.

Zimmerman (1997) states that in order to answer this fourth question – whether the agent is responsible for having committed the act – is subject to the same questions as moral responsibility in general. If an agent needs to be aware of what they are doing to be held morally responsible for their actions, they can only be held responsible for a benighting act if they were aware of what they were doing, what possible consequences of this act would be, what the moral significance of the act was, and what alternatives were available to them. This can then turn into a circle, especially if the

[8] Interesting to note here is that the idea of responsibility for ignorance is also something derivative.

benighting act was due to another benighting act, and so on. To end this regression, there would at some point have to be an act that the agent committed in full awareness of *all* the relevant (moral) facts.

Indeed, volitionists state that this does not only apply to moral ignorance, but also to other kinds of ignorance such as factual ignorance (Zimmerman, 1997; Rosen, 2004). If someone does not know that pushing a button which electrocutes someone is wrong – if they do not believe that harming someone is wrong – then they do not believe that their actions are morally wrong. According to volitionists, if the person is unaware of what they were doing wrong, then they are only blameworthy for this action if they are blameworthy for their moral ignorance, and they are only blameworthy for this ignorance if it is derived from a blameworthy benighting act. If these conditions are not satisfied, then the person is excused from blame. Yet most philosophers do not find this line of argumentation plausible, as it does not seem feasible that the requirements for the epistemic condition are so strict that most ordinary people who do wrong fail to meet them, and are thus not blameworthy for their actions (Talbert, 2022).

To briefly summarise, there are four main points of the regress argument (Rudy-Hiller, 2022). First, an agent is only blameworthy for a wrongful act they committed out of ignorance if they are culpable for that ignorance as well. Second, ignorance is only culpable if it comes from a benighting act which was performed in full awareness. Third, the agent needs to have a relevant awareness for both the benighting acts and their actions, and the action's moral significance. Fourth, these three principles mentioned above apply to all kinds of ignorance; not merely factual ignorance.

4.3.3. Responses to the issue of awareness

As set out above, volitionists state that responsibility requires awareness, and argue that awareness is the main component of blameworthiness, rather than for example causal reasons. Some philosophers (partially) agree with the statement that responsibility requires awareness, but make an exception for the necessity of the belief that the action needs to be morally wrong (Rudy-Hiller, 2022). Others claim that while an awareness of wrongdoing is required to hold someone responsible or culpable for their actions, this is not necessary when assessing the culpability for the ignorance itself. A fourth group states that the required awareness is only necessary for the factual aspects of the action, and there is another group that disregards the issue of awareness altogether and states that one can be held responsible for their actions so long as the other criteria of moral responsibility are satisfied. I will look at each position in more detail, and set out what this means for the epistemic condition of moral responsibility.

4.3.3.1. Weakened internalism

The most conservative strategy for refuting the volitionist interpretation of the epistemic condition – which looks at the question of ignorance in the debate on ascribing moral responsibility – would be to accept the main parts of the arguments, but to disagree with the last component which states that an *occurrent* awareness of wrongdoing is absolutely necessary to attribute (moral) blameworthiness to someone (Rudy-Hiller, 2022). A weaker kind of awareness, such as an unconscious belief or a dormant belief – would also be enough to satisfy the epistemic condition of moral responsibility (Levy, 2013). Others take a similar approach and deny entirely that the content of the awareness should include a belief that the action is overall considered to be wrong (or right). Instead, there are other beliefs that people consider to be sufficient for blameworthiness. These include a belief that there are reasons not to perform the action (Robichaud, 2014), a belief that one is acting based on morally reproachable reasons (Sartorio, 2017), or a belief that there is a non-negligible risk that one's action is wrong (Guerrero, 2007). This view has also been called 'weakened internalism', as it states that the epistemic condition requires the presence of a certain mental state that looks at the action's moral worth when the agent performs the action (Rudy-Hiller, 2022). It is called weakened because non-occurrent mental states would also be able to satisfy the epistemic condition, rather than only looking at beliefs where the agent decides to act despite knowing better.

Zimmerman (1997) defends the volitionist argument and states that if a belief is not occurrent, then one cannot act with the intention to listen to this belief (or with an intention to *not* listen to this belief). This means that if someone does not act deliberately with regards to this belief, then the belief does not play a role at all in the action. A response to this defence states that occurrent beliefs are not the only ones that play a role when determining the reasons for a certain action, and therefore it cannot be the case that an occurrent belief is necessary for knowingly doing wrong (Peels, 2011). Rudy-Hiller (2022) gives the example of making the decision to go to a park for a break from work – one can occurrently believe that this break from work is necessary, while at the same time (dispositionally) believing that the park is a nice place to go to for a break, that there are often amusing events taking place there, etcetera. These kinds of beliefs also play a role when making a moral decision; when someone is planning to push the button and electrocute someone else, the beliefs the agent is entertaining can focus on the different aspects of their plan and the different ways in which the plan can be carried out without anyone else knowing or noticing. There might be a dispositional belief that what they are doing is wrong, and this can also play a role in the way the agent acts, but this is not occurrent. This means that it would be possible to hold the agent responsible for their actions, despite the fact that they were not *occurrently* entertaining the belief that they were doing wrong (i.e. that electrocuting by pushing the button is wrong).

Weakened internalists also offer another line of argumentation to deny that the content of the awareness must involve a belief about the action's moral significance or moral wrongness. Guerrero (2007) argues that if an agent is not sure whether the action they are planning to perform is morally permissible – therefore admitting the possibility that it might be considered wrong – can be blameworthy for still choosing to perform the act, even if their ignorance (which here is presented as uncertainty) is blameless. Nelkin & Rickless (2017) present a similar argument, stating that being aware of doing something that poses the risk of forgetting essential information can still be considered as responsibility for unknown wrongdoing. Robichaud (2014) claims that if there is a sufficient (even though this might not be decisive) reason to perform an action this satisfies the requirement of awareness of moral significance. According to Robichaud, this means that it would also be possible to be considered blameworthy for an action for one which has decisive (even though these might not be sufficient) reasons to avoid doing.

4.3.3.2. Ignorance and epistemic vices

The second group I will discuss in this section states that awareness of wrongdoing is required to hold someone responsible or culpable for their actions; yet this is not necessary when assessing the culpability for the ignorance itself. Some philosophers state that wrongdoers can be culpable for their ignorance and the actions they performed in this ignorance, even if the benighting acts themselves were performed out of ignorance of their wrongness (Rudy-Hiller, 2022). An agent might have epistemic vices that blind them to reasons to improve their epistemic situation. This leads to an interesting position: while attributing blameworthiness for ordinary actions does require an awareness of wrongdoing from the agent, blameworthiness for benighting acts and ignorance stemming from these acts does not. Blameworthiness for actions performed out of ignorance is therefore always derivative, unless these actions were benighting acts.

FitzPatrick (2008) states that ignorance is culpable when it comes from the violation of epistemic obligations, which the agent could have reasonably been expected to remedy (such as arrogance, overconfidence, or laziness). Levy (2009) argues that agents are blameworthy only "if it is reasonable to expect them to conform their behaviour to the appropriate normative standards" so long as "conforming their behaviour to normative standards is something they can do rationally (and not merely by chance or accident)" (page 735). Montmarquet goes further than FitzPatrick by claiming that beliefs themselves can be "fundamental and underived" (1995, page 43). In his view, the agent always possesses a direct (though incomplete) control over the formation of their own beliefs. Montmarquet agrees with the volitionists' point that if an agent is considered culpably ignorant of a certain fact, there will always have been a certain way in which this could have been remedied, yet he argues that talking about benighting acts does not get us to the true source of culpable ignorance.

This, and the beliefs associated with this attitude, is according to him an “intellectually irresponsible attitude”. In claiming that agents are capable of controlling the formation of their beliefs, Montmarquet goes against Zimmerman (1997). Montmarquet gives the example of whistling: the agent directly controls the whistling itself, and the effort they exert in whistling (1995). If this is true for beliefs, the agent is directly responsible for the creation and formation of beliefs, as well as the beliefs themselves. Zimmerman (1997) states that whistling, unlike the formation of beliefs, is an action. Beliefs are a part of our mental state, which can be the result of the action that we have control over, but we do not have control over the beliefs themselves.

4.3.3.3. Quality of will

A third group of philosophers resisting the regress argument are those who appeal to the ‘quality of will’ argument (Rudy-Hiller, 2022). While there are different ways to define ‘quality of will’, the basic idea behind it is this: an agent is praiseworthy for an action (or attitude) that corresponds with the demands of morality so long as the performance of the action arises from proper regard or concern for another person’s morally significant interests (Shoemaker, 2013). Turning this upside down means that an agent can be considered blameworthy for an action that does not accord with the demands of morality so long as the performance of the action arises from a *lack* of proper regard or concern for another person’s morally significant interests. This means that an agent does not need to believe that an action is morally right or wrong for the action to display their quality of will, and thus for the agent to be praise- or blameworthy for the action. The permissibility of the action is dependent on the other person’s interests, and not the agent’s beliefs. A *de dicto* awareness of an action’s moral significance is therefore not required for blameworthiness or praiseworthiness.

The question then arises what kind of awareness is considered necessary by quality-of-will theorists; and Talbert (2013) states it should be considered “what a wrongdoer needs to know in order for her actions to express the attitudes and judgements that make blame appropriate” (page 242). For this, a factual awareness of the situation is necessary, as the agent has to be aware of what they are doing and what possible and probable consequences of their actions are. While quality-of-will theorists agree that a *de dicto* awareness of the moral significance of an action is not required, they disagree on the necessity of a *de re*[9] awareness (Rudy-Hiller, 2022). Some state that it is necessary, while others state it is not. Sliwa (2017) states that agents can only be considered praiseworthy if they *intentionally* do the right thing, and that without this intention present they cannot be praised for a certain action. Denying that a *de dicto* awareness of the action’s moral significance is necessary for praise- or blameworthiness means that quality-of-will theorists ultimately have to deny all four parts

[9] Awareness of the action’s right-/wrong-making features regardless of whether one conceives them as such

of the regress argument. After all, if moral knowledge is not required to ascribe praise or blame, then moral ignorance does not stand in the way of praise or blame either.

4.3.3.4. Capacitarianism

The fourth and final type of response to the regress argument that I will discuss here is one that does not only focus on moral ignorance, but also factual ignorance. According to this strand, agents can be directly blameworthy not only for actions they committed out of moral ignorance, but also for actions committed out of factual ignorance (Rudy-Hiller, 2022). Blameworthiness is explained largely through the presence of certain capacities which the agents possess, which makes the agent capable of acquiring the relevant awareness. This also accounts for the name of this position: capacitarianism. Agents satisfy the epistemic condition if they are aware of the relevant moral and factual considerations or if they could (and should) be aware of them. This is then further dependent on the evidence available to the agent, the opportunities the agent has to (adequately) process the information, and their own cognitive capacities.

This view does account for the cases where people seem to be blameworthy for unwitting omissions despite the fact that these omissions and attendant failures of awareness are not explainable in terms of ill intentions, and that blameworthiness cannot be traced back to a previous failure to execute some duty (Rudy-Hiller, 2022). These cases have also been called ‘forgetting cases’, where for example a dog was left inside a hot car during several hours. So long as the action (in this case, leaving the dog in the car) was not done with the intent to do wrong towards the dog or others who love the dog, it is not considered as blameworthy. Indeed, other unexpected commitments caused the dog to be forgotten in the first place. Yet because the agent does have a duty towards the dog that they fail to execute, they can be held responsible for the action. In this case, neither awareness nor ill will are necessary to state that someone can be considered culpable.

Yet this raises three questions for capacitarianists to answer: firstly, what kind of norms support the claim that an unwitting agent who has done wrong should have known better; secondly, what capacities justify the assumption that an unwitting agent could have done better; and thirdly, whether it is true that an unwitting agent should and could have done better are a sufficient basis for assigning responsibility to this agent (Rudy-Hiller, 2022).

The first question for capacitarianists is about the norms that are used to evaluate failures of awareness. When capacitarians speak of a failure of awareness, this does not mean that the agents involved should have known better and that they are automatically excused from their (unknowing) wrongdoing (Rudy-Hiller, 2017). Norms of awareness are dependent on the agent’s cognitive capacities and circumstances. Because awareness of certain considerations is necessary for fulfilling moral obligations, such an awareness can also be demanded of agents (Sher, 2009). Important to note here is that the norms of awareness, that are relied upon to support allegations that agents should have known, are not the same as duties of inquiry, as these concern actions rather than a state of mind (which is what awareness is). If failures of awareness are then outside of the agent's control – as becoming aware is then not something that we *do* – how can something as unvoluntary be seen as the basis for attributing responsibility? There are different responses to this issue; which range from letting go of the control condition of moral responsibility (Sher, 2009) to denying that norms of awareness give rise to moral obligations (Clarke, 2014). Rudy-Hiller (2017) makes use of a variant where the agents do have responsibility-relevant control over their states of awareness, and thus then can be morally obligated to remember or notice morally significant considerations.

The second question concerns the question of capacities; which of them make it true that an unwitting wrongdoer could have known better? Capacitarians make use of the idea of unexercised capacities to explain why certain unwitting agents could have been aware of the relevant considerations; they do have the cognitive capacities, yet have not exercised them despite the fact that there were no barriers to doing so. This means that they *could* have known better. There are several objections against this point, however. Firstly, the cognitive capacities that can issue intentional actions do not give the agent direct control over the awareness of the relevant considerations, but only make these considerations occurrent (Rosen, 2004). If control over the awareness of the relevant considerations is required for moral responsibility, then it remains unclear how failures of awareness can contribute to direct blameworthiness for unwitting wrongdoing. Secondly, capacitarians claim that the possession of these (cognitive) capacities grounds the expectations that are necessary to attribute moral responsibility to an agent. This can also be disputed, as the exercise of some of these capacities are not under our own control and whether they are used or not is dependent on chance (Clarke, 2017).

The third question centres around the 'should-and-could-have-known-better' clause, and whether this contributes to explaining blameworthiness for unwitting wrongdoing, or if something else is needed to ascribe moral responsibility (Sher, 2009). Since attributions of moral responsibility are usually based on some morally relevant feature of the agent to which awareness contributes – for example good will, having good intentions, or choice – accounting for unwitting wrongdoing can appear to be random (King, 2009). This also seems to rely on the fact that one has unexercised epistemic powers that lead to the wrongdoing. In fact, if someone does remember to do something once – for example ensuring that a dog is not left in a car on a hot day – and forgets to do so the next day, then it can be said that they did exercise these epistemic capacities the first time, and the second time they should and could have known better. Sher's (2009) response to this question focuses on *origination*, which he understands as an appropriate causal relationship between the agent and the 'wrongmaking' features of the act. In order to then blame or praise an agent for their actions, they need to be aware of the relevant moral features of the action and choose to act because or despite of these actions. Sher (2009) also adds that for wrong actions, this connection can also occur when the agent's unawareness of the action's wrong-making features is caused by the agent's disposition or character traits. If this is indeed the case and the agent also satisfies the other conditions of moral responsibility, then the agent can be held morally responsible. To go back to the example of the dog in the hot car; if the person forgetting the dog did so because of other pressing concerns, the failure to make sure the dog was comfortable does make this person blameworthy, because it is causally connected to this person. Origination then grounds moral responsibility in both known and unknown wrongdoing. It remains important, however, to show that the purely causal origination relation is a morally plausible basis for blaming someone.

Other authors do not agree with this view, and have even called it problematic. Levy (2014) for example states that while origination may be a condition on moral responsibility, moral responsibility cannot be based on origination alone. A failure of awareness does not automatically mean that a person can be blamed for an action's wrongness, because this does not say anything about this persons quality of will. Smith (2011) argues that one can only be held responsible in this case if there is a rational connection to the agents (moral) judgements, as only a causal connection would not be enough to hold someone morally responsible.

4.3.4. An overview

In the sections above, I have given an overview of the different positions on the epistemic condition of moral responsibility. Although one might have the intuition that being morally responsible for an action or decision requires awareness of certain things, there are several different positions on how the kind and content of this awareness is characterized. In the following section, I will focus on automated decision-making systems through the lens of the epistemic condition of moral responsibility.

5. The epistemic condition of moral responsibility and automated decision-making systems

In this section, I will be going through the different aspects of the epistemic condition of moral responsibility specifically related to automated decision-making systems, and explain whether or not this condition is met. As I mentioned earlier, the introduction of new technologies has made it more difficult to attribute moral responsibility, especially in the public sector where questions concerning (for example) financial aid need to be decided upon, and it is important to discuss the epistemic condition in relation to automated decision-making systems in detail. I will both look at the civil servant making the decision and the applicant. After all, the applicant is also entitled to the information grounding a decision, as this gives them the opportunity to firstly know and understand where a decision comes from, secondly allows them to contest the decision if they so wish, and thirdly gives them the opportunity to change aspects of their life if they found out this conflicted with the decision.

5.1. Content of awareness

First, I will look at questions regarding the content of the awareness – in other words, how the outcome of the system is presented, and how the civil servants understand the system they are working with. This part relies on general information on automated decision-making systems based on machine learning, as specific systems have as of yet not been made public. For the content of the awareness, it is important to first look at the different explanations I've outlined in the section explanations and see to which degree they might satisfy the epistemic condition.

Since there are currently no real options to make these machine learning systems transparent, ex post explanations are currently the best options available to better understand the system. These ex post explanations explain why a result is inferred, while looking at the system from the user's point of view (Xu et al., 2019). Each of these ex post explanations has advantages, especially when considering that they can help improve the understanding of the system, yet all of them also have several disadvantages which make it difficult to say that these do fulfil (an aspect of) the epistemic condition of moral responsibility.

Visual explanations make it easier to communicate the functioning of a model to an audience without a specific technical background (Belle & Papatonis, 2021), which includes civil servants working with the system without any kind of programming background. Most of these visual explanations are easy to interpret, and adding such an explanation to a certain system does not require a lot of work. Yet there is an upper boundary as to how many features can be added to such a visualisation, and how many different factors and features human beings are able to understand at the same time. Additionally, humans need to inspect the resulting visualisations to produce explanations – the system itself does not do so. This type of explanation therefore focuses more on how the model functions, than how the model got to a certain outcome. This means that while it makes it easier for civil servants to gain a general understanding, they are not able to fully trace back how the system got to a certain score and what parameters were used, merely because the system is rather too large and complex to be caught within this visual explanation (Belle & Papatonis, 2021).

Local explanations are capable of going into more detail than visual explanations, if only because they focus on a small part of the model – a local area of interest – rather than the model as a whole (Belle & Papatonis, 2021). This kind of explanation operates on instance-level explanations; which consists of a set of features that are considered to be the most responsible for the prediction of an small part of the model. In other words, this explanation focuses on the smallest set of features which have to be changed in the binary vector in order to alter the predicted label (Tamagnini et al., 2017). As mentioned, this type of model does not generalise on a larger scale, which means that it is only useful when considering a small set of parameters. Small changes in these parameters might result in very different explanations, and very different outcomes. Additionally, it is very difficult to define locality – what parameters exactly are considered, and to what degree should connected parameters be taken into account as well? For smaller parts of the machine learning application this can be useful, but it requires a profound understanding of the model and how a small change can affect the entire functioning of the model – which is not knowledge currently required of civil servants.

Explanations by example can provide insights about the model's internal functioning, as through the use of an example it would be possible to trace back a couple of steps (Belle & Papatonis, 2021). These kinds of explanations can uncover the most influential training datapoints, which have led to the predictions the model makes. In other words, this can give the user of the model insight into the different parameters and the numerical value attached to them. This is often done with the use of the training data, which can make the different established patterns and connections visible. These kinds of explanation do require a human inspection, and they do not explicitly state what parts of the example data influence how the model functions. Explanations by example can therefore give insight into the models functioning, but often do not go into specifics. While this, like the other examples, can be very helpful for the overall understanding of the model, it does not help in specific situations where the applicant demands to know how the decision was made.

Explanations by simplification focus on approximating the whole of the model while using (a mixture of) simpler models (Belle & Papatonis, 2021). The explanations – such as decision rules for example –

are easier to understand for those working with the model. The choice to use a complex model is usually based on the fact that a simple model is not capable of functioning on the same level, or because it does not have the same features. The approximation by simpler models can therefore often not capture the entire functioning of the model, and these surrogate models often come with their own limitations as well. Again this type of explanation can be used to further the understanding of the system one works with, but leaves much to be desired when relaying this information to the applicant.

Feature relevance explanations, like local explanations, operate on an instance level where the importance of each of the features and parameters is calculated (Belle & Papatonis, 2021). This can give insight into how the different data inputs are weighed, and provides an opportunity to look into the parameters which have been given a disproportionate weight. However, in cases where the different features or parameters are heavily correlated, these types of explanations are highly sensitive. In many cases the exact weights of the parameters are approximated, meaning that these do not correspond one hundred percent with what happens in the model. The (incorrect) ordering of these weights then impacts the outcome. So long as the different parameters are then not correlated, this can indicate to civil servants and applicants alike what factors are deemed more important than others by the model. If these parameters are correlated and it is difficult to say what their importance truly is, both groups would have to be cautious before taking this as sufficient to satisfy epistemic claims.

While these different kinds of explanations therefore seem to add to the content of the awareness of the action – as this in more detail describes how the model works, and through which connections between data points become more clear – they remain approximations of how the system actually works. The awareness of the action seems pretty straightforward. If person A pushes a button and knows what the button is for, then they have an awareness of the action, and can be held responsible for it. Alternatively, if person A does not know what the purpose of the button is, then the issue of attributing responsibility is more complicated. Yet many philosophers deny that one needs to know in detail what a person is doing in order to be held responsible for it. In relation to decision-making with machine learning systems in the public sector, it is important to look at what it is that the civil servant is doing and whether they are aware of the content of their actions. Here the distinction between an automated decision-making system and a decision-support system is important to note. In an automated decision-making machine, the score from the system is fed into a decision rule and human beings often only have an input when checking whether this has been applied accurately. With decision support systems, the output from the system is merely one of the factors which are used to determine whether or not someone – in this specific case – is eligible for financial aid. The civil servant is aware that they are deciding on whether or not to grant financial aid, based on a number of factors. For fully automated systems, understanding where the score is coming from is crucial to understand the final result. For decision support systems, as long as the other factors are taken into account as well (such as information handed in by the applicant, historical information on income, etc.) and are combined with the result from the system, the civil servant is capable of weighing them against one another. Yet in most cases, the result or recommendation from the system is taken as more reliable and more accurate (Marcus & Davis, 2019), which means that this can be given disproportionate weight by the person in charge of the final decision. If this is the case, then understanding where the recommendation of the system is based on is essential as well.

The awareness of the moral significance of the action is the second question to answer with regard to the content of the awareness and algorithm supported decision-making, and this again lies more with the civil servant than with the system itself. The human being is capable of interpreting and adding context to the (recommended) decision, for example that a certain amount of money might be sufficient legally to survive on, but whether the agent beliefs this is another matter. Important to note here is the distinction between a *de dicto* awareness that an action is morally wrong, and a *de re* awareness. Leaving aside the question whether or not the government should be deciding on financial aid, there can be aspects which can be considered morally wrong – such as the fact to not label someone as eligible, or to have done so on morally wrong reasons. This is where the ambiguity starts, for how can one know what the decision is based on exactly, if they are not aware of how the machine learning system came to its decision?

Similar to the question with regard to the moral significance of the action, the awareness of possible consequences of the action is very much dependent on whether or not the civil servant (sufficiently) understands the system they are working with. If a civil servant does not know what factors the decision is based on, the consequences are more difficult to predict, if not impossible. Through the use of explanations by example, for example, it would be possible to get an approximation of the consequences. Zimmerman (1997) for example seems to state that if the agent has taken into account some of the possible consequences, one can already be held responsible (so long as all the other specific conditions are also met). As predicting the future is impossible, this is also not a hard requirement for those working with automated decision-making systems.

The awareness of alternative options – that is, in this case, alternative options to using a machine learning system – might have been there, as the use of machine learning in the public sector is still relatively recent. Collecting and going through all the information by hand would be an option, though perhaps not the most (cost) effective. Yet with the call for more smart and integrated solutions for citizens – meaning a decision fully tailored to their personal, individual circumstances – governments often feel the pressure to start working with these systems (Diakopoulos, 2016; Frissen, 2023). Working with the system was therefore a considered a necessity, and this brings us to the different attempts that have been made to make the system intelligible. Of course, it could be argued that all civil servants working with the system would need to have a detailed understanding of the system, but as this is not a rule-based decision aid where the mathematical formula is established in advance, transparency is not as easy to attain in this case. Understanding the different steps to get to such an outcome – as I have outlined earlier – would already help, as this would also give an indication to those working with the system what exactly the outcomes mean. As Babushkina & Votsis (2022) state in their article, what these recommendations or outputs from automated decision-making systems indicate is focused on a similarity between the training data and the data entered. Even when these civil servants do not know what datasets were used to train the model, this might already help in better understanding what a decision is based on.

With regard to the content of the awareness, the epistemic condition of moral responsibility seems to undermined, especially when taking into account the accountability measures in place for government decisions. This also means that the explanations offered by the government to the applicants would not be sufficient for the applicants to satisfy their own epistemic condition – without this knowledge, they do not have the necessary means to make decisions about their own life. For an overview of the different automated decision-making systems and decision support systems, see table 1 below.

Table 1: Automated decision-making systems/Decision support systems and the content of awareness

Conditions of moral responsibility *Type of ADM system*	*Awareness of action*	*Awareness of the moral significance*	*Awareness of the consequences*	*Awareness of alternative actions*	*Content of awareness*
Rule-based decision-support systems	●●●●●	●●●●●	●●●●●	●●●●●	●●●●●
Rule-based automated decision-making systems	●●●○○	●●●○○	●●●○○	●●●●●	●●●○○
Machine learning decision-support systems	●●●●●	●●○○○	●●○○○	●●●●●	●●○○○
Machine learning automated decision-making systems	●○○○○	○○○○○	○○○○○	●●●●●	●○○○○

For this table I assume an ideal situation, namely a civil servant who has the necessary knowledge of these decision aid systems and (if there are several factors to be weighed) gives all factors appropriate weight. For some categories – for example such as the awareness of the consequences – there can be huge differences between different decisions, and to what degree someone is aware of the consequences an individual decision can have on the applicant. In this case as well, if a civil servant *can* know what the consequences of the action for an individual would be, it is assumed that they in fact *do* know.

5.2. Kinds of awareness

The questions with regard to the kinds of awareness focus on the mental state of the person making the decisions, and asks what kind of awareness they should possess in order to be aware of their actions. If a person is *not* aware of what they are doing or had a false belief about what the action they were performing meant, then the requirement of awareness has not been satisfied. Other requirements on the moral significance or possible consequences are then also not able to be fulfilled. This ignorance or false belief would mean that the agent is not considered culpable or blameworthy for their actions, simply because they did not know. There is one caveat to this – the ignorance itself should not be culpable. If the civil servant working with the system does not understand the system, this should follow from the fact that the system is truly black boxed or that they do not have access to this system whatsoever, in order to not be culpable.

According to the volitionist argument, the question whether or not someone is culpable for their ignorance depends on the same requirements as for moral responsibility in general (Zimmerman, 1997). Additionally, this counts for both factual ignorance – not knowing how the automated decision-making system worked exactly – and moral ignorance – not knowing what the moral significance was of (the decision to) use the system.

Whether or not the epistemic condition of moral responsibility would be fulfilled for the civil servants would then depend on whether or not they are culpable for their ignorance. I argue here that this is not the case. After all, the whole set-up of the automated decision-making systems based on machine learning is that they find patterns and connections that human beings might not, and that they are capable of analysing more data in a shorter period of time than humans can. Ignorance then seems to be built into the systems used to deny or approve applicants, and apart from understanding the general steps by which such a system is built and deployed (and of course through the use of ex post examples) there are limited options for the civil servants to dive into the system and figure out how this works in detail.

If the civil servants have difficulty understanding the system and the policy documents which should have more details on the decision making process, to such a degree that it is doubtful whether or not they satisfy the epistemic condition, then for applicants this situation might be even worse. Policy documents are often said to be incomprehensible for those not working with the jargon (Frissen, 2023), and applicants often do not even get access to the system itself; they have to rely on the communication from the government to hear what the final decision is. Oftentimes, they might not even know that their case was subject to a decision support system or automated decision-making system. They are then even less able to satisfy the epistemic condition for decisions they make about their own lives – for which the explanation about the decision is crucial. Decisions about financial aid, educational opportunities, etc. can have a life-changing impact.

5.3. Legal mechanisms to overcome the lack of understanding

The question arises what could be done to remedy the situation, knowing that the epistemic condition is not met by the civil servants working with the system, and thus also not by the applicants who later have to use the outcome (and the basis for this outcome) to decide on matters in their personal lives. The first solution would take away worries that what happens in the system is not accessible to those working with the system or those tasked with supervising these processes, by focusing only on aids for decision-making which are rule-based such as decision trees. The advantage of rule-based systems lies in the fact that the mathematical formula or algorithm to make the decision or recommendation are known in advance, as are the different parameters necessary for the decision and the weight that these parameters get. Yet as I have mentioned earlier, making decisions with only rule-based systems would require a lot more manpower and time, which in turn can cause difficulties. Additionally, the call for more sophisticated systems to help out with (relatively) routine decision-making systems (Ivanov, 2022; Diakopoulos, 2016) means that it could be seen as a step backwards to go back to working with simpler systems.

There are a couple of legal mechanisms currently under discussion which might help with the issue, such as the Right to an Explanation, the Right to a Justification, and the Right to Contest. Each of these has difficulties with the execution and some legal scholars have argued that these mechanisms would not solve the issues in the first place. Yet as some of them have been (implicitly) introduced in recent (inter)national legislation such as the General Data Protection Regulation, I will briefly go over them and present the opportunities and challenges. All of these legal mechanisms rely on the assumption that it would be possible to create systems which are transparent, traceable, and/or explainable, which is the first difficulty found with each of the systems. To make sure that citizens are not subject to automated decision-making systems alone, the GDPR has also established that decisions made by such systems always have to be checked by humans, before being finalised, in order to keep an extra mechanism of control in the loop.

5.3.1. The right to an explanation

The right to an explanation refers to the idea (and in certain cases, legal mechanisms) that individuals subject to automated decision-making or profiling algorithms should have the right to receive an explanation of specific automated decisions that have been made about them (Wachter et al., 2017). The specific type of explanation that should be offered is under debate, as is the fact whether or not this is actually included in the GDPR. Kaminski (2019) states that the right to an explanation is a fundamental aspect of algorithmic accountability, which has the aim to ensure that individuals have transparency and insight into the decisions, and that they are empowered to understand and challenge the decisions made by the algorithms. This, in turn, promotes fairness, accountability, and responsibility.

Wachter et al. (2017) state that this right to an explanation cannot be found within the GDPR, and have several reasons as to why this is not made explicit. There are three possible legal bases on which the right to an explanation can be based: the right not to be subject to automated decision-making and safeguards to ensure that this is indeed the case, notification duties of the data controllers, and the right to access. Wachter et al. (2017) therefore focus on the right to be informed, rather than the right to an explanation, where the subjects to the automated decision-making systems are provided with meaningful information about the significance and the consequences of the automated decision-making process. This seems to strike a balance between transparency, and protecting intellectual property.

If it is the case that the right to an explanation is included in the GDPR, then it provides citizens of the EU the right to demand a "meaningful explanation about how their automated decision making and/or profiling systems reach final decisions on involved data subjects" (Kim & Routledge, 2022). One of the main questions to answer here, is what kind of explanation is seen as to satisfy this right to an explanation. In general, there are two different kinds of explanations that can be used within the context of algorithmic decisions: an ex ante, generic explanation, or an ex post explanation about a specific decision.

The right to an ex ante explanation seems to be the equivalent to an already widely accepted right, namely the right to be informed, or the right to informed consent. Indeed, within the Dutch context, this can also be said to be satisfied by the principles of legal certainty and legitimate expectations – citizens can count on the government to act in a consistent manner, and the government has certain rights and duties when it comes to decisions made involving citizens (Jaspers, 2019; Rijkswaterstaat, 2019). In this sense, the GDPR does not add to the obligations of the government to make the reasoning behind certain decisions transparent as they are already legally obliged to do so. Yet it is the ex post explanation that does present a new kind of right to an explanation, as this requires the decision-maker to explain the reasoning behind a specific decision, and explain the process of decision-making (Kim & Routledge, 2022). While it can be debatable whether private organisations and companies are able to satisfy this interpretation of a right to an explanation, the Dutch government is already legally obliged to do so by one of the General Principles of Good Governance, namely the motivation principle (Van Goud, 2016).

The provisions of the GDPR mostly focus on regulations for businesses and governments to act in certain ways before collecting, storing, and making use of personal data (Kim & Routledge, 2022). Initially, therefore, the GDPR seems to focus on the ex ante explanation aimed at a generic explanation of the system's functions, which has also been equated to a right to informed consent (Wachter et al, 2017). Not all scholars agree with this rather narrow interpretation, however, as for example Selbst and Powles (2017) argue with their interpretation of the GDPR. They state that there is a right to an ex post explanation embedded within the GDPR, and that this should also be specific to the particular situation, rather than an approximation or a generic explanation.

Yet with the current possibilities for explaining machine learning systems and outcomes, looking at more specific situations is not yet possible. Implementing the right to an explanation (with specific, ex post explanations) faces several challenges and limitations; one of which concerns the complexity of the majority of automated decision-making systems (Belle & Papatonis, 2021; Kaminski, 2019). As introduced earlier, within opaque systems, it is not possible to get to a specific ex post explanation. Additionally, there is a risk that an explanation can be used to legitimise or justify other decisions, as this can be seen as an example of how the system works (Kaminski, 2019).

5.3.2. The right to a justification

In the current debate on artificial intelligence, machine learning, and explanations, there are some who state that it is not in fact the explanation (as an extension of transparency) that should be looked for, but rather that a justification of a decision is the important aspect of being able to attribute responsibility to someone (Malgieri, 2021). Explanations are said to valuable because it is important for human beings to understand the systems they are using and are subject to, especially when individuals want to challenge certain decisions or identify if and where there were biases present within the system (Gillis & Simons, 2019). Yet Gillis & Simons (2019) state that explanations in themselves are only valuable if they are a means to provide a justification of the broader decision-making procedure: "What matters is justifying why the rules are the way they are; explaining what the rules are must further this end" (page 76). Furthermore, they state that the focus on technical explanations only matter for individuals, and that this is a sign of an too strong focus on transparency. Indeed, while transparency is a necessary part of offering an explanation or justification, Gillis and Simons (2019) argue that transparency in itself is not a goal, but rather a means to achieve a goal.

While questioning the rules and determining whether or not they still (should) apply is a worthwhile exercise, the problem lies in the fact that in automated decision-making or decision support systems the rules themselves are often unclear (Malgieri, 2021). If the machine learning algorithm is the one establishing patterns and finding connections between different datapoints, then the rules themselves have become opaque, even though the system has been constructed to help out with certain decision-making tasks. Explanations, then, especially on an individual level, are necessary to find out what exactly the outcome of a system is based on and how that is used in human decision-making (Giovanola & Tiribelli, 2022). The justification question Gillis and Simons (2019) are focusing on is more a question of accountability. They state that the justification is not only meant to explain the rules, but also why the rules were applicable to a certain situation in the first place. Attributing responsibility to someone and, as a step further, hold them responsible for the decision is something different than wanting an explanation as an epistemic basis for making further decisions. A justification without explanation of the system and the patterns and connections that were used to get to a certain decision or recommendation, would therefore not be sufficient to hold someone (morally) responsible for the decision.

5.3.3. The right to contest

Being able to contest decisions made by automated decision-making[10] systems is a third legal mechanism introduced to make automated decision-making and decision support systems more accountable. Contestability is required by law within the European Union when it comes to automated decision-making systems and profiling systems (art. 22(3) GDPR, European Union, 2016). Several scholars have written guidelines on how to create contestable systems – often, there needs to be an option to contest a system written into the code – and have argued that contestability should make it easier to intervene in an automated decision-making system (Almada, 2019; Lyons et al., 2021). Others, such as Kluttz et al. (2019) go even further and state that contestability in itself should be the main focus of programmers and policy makers to ensure accountable automated decision-making systems.

Kluttz et al. (2019) start their argument with determining what exactly it is that is being made transparent, when the discussion focuses on making both automated decision-making systems and decision-making systems in general more accountable. They state that this concerns three different aspects, which are protected in privacy laws and laws directed at consumer protection. The first aspect focuses on information about individuals which is used by these systems to make the decisions or get to a recommendation. After all, citizens have the right to request the information a public institution has on them. The second aspect Kluttz et al. (2019) focus on is the existence of these algorithms, and the scope and purpose to which they are used within public organisations – though not the algorithms themselves. The third aspect focuses on the output of these algorithms and the

[10] Interestingly, this seems to apply more to the private sector than the public sector, as all public decisions need to be contestable anyways (Rijkswaterstaat, 2019).

decision rules within certain processes. This also means that most algorithms are not made public. The issue Kluttz et al., (2019) have with the focus on solely explainability in the debate on the accountable use of automated decision-making systems is that most of the correlational patterns found by these systems are taken to be causal by the people working with these systems, even though they are not.

Kluttz et al. (2019) therefore propose to focus on contestability rather than explainability or transparency – they state that being able to challenge the predictions of the system is more useful. This should be built into the system as a standard design feature (see for example the guidelines Lyons et al. (2021) have written for including contestability) and should be consistent with national and international privacy laws and consumer protection regulations. This would decrease the dependency on automated decision-making systems when these systems are not working as they were intended to, and could limit the amount of 'bias' built or found in the patterns in the data. Contestability would also increase the legal options people have available to them, were they to be subject to automated decision-making processes, more than explainability or transparency would according to Kluttz et al. (2019). Wachter et al. (2017) share this view, and state that the GDPR does not offer a 'right to an explanation' as it has been explained by several legal scholars, but rather a right to information or a right to be informed.

Important in Kluttz et al.'s (2019) argument on why transparency would not be sufficient to create responsible automated decision-making systems is the three aspects on transparency they focus on. As I briefly mentioned above, they focus mostly on individual information in the system, the existence and scope of the system, and outputs of the system. While this does allow citizens to request information about themselves, it does not mean that they get an insight into how the system works. Transparency therefore remains important when talking about accountability and automated decision-making – indeed, contestability itself relies in part on transparency, traceability, and explainability. Combining contestability with these three concepts, especially when considering design requirements for automated decision-making systems, would increase accountable decision-making. In order to contest a decision, one needs to know what exactly there is to contest.

6. Knowing and explaining

Using automated decision-making systems and decision support systems within the public sector means that the civil servants working with these systems are unable – no matter how much effort they put into remedying their ignorance – to fully understand the system and the way it functions. While the main factors of the decision-making process have been set out in regulations and policies, any additional information used by the system is not on the radar of those who work with the system. Additionally, as ex post explanations are mostly generic explanations (meaning that they provide examples, but do not go into specific cases) it is also not possible for the civil servant to satisfy the motivation principle of the Dutch General Principles of Good Governance, or indeed the idea behind the Right to an Explanation or the Right to a Justification. Introducing these concepts would therefore also allow the civil servant working with these systems to gain a better understanding of what exactly they are working with, and what possible consequences using the system can have for individual citizens.

Additionally, ensuring that the civil servants have access to the exact information as to how a decision or a recommendation are established means that they can be considered as morally responsible for the decisions made with the aid of these systems. As mentioned earlier, with the introduction of the GDPR citizens of the EU have the right not to be subject to automated decision-making systems, which means that a decision-support system can be used, while full automation is not allowed. A human being is therefore always in the loop, and can therefore be considered as morally responsible for the decision as they have the final say. They have the powers and capacities, have a causal relationship to the outcome of the decision and therefore the consequences for the applicant, and can be doing wrong, in the case of making a decision when they should not or vice versa. With the right to an explanation specifically, it is guaranteed that they can have access to the information (on both a general basis how the system works, and a specific and individual basis). The question to answer then is whether they should look into each specific instance, and whether this is practically possible.

In an ideal world, the answer to the question whether or not civil servants should be aware of each decision, its (possible) consequences, and moral significance would be an unequivocal 'yes'. Especially when talking about moral responsibility and considering the fact that these decisions can profoundly affect people's lives, it should be considered a necessity that the civil servants have the knowledge necessary to make these decisions responsibly. Yet with the current systems used for decision-making, and the fact that legal protections such as the right to an explanation, the right to a justification, and the right to contest cannot yet be used optimally to challenge this situation, it seems prudent to reassess the use of these systems, and determine in which situations they can be used responsibly.

The fact that the civil servant is not capable of explaining a certain decision also has consequences for the applicant, or moral patient, of the decision. After all, they are moral agents in their own right, and often need the information the decision is based upon to consider what options are open to them, and what would be the correct decision for them to take. Kim and Routledge (2022) give the example of informed consent, and argue that in the case of informed consent, the applicant or moral patient is aware of what they are consenting to, rather than merely clicking a button to access a certain website. When a person consents to another person's or organisation's actions, the first person often only consents when they have some kind of explanation about what the other person's actions will actually mean, leaving aside what kind of explanation is offered. In this case, it will often be an ex ante explanation, as the applicant has to supply the information before it can be used within the algorithmic system. Here, the right to be informed or the right to an explanation overlaps with the right to informed consent, as the explanation is given in advance – users of the website or system know in advance what kind of personal information they should supply or is collected. The consent is only meaningful so long as the consent is informed, as it is difficult to know what one is consenting to otherwise. This is similar to the right to contest, yet here the information is necessary to be disclosed afterwards, as it is difficult to contest a decision if it is not known what it is based on.

The right to a justification would also not solve the issue, as the justification is offered by the civil servant to the applicant is in most cases based on ex post explanations or approximations rather than an actual explanation. Apart from the fact that the civil servant then justifies something they do not actually know or are aware of, the justification offered to the applicant would then also not be sufficient grounds for them to base decisions on or take action on. In other words, the motivation principle is not satisfied, and the epistemic condition undermined. This then also begs the question if an applicant can be held morally responsible for decisions based on this information, as it is not possible for the applicant to gain access to the systems behind the decision nor the information used by these systems. In other words, they are not aware whether or not the knowledge is correct, and are not able to verify this information. The right to an explanation, i.e. a specific explanation, focused on the particular situation and specific individual, is therefore the only way to ensure that the moral patient also can become a morally responsible agent.

With regards to the principle of legal certainty, there are also several questions. As the applicant is not fully aware of what type of data is used by the automated decision-making system – after all, more data is collected than what is actually necessary for many decisions made in the public sector –

then it is also difficult to find out whether or not the system focused only on those aspects defined in the regulation (as per the principle of legal certainty) or if there were other factors involved in the decision. It can be the case only those datapoints are entered which are necessary for the decision and that a civil servant following a simple decision tree would come to the same conclusion, but this does not have to be so and it is not easy to find out.

The motivation principle is then difficult to enact by the civil servant, who does not have the access to the information nor the processes in the system to provide a full explanation. As the applicant gets this information via the civil servant, they are not aware of the reasoning behind certain decisions, nor whether the decision-rules established by the government (and made public) have actually been followed.

7. Conclusion and discussion

Decision-making in the public sector has in some ways changed tremendously with the introduction of digital systems, and in other ways stayed the same. As these decisions have to be based on legislation or policies, citizens are able to know in advance what is expected of them, or what they are not allowed to do. Indeed, following the policies can be compared to following a decision-tree, as this is also a rule-based system for arriving at the appropriate decision for the appropriate situation. Frissen (2023) adds here that the expectations for public sector decision-making have only grown, as with the inclusion of these digital systems, citizens have started to expect 'smart' decisions, tailored to their own personal situations. There are however several additional requirements for public sector decision-making, which in the Dutch context have been codified into the General Principles of Good Governance (Rijkswaterstaat, 2019). While all principles play an important role in ensuring that decision-making is accountable, there are two in particular that have to be taken into account when making decisions with automated decision-making systems or decision support systems: the motivation principle, and the principle of legal certainty. The motivation principle states that civil servants (or the government in general) needs to be able to explain the reasoning behind the decisions. The principle of legal certainty states that citizens (and thus civil servants as well) need to be aware of their rights and duties, and that they can expect that the laws and regulations concerning those rights and duties will be honoured.

As I mentioned above, policies and regulations can be compared to decision-trees in the way that both of them have a clear path to follow, to get to the decision. While there are also several decision-making aids which are more complex, in general these rule-based systems are relatively easy to follow. The issue arises when the systems get more complex – opaque, to use the terminology from Belle and Papatonis (2021) – and when it is no longer possible to trace the decision back to its starting point. Indeed, this is where issues with transparency, traceability, and explainability arrive. All these are necessary to satisfy the motivation principle I mentioned above, as well as the principle of legal certainty. The explanations that one than can get from these systems are not the explanations one can get with a rule-based system, as the closest one can get for a particular situation is an approximation of the decision-making process, after which the decision can be justified (though not fully explained). These are called ex post explanations, and are generic explanations that would not work for satisfying the motivation principle – citizens have the right to know what exactly happened in their particular case.

Leaving aside the questions of legal responsibility or causal responsibility, the issue of moral responsibility remains interesting for those working with these systems. In most cases, civil servants have the powers and capacities to work with them, are causally related to the decision (and outcome of the decision), and are capable of wrongdoing (or rightdoing). It is the epistemic condition that is causing problems here, as it is unclear what kind of awareness civil servants have of the decision they are making, if the information for this decision comes from an automated decision-making system or decision support system. This has been further divided into the awareness of the action, the awareness of the moral significance, the awareness of the consequences, and the awareness of alternative actions or options.

In most cases, the different kind of ex post explanations that we do have for these systems seem to add to the content of the awareness of the action, as the civil servants are more aware of how the system works even if they are unable to trace back the different steps. Yet it is precisely this issue of specific situations where the use of these systems falls short, and where ex post explanations are not capable of remedying this. There are several (conceptualisations of) legal mechanisms which might help in these cases, such as the right to an explanation, the right to a justification, and the right to contest. As I have outlined above, due to difficulties in realising the right to an explanation, it is currently not possible for civil servants to go into detail on specific situations or individual circumstances if the decision was made with machine learning decision aid. The right to a justification, which can be very helpful for determining which parameters were considered to be more important than others, does not fill this void as a justification is always backwards looking, and does not focus on the process as it happened. The right to contest, already present in most cases of public sector decision-making, also does not solve the issue entirely as one needs to know what it is exactly that one is contesting.

This then leads to several difficulties: because it is not possible to get a particular explanation from the system, civil servants do not know what the output from the system is based on and are unable to satisfy the epistemic condition of moral responsibility. This means that they are also unable to satisfy the motivation principle – meaning that they are unable to explain the complete reasoning behind the decision. This in turn means that the applicant whom the decision was about is not aware of the particulars, and cannot use this information to base their own decisions on, nor use this information to contest the situation in case something in the process has gone wrong (which was for example the case in the Dutch childcare benefits scandal). Furthermore, because it is not possible for the citizen to dive deeper into the decision and determine on what factors this decision was made, the principle of legal certainty can also not be satisfied. After all, whether or not the decision was made on the information outlined in the policies is unclear, as the machine learning application could have made different connections, and could have given those priority.

As an answer to the research question asked at the beginning of this thesis, I would therefore argue that the use of automated decision-making systems does indeed undermine the epistemic condition of moral responsibility, and that the current conceptualisations of legal mechanisms would only remedy this if the right to an explanation would be understood as a specific ex post explanation. While the right to an explanation would solve the issue as both the civil servant and the citizen is aware of the input and the process through which this input goes, it is currently not possible to realise this on an individual basis. Using automated decision-making systems and decision-support systems within the public sector thus needs additional requirements, and legal protections for citizens, before using these systems can be considered as morally responsible.

Of course, it is important to take into account that there are different types of automated decision-making systems – even those that can be considered as transparent by some can still cause confusion by those having to use it – and that the developments in the field of machine learning with the aim to facilitate decision-making have not yet finished. Introducing workshops, explanatory documents for the systems, or even courses on how to work with automated decision-making systems on a general level would already remedy part of the ignorance on the part of the civil servant. Indeed, there might even be a case for introducing mandatory testing for civil servants, in order to ensure that they have a basic understanding of the system they are working with, and what exactly the outcome of the system represents.

Further research could focus on new developments in the field of decision aids and automated decision-making systems, and additionally look for the use of these systems outside of the public sector. While the GDPR does impact how these systems are used in the private sector, additional requirements might be necessary to ensure morally responsible decision-making within specifically this sector. Another possibility for further research lies in the fact that there are other conceptualisations of moral responsibility, which can focus on different implications of the use of automated decision-making systems. Indeed, looking at the different power relationships between the civil servant and the applicant would also add to the discussion, as this would further highlight the need for legal protections, and a need for a more equal knowledge base.

Another direction for further research could focus on questions of distributed moral responsibility. While I have not focused upon the problem of many hands with regards to responsibility or accountability in this thesis, and have made the assumption that decisions are made by individual civil servants, this does not take away that in practice, there is often a group of people working on a decision or several aspects of a decision. After all, one department might be dealing with the information to add, another working on processing, and another working on the design of the automated decision-making system itself. All of them could be implicated in how the system gets to a certain recommendation, and thus they can be considered responsible (to a certain degree) for the decision that was made. This would be an exciting question to continue this line of research.

Bibliography

Almada, M. (2019). Human intervention in automated decision-making: Toward the construction of contestable systems. *Proceedings of the Seventeenth International Conference on Artificial Intelligence and Law*, 2–11.

Amaya, S., & Doris, J. M. (2015). *No excuses: Performance mistakes in morality*.

Amnesty International. (n.d.). *Algoritmes, Big Data en de overheid*. Amnesty International. Retrieved 30 June 2023, from https://www.amnesty.nl/wat-we-doen/tech-en-mensenrechten/algoritmes-big-data-overheid

Amnesty International. (2021, October 25). *Xenofobe machines: Discriminatie door ongereguleerd gebruik van algoritmen in het Nederlandse toeslagenschandaal*. Amnesty International. https://www.amnesty.org/en/documents/eur35/4686/2021/nl/

Anaeko. (2019, August 3). The Application of Artificial Neural Networks in Government. Anaeko — Trusted Data Partner. https://medium.com/anaeko-trusted-data-partner/the-application-of-artificial-neural-networks-in-government-6a7fdeb475bd

Arrieta, A. B., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., García, S., Gil-López, S., Molina, D., & Benjamins, R. (2020). Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion*, *58*, 82–115.

Atkinson, K., Bench-Capon, T., & Bollegala, D. (2020). Explanation in AI and law: Past, present and future. *Artificial Intelligence*, *289*, 103387.

Babushkina, D., & Votsis, A. (2022). Epistemo-ethical constraints on AI-human decision making for diagnostic purposes. *Ethics and Information Technology*, *24*(2), 1–15.

Belle, V., & Papantonis, I. (2021). Principles and practice of explainable machine learning. *Frontiers in Big Data*, 39.

Brockmann, E. N., & Anthony, W. P. (2002). Tacit Knowledge and Strategic Decision Making. *Group &*

Organization Management, *27*(4), 436–455. https://doi.org/10.1177/1059601102238356

Burggraaf, C. M. J. (2021, December 1). *Het materiële zorgvuldigheidsbeginsel*. https://www.navigator.nl/document/id1df151b48693419bb2fc331da7394d45/toetsingsintensiteit-een-vergelijkende-studie-naar-het-varieren-van-de-toetsingsintensiteit-door-de-rechter-staat-en-recht-nr-54-2545-het-materiele-zorgvuldigheidsbeginsel

Chauvet, C. (2015). Obedience/Disobedience and Civil Servants. *Pouvoirs*, *155*(4), 149–160.

Clarke, R. (2017). Ignorance, Revision, and Commonsense. *Responsibility: The Epistemic Condition*.

Clarke, R. K. (2014). *Omissions: Agency, metaphysics, and responsibility*. Oxford University Press.

Coeckelbergh, M. (2020). Artificial intelligence, responsibility attribution, and a relational justification of explainability. *Science and Engineering Ethics*, *26*(4), 2051–2068.

Cutler, A., Cutler, D. R., & Stevens, J. R. (2012). Random forests. *Ensemble Machine Learning: Methods and Applications*, 157–175.

de Jonge, E. J. (2017, October 26). *Ministeriële verantwoordelijkheid*. Nederland Rechtsstaat. https://www.nederlandrechtsstaat.nl/grondwet/inleiding-hoofdstuk-2-regering/artikel-42-ministeriele-verantwoordelijkheid/

Dhiraj, K. (2019, June 13). Top 4 advantages and disadvantages of Support Vector Machine or SVM. *Medium*. https://dhirajkumarblog.medium.com/top-4-advantages-and-disadvantages-of-support-vector-machine-or-svm-a3c06a2b107

Diakopoulos, N. (2016). Accountability in algorithmic decision making. *Communications of the ACM*, *59*(2), 56–62. https://doi.org/10.1145/2844110

Douglas, D. M., Howard, D., & Lacey, J. (2021). Moral responsibility for computationally designed products. *AI and Ethics*, 1–9.

European Union. (2016). *General Data Protection Regulation*. GDPR.eu. https://gdpr.eu/tag/gdpr/

Engstad, N. (2017, September 29). Consistency of the case law as a prerequisite of legal certainty: European and National perspectives. *Portal*. https://www.coe.int/en/web/portal/news-

2017/-/asset_publisher/StEVosr24HJ2/content/conference-on-harmonisation-of-case-law-and-judicial-practice

Everitt, B. S., & Skrondal, A. (2010). *The Cambridge dictionary of statistics*.

FitzPatrick, W. J. (2008). Moral responsibility and normative ignorance: Answering a new skeptical challenge. *Ethics*, *118*(4), 589–613.

Frankfurt, H. G. (1969). Alternate Possibilities and Moral Responsibility. *The Journal of Philosophy*, *66*(23), 829–839. https://doi.org/10.2307/2023833

Fricker, M. (2010). The Relativism of Blame and Williams's Relativism of Distance. *Aristotelian Society Supplementary Volume*, *84*(1), 151–177.

Frissen, P. (2023). *De Integrale Staat—Kritiek van de Samenhang* (1st ed.). Boom Uitgevers Amsterdam.

Gillis, T. B., & Simons, J. (2019). Explanation < Justification: GDPR and the Perils of Privacy. *Journal of Law & Innovation (JLI)*, *2*, 71.

Giovanola, B., & Tiribelli, S. (2022). Weapons of moral construction? On the value of fairness in algorithmic decision-making. *Ethics and Information Technology*, *24*(1), 3. https://doi.org/10.1007/s10676-022-09622-5

Guerrero, A. A. (2007). Don't know, don't kill: Moral ignorance, culpability, and caution. *Philosophical Studies*, *136*, 59–97.

Gunnemyr, M., & Touborg, C. T. (2023). You just didn't care enough. *SOCIAL PHILOSOPHY*, 1.

Günther, M., & Kasirzadeh, A. (2022). Algorithmic and human decision making: For a double standard of transparency. *AI & SOCIETY*, 1–7.

Hamon, R., Junklewitz, H., Malgieri, G., Hert, P. D., Beslay, L., & Sanchez, I. (2021). Impossible Explanations? Beyond explainable AI in the GDPR from a COVID-19 use case scenario. *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 549–559.

Heilbron, B., Kootstra, A., & van Nuland, M. (2023, June 21). *Zorgen over fraudebestrijding Duo: Advocaten zien opvallend veel beschuldigde studenten met een migratie-achtergrond*. Trouw. https://www.trouw.nl/onderwijs/zorgen-over-fraudebestrijding-duo-advocaten-zien-opvallend-veel-beschuldigde-studenten-met-een-migratie-achtergrond~bf3d5cd6/

Hyde, D. (2013, September 23). *Who to blame when 'computer says no'*. https://www.telegraph.co.uk/finance/personalfinance/bank-accounts/10324271/Who-to-blame-when-computer-says-no.html

IBM. (2018, August 16). *About Linear Regression*. https://www.ibm.com/topics/linear-regression

IBM. (2022a, February 6). *What is the k-nearest neighbors algorithm?* https://www.ibm.com/topics/knn

IBM. (2022b, May 20). *What is Logistic regression?* https://www.ibm.com/topics/logistic-regression

Ivanov, S. H. (2022). Automated decision-making. *Foresight*, *ahead-of-print*.

Jansen, J. E. (2018, January 2). *Commentaar op Algemene wet bestuursrecht Artikel 3:3*. https://www.sdu.nl/content/commentaar-op-algemene-wet-bestuursrecht-artikel-33-detournement-de-pouvoir-artikeltekst-geldig

Jaspers, J. (2018, October 29). *Algemene beginselen van behoorlijk bestuur—Jaspers Overheidsrecht*. https://jaspersadvocaat.nl/kennisbank/begrippen/algemene-beginselen-behoorlijk-bestuur/

Jaspers, J. (2019a, August 5). *Wat is het beroep op gelijkheidsbeginsel?* https://jaspersadvocaat.nl/conflict-gemeente/beroep-gelijkheidsbeginsel/

Jaspers, J. (2019b, October 4). *Het vertrouwensbeginsel: Vertrouwen tussen burger en overheid.* https://jaspersadvocaat.nl/vertrouwensbeginsel/wat-is-vertrouwensbeginsel/

Jaspers, J. (2020a, January 13). *Zorgvuldigheidsbeginsel: Definitie en uitleg*. https://jaspersadvocaat.nl/recht-algemeen/zorgvuldigheidsbeginsel/

Jaspers, J. (2020b, February 20). *Verbod op détournement de procédure: Wat houdt dit in*. https://jaspersadvocaat.nl/recht-algemeen/detournement-de-procedure/

Kaminski, M. E. (2019). The right to explanation, explained. *Berkeley Technology Law Journal*, *34*(1), 189–218.

Khalilia, M., Chakraborty, S., & Popescu, M. (2011). Predicting disease risks from highly imbalanced data using random forest. *BMC Medical Informatics and Decision Making*, *11*(1), 51. https://doi.org/10.1186/1472-6947-11-51

Kim, T. W., & Routledge, B. R. (2018). Informational Privacy, A Right to Explanation, and Interpretable AI. *2018 IEEE Symposium on Privacy-Aware Computing (PAC)*, 64–74. https://doi.org/10.1109/PAC.2018.00013

Kim, T. W., & Routledge, B. R. (2022). Why a right to an explanation of algorithmic decision-making should exist: A trust-based approach. *Business Ethics Quarterly*, *32*(1), 75–102.

King, M. (2009). The problem with negligence. *Social Theory and Practice*, *35*(4), 577–595.

Klein, P. (2019, July 8). *Een ongekende heksenjacht*. RTL Nieuws. https://www.rtlnieuws.nl/columns/column/4773721/menno-snel-belastingdienst-toeslagenaffaire-ministerie-van-financien

Kluttz, D. N., Kohli, N., & Mulligan, D. K. (2022). Shaping our tools: Contestability as a means to promote responsible algorithmic decision making in the professions. In *Ethics of Data and Analytics* (pp. 420–428). Auerbach Publications.

Kosmas, I., Papadopoulos, T., Dede, G., & Michalakelis, C. (2023). The Use of Artificial Neural Networks in the Public Sector. *FinTech*, *2*(1), Article 1. https://doi.org/10.3390/fintech2010010

Krysovatyy, A., Lipyanina-Goncharenko, H., Sachenko, S., & Desyatnyuk, O. (2021). Economic Crime Detection Using Support Vector Machine Classification. *MoMLeT+ DS*, 830–840.

Levy, K., Chasalow, K. E., & Riley, S. (2021). Algorithms and decision-making in the public sector. *Annual Review of Law and Social Science*, *17*, 309–334.

Levy, N. (2003). Cultural membership and moral responsibility. *The Monist*, *86*(2), 145–163.

Levy, N. (2007). The responsibility of the psychopath revisited. *Philosophy, Psychiatry, & Psychology*, *14*(2), 129–138.

Levy, N. (2009). Culpable ignorance and moral responsibility: A reply to FitzPatrick. *Ethics*, *119*(4), 729–741.

Levy, N. (2011). *Hard luck: How luck undermines free will and moral responsibility*. OUP Oxford.

Levy, N. (2013). Moral responsibility and consciousness: Two challenges, one solution. In *Neuroscience and legal responsibility* (pp. 163–180). Oxford University Press.

Levy, N. (2014). *Consciousness and Moral Responsibility*. Taylor & Francis.

Lipton, Z. C. (2018). The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery. *Queue*, *16*(3), 31–57.

Lyons, H., Velloso, E., & Miller, T. (2021). Fair and Responsible AI: A focus on the ability to contest. *ArXiv Preprint ArXiv:2102.10787*.

Malgieri, G. (2021). “Just” Algorithms: Justification (Beyond Explanation) of Automated Decisions Under the General Data Protection Regulation. *Law and Business*, *1*(1), 16–28. https://doi.org/10.2478/law-2021-0003

Malgieri, G., & Pasquale, F. A. (2022). From transparency to justification: Toward ex ante accountability for AI. *Brooklyn Law School, Legal Studies Paper*, *712*.

Marcus, G., & Davis, E. (2019). *Rebooting AI: Building artificial intelligence we can trust*. Vintage.

Montmarquet, J. A. (1995). Culpable ignorance and excuses. *Philosophical Studies: An International Journal for Philosophy in the Analytic Tradition*, *80*(1), 41–49.

Nelkin, D. K., & Rickless, S. C. (2017). Moral responsibility for unwitting omissions. *The Ethics and Law of Omissions*, 1–32.

Oost. (2023, June 23). *DUO stopt met omstreden algoritme dat vooral studenten met migratie-achtergrond als fraudeur bestempelde*. NRC. https://www.nrc.nl/nieuws/2023/06/23/duo-stopt-met-omstreden-algoritme-dat-vooral-studenten-met-migratie-achtergrond-als-fraudeur-bestempelde-a4168024

Oswald, M. (2018). Algorithm-assisted decision-making in the public sector: Framing the issues using administrative law rules governing discretionary power. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, *376*(2128), 20170359.

Oswald, M., Grace, J., Urwin, S., & Barnes, G. C. (2018). Algorithmic risk assessment policing models: Lessons from the Durham HART model and 'Experimental' proportionality. *Information & Communications Technology Law*, *27*(2), 223–250. https://doi.org/10.1080/13600834.2018.1458455

Oxford Learners Dictionary. (n.d.). *Explanation*. Retrieved 27 May 2023, from https://www.oxfordlearnersdictionaries.com/definition/english/explanation

Parlement.com. (n.d.). *Ministeriële verantwoordelijkheid*. Retrieved 11 April 2023, from https://www.parlement.com/id/vh47j7avxf09/ministeriele_verantwoordelijkheid

Peels, R. (2011). Tracing culpable ignorance. *Logos & Episteme*, *2*(4), 575–582.

RapidMiner. (2021, December 2). *Bayesian Modeling*. RapidMiner. https://rapidminer.com/glossary/bayesian-modeling/

Rawat, S. (2022, July 21). *Advantages and Disadvantages of Neural Networks | Analytics Steps*. https://www.analyticssteps.com/blogs/advantages-and-disadvantages-neural-networks

Redactie Trouw. (2023, June 29). *Algoritmes moeten vaker kritisch onderzocht worden*. Trouw. https://www.trouw.nl/opinie/algoritmes-moeten-vaker-kritisch-onderzocht-worden~b39a09e8/

Redactie Volkskrant. (2021, November 23). *'Belastingdienst gebruikte algoritme dat lage inkomens selecteerde voor extra fraudecontroles'*. de Volkskrant. https://www.volkskrant.nl/nieuws-achtergrond/belastingdienst-gebruikte-algoritme-dat-lage-inkomens-selecteerde-voor-extra-fraudecontroles~bac84336/

Rijkswaterstaat. (2019, November 16). *Algemene beginselen van behoorlijk bestuur*. Kenniscentrum

InfoMil. https://www.infomil.nl/vaste-onderdelen/onderwerpen/lucht-water/handboek-water/handreiking-lozingen/procedure-keuze-rechtsbescherming/algemene-beginselen-behoorlijk-bestuur/

Robichaud, P. (2014). On culpable ignorance and akrasia. *Ethics*, *125*(1), 137–151.

Rosen, G. (2004). Skepticism about moral responsibility. *Philosophical Perspectives*, *18*, 295–313.

Rosen, G. (2008). Kleinbart the oblivious and other tales of ignorance and responsibility. *The Journal of Philosophy*, *105*(10), 591–610.

Rudy-Hiller, F. (2017). A capacitarian account of culpable ignorance. *Pacific Philosophical Quarterly*, *98*, 398–426.

Rudy-Hiller, F. (2022). The Epistemic Condition for Moral Responsibility. In E. N. Zalta & U. Nodelman (Eds.), *The Stanford Encyclopedia of Philosophy* (Winter 2022). Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/archives/win2022/entries/moral-responsibility-epistemic/

Rutten. (2022, January 25). *Belastingdienst schatte frauderisico's in op basis van nationaliteit en uiterlijk*. NRC. https://www.nrc.nl/nieuws/2022/01/25/belastingdienst-schatte-frauderisicos-in-op-basis-van-nationaliteit-en-uiterlijk-a4083150

Salles, R. (2005). *The Stoics on determinism and compatibilism*. Routledge.

Sartorio, C. (2016). Ignorance, Alternative Possibilities, and the Epistemic Conditions for Responsibility. In *Perspectives on Ignorance from Moral and Social Philosophy* (pp. 15–29). Routledge.

Schlick, M. (1966). *Problems of Ethics* (D. Rynin, Trans.; Vol. 14). Cambridge University Press.

Schravesande, F. (2023, June 21). *DUO keek door een gekleurd filter naar uitwonende studenten: Het algoritme zei 'fraude'*. NRC. https://www.nrc.nl/nieuws/2023/06/21/duo-keek-door-een-gekleurd-filter-naar-uitwonende-studenten-het-algoritme-zei-fraude-a4167810

SciKit. (2012, January 11). *Decision Trees*. Scikit-Learn. https://scikit-learn/stable/modules/tree.html

Selbst, A., & Powles, J. (2018). “Meaningful information” and the right to explanation. *Conference on Fairness, Accountability and Transparency*, 48–48.

Shafi, A. (2021, May 18). *What is a Generalised Additive Model?* Medium. https://towardsdatascience.com/generalised-additive-models-6dfbedf1350a

Sher, G. (2009). *Who knew?: Responsibility without awareness*. Oxford University Press.

Shoemaker, D. (2013). Qualities of Will. *Social Philosophy and Policy*, *30*(1–2), 95–120. https://doi.org/10.1017/S0265052513000058

Sliwa, P., Robichaud, P., & Wieland, J. (2017). On knowing what’s right and being responsible for it. *Responsibility: The Epistemic Condition*, 127–145.

Smith, H. M. (2011). Non-tracing cases of culpable ignorance. *Criminal Law and Philosophy*, *5*, 115–146.

Smith, R. (2020, July 23). *The Key Differences Between Rule-Based AI And Machine Learning*. Medium. https://becominghuman.ai/the-key-differences-between-rule-based-ai-and-machine-learning-8792e545e6

Talbert, M. (2013). Unwitting wrongdoers and the role of moral disagreement in blame. *Oxford Studies in Agency and Responsibility*, *1*, 225–245.

Talbert, M. (2022). Moral Responsibility. In E. N. Zalta & U. Nodelman (Eds.), *The Stanford Encyclopedia of Philosophy* (Fall 2022). Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/archives/fall2022/entries/moral-responsibility/

Tamagnini, P., Krause, J., Dasgupta, A., & Bertini, E. (2017a). Interpreting black-box classifiers using instance-level visual explanations. *Proceedings of the 2nd Workshop on Human-in-the-Loop Data Analytics*, 1–6.

Tamagnini, P., Krause, J., Dasgupta, A., & Bertini, E. (2017b). Interpreting black-box classifiers using instance-level visual explanations. *Proceedings of the 2nd Workshop on Human-in-the-Loop Data Analytics*, 1–6.

Timpe, K. (2011). Tracing and the epistemic condition on moral responsibility. *The Modern Schoolman*, *88*(1/2), 5–28.

Tversky, A., & Kahneman, D. (2000). *Choices, values, and frames*. Cambridge University Press.

Van Bekkum, D. (2023, June 23). *Minister Dijkgraaf stelt mogelijk discriminerend algoritme van Duo buiten werking*. de Volkskrant. https://www.volkskrant.nl/nieuws-achtergrond/minister-dijkgraaf-stelt-mogelijk-discriminerend-algoritme-van-duo-buiten-werking~b7b9dd55/

Van de Poel, I., Royakkers, L., & Zwart, S. D. (2015). *Moral responsibility and the problem of many hands*. Routledge.

Van den Oord, A., Dieleman, S., & Schrauwen, B. (2013). Deep content-based music recommendation. *Advances in Neural Information Processing Systems*, *26*.

van Goud. (2016, August 21). *Algemene beginselen van behoorlijk bestuur*. https://advocatenvastgoed.nl/specialismen/algemeen-bestuursprocesrecht/algemene-beginselen-van-behoorlijk-bestuur

Van Wart, M. (1996). The sources of ethical decision making for individuals in the public sector. *Public Administration Review*, 525–533.

Wachter, S., Mittelstadt, B., & Floridi, L. (2017). Why a right to explanation of automated decision-making does not exist in the general data protection regulation. *International Data Privacy Law*, *7*(2), 76–99.

Watson, H. J., & Nations, C. (2019). Addressing the growing need for algorithmic transparency. *Communications of the Association for Information Systems*, *45*(1), 26.

Wihlborg, E., Larsson, H., & Hedström, K. (2016). ‘ The Computer Says No!’–A Case Study on Automated Decision-Making in Public Authorities. *2016 49th Hawaii International Conference on System Sciences (HICSS)*, 2903–2912.

Wirtz, B. W., Weyerer, J. C., & Geyer, C. (2019). Artificial intelligence and the public sector—Applications and challenges. *International Journal of Public Administration*, *42*(7), 596–615.

Xu, F., Uszkoreit, H., Du, Y., Fan, W., Zhao, D., & Zhu, J. (2019). Explainable AI: A brief survey on history, research areas, approaches and challenges. *Natural Language Processing and Chinese Computing: 8th CCF International Conference, NLPCC 2019, Dunhuang, China, October 9–14, 2019, Proceedings, Part II 8*, 563–574.

Zimmerman, M. J. (1997). Moral responsibility and ignorance. *Ethics*, *107*(3), 410–426.

Appendices

Appendix 1: overview of rule-based decision aids

Model/system	*Brief explanation of how the system works*
Linear regression	Linear regression models are relatively simple, and can provide users with an easy to interpret mathematical formula which is used to generate predictions based on the values used (IBM, 2018).
Logical regression	Logical regression or logistical regression is often used for classification and predictive analysis (IBM, 2022). Logical regression is used to estimate the probability of the occurrence of a certain event, such as voting, based on a set of independent variables, through the log odds, or the natural logarithm of odds. The results are often shared in an odds ratio (OR), which makes interpreting the results easier.
Decision trees	Decision trees are a non-parametric supervised learning method, which can be used for both classification and regression purposes (SciKit, 2012). The aim is to create a model that can predict the value of certain targets, through the use of simple decision rules which have been inferred from the data.
K-nearest neighbours	The k-nearest neighbours algorithm is a non-parametric supervised learning classifier, which uses proximity to create classifications or predictions about the grouping of individual data points (IMB, 2022). For classification tasks, the label that is most frequently represented around a given data point is used.
Rule based learners	Rule based AI systems produces pre-defined outcomes based on a certain set of rules created by humans. These systems are simple AI systems, which are using if-then coding statements (Smith, 2020).
General additive models	General additive models are an adaptation of linear models, which allows the modelling of non-linear data while still maintaining explainability (Shafi, 2021). The equation used in these kinds of models is defined by the sum of a linear combination of variables, which are all given weight.

Bayesian models	Bayesian modelling is a statistical method of modelling where probability is influenced by the belief of the likelihood of a certain outcome (RapidMiner, 2021). Prior probability is used to inform the outcome, and this is updated while new evidence is received. This model assumes that the data used as input is all independent from each other.