Digital Future Society

Towards meaningful oversight of automated decision-making systems

A programme of

GOBIERNO DE ESPAÑA
VICEPRESIDENCIA PRIMERA DE GOBIERNO
MINISTERIO DE ASUNTOS ECONÓMICOS Y TRANSFORMACIÓN DIGITAL
SECRETARÍA DE ESTADO DE DIGITALIZACIÓN E INTELIGENCIA ARTIFICIAL

red.es

MOBILE WORLD CAPITAL BARCELONA

About Digital Future Society

Digital Future Society is a non-profit transnational initiative that engages policymakers, civic society organisations, academic experts and entrepreneurs from around the world to explore, experiment and explain how technologies can be designed, used and governed in ways that create the conditions for a more inclusive and equitable society.

Our aim is to help policymakers identify, understand and prioritise key challenges and opportunities now and in the next ten years in the areas of public innovation, digital trust and equitable growth.

Visit digitalfuturesociety.com to learn more

A programme of

GOBIERNO DE ESPAÑA
VICEPRESIDENCIA PRIMERA DE GOBIERNO
MINISTERIO DE ASUNTOS ECONÓMICOS Y TRANSFORMACIÓN DIGITAL
SECRETARÍA DE ESTADO DE DIGITALIZACIÓN E INTELIGENCIA ARTIFICIAL

red.es

MOBILE WORLD CAPITAL™ BARCELONA

Published
October 2022

Disclaimer
The information and views set out in this report do not necessarily reflect the official opinion of Mobile World Capital Foundation. The Foundation does not guarantee the accuracy of the data included in this report. Neither the Foundation nor any person acting on the Foundation's behalf may be held responsible for the use which may be made of the information contained herein.

Digital Future Society

Table of contents

1. Introduction

Artificial Intelligence (AI) has reached almost all sectors all over the world. Thanks to the promise of accelerated efficiency and the digital fulfilment of repetitive tasks, the number of public administrations using automated decision-making systems (ADMS) has been increasing year on year (Misuraca and Noordt 2020; Zuiderwijk et al. 2021). There are risks, however, including the potential for historical inequalities, biases and discriminations becoming enshrined in algorithms (Digital Future Society 2020). Furthermore, automating decisions raises questions around liability. Who is responsible if an algorithm makes a discriminatory decision?

Clearly, governments face challenges as they attempt to increase efficiency through the automation of processes and the digitisation of society, and Digital Future Society (DFS) has been contributing to the debates relating to these challenges.[1] The DFS white paper Governing algorithms: perils and powers of AI in the public sector, analyses the attempts made by governments to implement ADMS (Digital Future Society 2021). It seeks to understand the different levels of governing algorithms and raises awareness of the false promises made about these systems. Rather than being governed by AI, the paper talks of a process of governance with AI, seeking "the classical situation of using and controlling a technology that reinforces our capacity, through a process that requires human supervision" (Ibid.).

ADMS can be problematic due to the reproduction of (already existing) biases leading to discriminatory outcomes caused by, for example, the design of these systems or the data used to train them. A key question then, needs to be how can we ensure a trustworthy use of ADMS in public administration?

Human agency and oversight are expected to solve this issue by supervising and taking the last word on a decision or mitigating errors in the data or the algorithm's output. However, current regulation poses human oversight as a safeguard for algorithmic systems, which is ambiguous at best.

At this moment, the European General Data Protection Regulation, commonly referred to as the GDPR, marks the only existing European regulation in this regard. However, it is not an AI-specific regulation, focusing primarily on data protection. The GDPR requires algorithm developers, organisations, and administrations to implement human supervision mechanisms, known as human oversight. These mechanisms see a human taking measures in the case of errors or discriminatory outcomes from the algorithm. However, the proposed mechanisms are too vague and do not cover the complexities of real-world use cases (Green 2021).

1. See DFS reports and white papers including Governing algorithms: perils and powers of AI in the public sector (https://digitalfuturesociety.com/es/report/governing-algorithms/), Gender bias in data: Towards gender equality in digital welfare (https://digitalfuturesociety.com/es/report/hacia-la-igualdad-de-genero-en-el-estado-de-bienestar-digital/), Inclusion by design: exploring gender responsive designs in digital welfare (https://digitalfuturesociety.com/es/report/exploring-gender-responsive-designs-in-digital-welfare/) and Where emerging tech meets government (https://digitalfuturesociety.com/es/report/donde-la-tecnologia-emergente-se-encuentra-con-el-gobierno/)

Furthermore, the regulation assumes that automated systems are free of human agency when, in reality, human interaction within such systems can take many forms and the implications of each of these different types of interaction need to be taken into account (Binns and Veale 2021).

As GDPR is not an AI-specific regulation, it is perhaps understandable that the complexity of the issue is overlooked, yet the problem remains that current legislation does not contain an understanding of the complex nature of human oversight of ADMS. There is no clarity on when or under which conditions human oversight can be a satisfactory response to bias, harm and problematic experiences as witnessed in the use of ADMS. Unfortunately, the only thing that is clear so far is that what seems like a simple endeavour, humans overseeing automated decisions, is much more complex and sometimes counterproductive (Campolo and Crawford 2020). A positive aspect, however, is that there is further European AI regulation on the horizon, but it remains to be seen whether it will truly grasp the complexity of implementing meaningful human oversight of ADMS.

Figure 1. **Human Oversight**

Image source: Digital Future Society.

About this Policy Brief

This policy brief seeks to inform the audience about the complexities behind human oversight, in its definition, regulation and practice. Given that there is a lot of debate on how human-algorithm interaction should be regulated and if human supervision as required presently is enough to mitigate algorithmic harms, this policy brief contributes to the debate exploring both regulation and the field of studying human-computer interaction, in an effort to propose recommendations that can help create meaningful human involvement.

In the first section, the document explores how EU regulation defines human supervision of automated decision-making systems. It provides an explanation of how existing regulation, the GDPR, requires human supervision and later on, it takes a look at the AI Act, foreseeing areas where the proposed regulation may fall short in providing sufficient protection.

The second section draws from the field of studying human-computer interaction to provide a better understanding of the many different layers encompassing automated decision contexts. First, the section looks at the human aspect of the interaction, specifically, what biases influence humans in the decision-making process. Then, the section moves to the different roles that automation plays in assisting humans in the decision-making process, by referring to the simplified structure proposed by academics Reuben Binns and Michael Veale.

Using these simplified typologies, the third section presents case studies of automated decision-making systems in Europe. Through the analysis of these case studies, the document aims to build a clearer picture of under what circumstances human oversight can be effective or not.

The policy brief then proposes a set of recommendations in an effort to enhance human oversight as a meaningful tool in ADMS.

Why now?

"Policymakers and companies eager to find a "regulatory fix" to harmful uses of technology must acknowledge and engage with the limits of human oversight rather than presenting human involvement — even "meaningful" human involvement — as an antidote to algorithmic harms. This requires moving away from abstract understandings of both the machine and the human in isolation, and instead considering the precise nature of human-algorithm interactions." [2]
— Ben Green and Amba Kak

With the increasing uptake of automation tools in the public sector, policymakers, government officials and administrators need to understand how automation impacts decision-making contexts. Although human oversight is generally promoted as a safeguard by regulation, experts caution against the false sense of security that human oversight promises, as the risk

2. See: https://slate.com/technology/2021/06/human-oversight-artificial-intelligence-laws.html

ADMS pose lie beyond the discretion of frontline workers. Regulation, as of now, reflects a superficial understanding of the human machine interaction. Therefore, to effectively minimise the harms of bias and discrimination in decision-making, whether it be from humans or algorithms, policymakers should first understand the risks and complexities behind the use of ADMS and how human oversight can play a meaningful role.

Complexities to consider

Moreover, when talking about human oversight there is often a misplaced assumption that human intervention should only occur during critical moments. In the context of ADMS, this critical moment is often interpreted as the final decision. This misunderstanding has, for example, led to current European regulation that does not understand that the system can have different relationships with different types of users. Another factor is that in an effort to streamline processes, system designs may often bypass human input altogether meaning there are systems in place today that do not have mechanisms to ensure effective oversight. Other considerations include the possibility that humans could fail to detect when they are being influenced by machines or when the machines are coming to incorrect conclusions, or that the operators themselves may be biased against a particular decision or lack the information, authority, or understanding to correctly intervene in a process. These represent some of the overlooked tensions that proposed AI regulation should clarify so that it will ensure successful human oversight. The case studies presented later on will explore the consequences that result from failing to take them into consideration.

Methodology

The content of this policy brief draws on an extensive literature review of three specified fields and a consistent review of case studies. The literature review explored risk communication, algorithmic audit systems and human-computer interaction. The review gave particular attention to publications in well-known journals and from conferences regarding algorithmic fairness and ethics.[3]

To select the case studies, the research explored different algorithmic observatories[4] for cases involving the use of ADMS based in Europe that include any type of human interaction. The case studies were then analysed against the framework developed by Reuben Binns and Michael Veale (2021) to select cases that illustrate each of the identified instances.

The research also studied documents available from the selected cases (official reports, research papers, available news articles, etc.) to identify the process of design, implementation, and development to try and identify the challenges and opportunities related to each experience and draw conclusions.

3. For example, see: FaccT conference (https://facctconference.org), AIES Conference (https://www.aies-conference.com) and CHI conference (https://chi2022.acm.org).

4. For example, see: the AlgorithWatch Automating Society report (https://automatingsociety.algorithmwatch.org) and Etica's OASI Report (https://eticasfoundation.org/oasi/).

2. Putting human oversight in context

Defining Human Oversight

In general, the term *oversight* is used in public policy at different levels, implying institutional transparency, public accountability, or agency over outcomes.[5] Current and proposed regulation does not help define such diffuse implications.

> In the context of this policy brief, we propose the following definition: **human oversight is mainly referred to as the agency that a human operator or supervisor of an (algorithm-based) system can pose to mitigate any harm or malfunction caused by the system.**

A common understanding of human oversight refers to placing humans back in control of a process that has been automated. That is, humans should be empowered to have enough agency over the system (i.e., be able to control or override its decision) as a form of risk mitigation. One of the most well-known approaches is known in the academic and technical sectors as the "human-in-the-loop" solution (HITL).

> **HITL refers to the capability for human intervention at every decision cycle of the system, placing the human back at the centre of the decision-making process.**

While the field of human-computer interaction has been analysing HITL for several decades, HITL was initially a form of operationalising human intervention in critical systems that required human discretion, like aviation or robotics (Dourish 2001). As this document will explore, HITL has broadened in scope as automated decisions have increasingly become an integral part of public services. This includes examples that highlight how human agency is neither possible nor desirable in many cases as it may cease to become an effective measure in curbing algorithmic harms or errors.

5. For example, Facebook has recently created an Oversight Board to provide answers and increase accountability to specific decisions made by the company's algorithm (see https://oversightboard.com).

Human Oversight in European regulation

Human Oversight in the European General Data Protection Regulation (GDPR)

Article 22 of the GDPR hints at human oversight with a prohibition on fully automated individual decision-making (European Parliament and Council of the European Union 2016).

The term human oversight does not appear in the regulation, but it does require human intervention in the case of ADMS. Article 22 states, a person "shall have the right not to be subject to a decision based solely on automated processing [...] which produces legal effects concerning him or her or similarly significantly affects him or her".

That statement includes those systems used for profiling or providing scoring to access social benefits, but only if systems are fully automated. In the third paragraph, "[the caseworker or government representative] shall implement suitable measures to safeguard the data subject's rights and freedoms and legitimate interests, at least the right to obtain human intervention on the part of the controller, to express his or her point of view and to contest the decision." (Ibid.).

Nevertheless, Article 22 appears confusing about what can be defined as "solely an automated process" and what type of human intervention is sufficient as a "suitable measure". Furthermore, it is not clear to what extent a system should be designed to include differing points of view and to allow the human (human reviewer, overseer) to contest the automated decision. In other words, the article only focuses on decisions made by automated systems and does not contemplate the whole range of possibilities where humans and organisations can participate in a decision-making system (Brkan 2017).

There are many different possible situations with examples including when humans gather the data used to train an algorithm, meaning it is not fully automated; when an automated system does not make a final decision but only classifies information about the citizen; or when a caseworker needs to approve the final decision. The proposed regulation overlooks this broad range of potential human oversight opportunities. Unfortunately, these opportunities allow humans, intentionally or not, to embed values, biases and assumptions into decision-making and so require much more attention and consideration. Without acknowledging that ambiguity, incompletion, and uncertainty are also part of the human decision, including humans in algorithm-supported decisions does not appear to be a direct path to a solution (Birhane 2021).

In addition, automated decisions involving the data subject's (e.g., a citizen's) personal data should also comply with Articles 13 and 14 of the GDPR that require the data subject to have access to "meaningful information about the logic involved" (European Parliament and Council of the European Union 2016). It is well known that many algorithms are black boxes, and remain challenging to explain even for technical specialists (Miron 2018). Yet, because the GDPR is created from the point of view of the data ecosystem, these articles do not

contemplate the complexity of algorithmic systems and ignore the difficulties the decision-making processes present to the human operators of ADMS. For example, current regulation does not contemplate the end users need for explainability of an ADMS nor does it establish a requirement for these systems to be contestable.

Human oversight in the High-Level Expert Group on AI's White Paper on Artificial Intelligence

Following on from the articles present in the GDPR, the High-Level Expert Group on AI (AI HLEG) has been defining requirements for establishing a more concrete regulation on AI. In 2020, the White Paper on Artificial Intelligence collected suggestions from the AI HLEG for an EU regulation (European Commission 2020). On page 21 of the white paper, the following "non-exhaustive manifestations" of human oversight appear in Section D(e):

- The output of the AI system does not become effective unless it has been previously reviewed and validated by a human to confirm the decision (e.g., the rejection of an application for social security benefits may be taken by a human only).
- The output of the AI system becomes immediately effective, but human intervention is ensured afterwards to override the decision (e.g., the rejection of an application for a credit card may be processed by an AI system, but human review must be possible afterwards).
- Monitoring of the AI system while in operation and the ability to intervene in real-time and deactivate (e.g., a stop button or procedure is available in a driverless car when a human determines that car operation is not safe);
- In the design phase, by imposing operational constraints on the AI system (e.g. a driverless car shall stop operating in certain conditions of low visibility when sensors may become less reliable or shall maintain a certain distance in any given condition from the preceding vehicle).

Unfortunately, although these manifestations represent clear examples of how to conceive human oversight in different scenarios, the white paper fails to define high-risk AI systems, even though the European Commission previously published a recommendation addressing the issue.

Human oversight in the proposed Artificial Intelligence Act

In early 2021, the European Commission published a recommendation for AI regulation (known as AI Act or AIA) (European Commission 2021). This included the need for human oversight in high-risk AI systems while, human supervision is optional at lower levels of risk.

A high-risk AI classification depends on what is at stake, considering whether the sector and the intended use involve significant risks. The high-risk AI systems are generally defined in Article 6 and specified in Annex III of the AI Act. These include scenarios where automation can be applied, like criminal justice, child welfare, human resource recruitment, or the use of biometric data for individual identification.

However, definitions of human oversight in the context of high-risk AI systems disappeared from the AI Act, leaving only a limited description in Article 14 (Ibid.):

> **Human oversight shall aim at preventing or minimising the risks to health, safety or fundamental rights that may emerge when a high-risk AI system is used in accordance with its intended purpose or under conditions of reasonably foreseeable misuse, in particular when such risks persist notwithstanding the application of other requirements set out in this Chapter.**

According to the AI Act, human overseers should be able to (Ibid.):

- understand the capacities and limitations of the high-risk AI system,
- remain aware of the possible tendency of automatically relying or over-relying on the output produced by a high-risk AI system,
- be able to correctly interpret the high-risk AI system's output,
- be able to decide, in any particular situation, not to use the high-risk AI system or otherwise disregard, override or reverse the output of the high-risk AI system,
- be able to intervene in the operation of the high-risk AI system or interrupt the system.

Human oversight blindspots in EU regulation

Presently, the GDPR is the only applicable regulation that requires human oversight for automated decision-making systems. However, much is expected to change with the AI Act, which is promised to go further and ensure that AI systems do not pose a risk to health, safety, or fundamental rights. Even so, there are debates around whether the proposed AI Act effectively contemplates the complexities of human oversight of algorithmic systems.

The regulation does face some challenges and may be difficult for actors to comply with because of the terminology it uses, unavoidable ambiguity in the scenarios it addresses, and the rapidly changing contexts of AI development (Green 2021).

Below are a few points which further explain the tensions between the two:

1. The proposed AI Regulation adopts technical jargon and terminology to refer to particular roles in the algorithmic systems (e.g., user, provider, the data controller, the data subject). **It fails to address the contextual complexities of how these systems interact with other systems and institutional structures**, leaving an ambiguous antecedent when it comes to real world use and implementations.

2. **The AI Act assumes that ADMS are sold by third parties (providers) and adopted by users, yet users are not clearly identified in the context of particular use cases**. For example, are users the representatives of a government, the citizen, or the IT staff? Also, Article 29, titled "Obligations of users of high-risk AI systems" states that the provider should embed human oversight into the system. In case of system errors, the only action provided is that the user (again, the user is not clearly defined) should inform the provider or distributor and suspend the use of the system. No other mitigation is suggested. However, it is unknown if the providers are external companies or agencies and if they know how the current analogue system works and how decisions are made.

3. Despite efforts to provide guidance on defining scenarios to implement AI and ADMS, AI is provoking changes in new contexts that are increasing every year. Contexts evolve, and algorithms are constantly updated, acquiring new capabilities. It seems that **regulatory mechanisms are limited because they address algorithms that have been designed only for particular uses** of automated decision-making (see Annex III) and might fail in many others. Therefore, the regulation exists within grey areas that can be difficult to regulate or fall into unprecedented and ambiguous situations where these rules become unclear to organisations and governments.

Broader blindspots to consider

More specifically, with regards to human oversight, the regulation may also prove to be too vague or ambiguous to effectively ensure the "meaningful" human oversight it wishes to promote. Furthermore, in some cases, human oversight might not be the adequate measure to mitigate the risk that automation poses. Provisions like the GDPR and the AI Act introduce **human oversight of automated decisions into implementation grey areas**. Below are a few areas which may be problematic to act on in practice:

- **As the GDPR prohibits the use of solely automated decision-making systems, it gives the impression that current systems do not have any intervention or supervision**. But, as researchers Reuben Binns and Michael Veale argue, it is hard to find real-world cases where decisions are made solely by an automated system or operated without any human involvement or supervision (Binns and Veale 2021). In the context of ADMS and high-risk systems, it is common to see humans playing a part in systems that see algorithmic implementation, either introducing the data that feeds the algorithm or actually taking the final decision. For example, judges use risk assessments to inform pretrial

and sentencing decisions, child services workers use predictive models to inform which families to investigate for child neglect and abuse, and welfare agencies use algorithms to determine eligibility for benefits to be confirmed by agents. As explained earlier, these implementations always include humans in the course of the decision, but humans do not always intervene in the final decision. The result is a misguided prohibition that may never be possible to implement in real world situations.

- As the GDPR and the AI Act lay out, more "meaningful" forms of human oversight and discretion are essential for protecting values like human rights. But **these "meaningful" interventions are ambiguous and difficult to accomplish in practice**. Many scenarios that see humans supervising these systems occur without the human operators having sufficient training, motivation, agency to redress, authority, or competencies to provide any kind of "meaningful" form of oversight.

- Application of Article 22 of the GDPR depends on whether an automated decision has legal effects concerning the data subject (i.e., the citizen) or, alternately, if it is "significant" (i.e., can have a remarkable impact on the life of the individual). In a case of legal effect, a human should oversee the decision. However, **legal effects are restricted to cases where legal status is altered or legal duties created** (e.g., assessment of immigration status or authentication of a legal contract), but "significant" effects are much vaguer.

- Most importantly, **presenting human oversight as a solution for potential harms can lead to a blurring of responsibility**. On the one hand, even when a system is assumed to have mitigation mechanisms, humans can over-rely on such mechanisms and not supervise the systems appropriately (e.g., rubber stamping). This offers a way to bypass scrutiny and consequences. On the other hand, human oversight could be an excuse to push attention onto human operators, enabling developers and companies to enhance promises like efficiency and optimisation, while leaving governmental bodies and civil servants with the responsibility to correct errors and mistakes.

To sum up, it is unclear in which scenarios and under which conditions human oversight can be a satisfactory response to bias, harm and problematic experiences as witnessed in the use of ADMS. In the following section, the policy brief addresses ADMS at the intersection of several disciplines to understand the extent to which human oversight can be helpful and effective.

3. Understanding context surrounding decision-making

The role of human operators in decision-making can be critical in certain scenarios, such as border control, social welfare or criminal justice where the decision could limit the rights and benefits of an individual or an entire social group. When interacting with an ADMS, human operators are expected to give specific responses, which can range from a quick reaction in particular contexts, like bypassing an alert in an airport security control, to an in-depth and elaborated answer for high-risk settings, like judges in pretrials and sentencing.

In order to understand the complex context of decision-making, academics highlight the importance to consider the many layers that encompass the use of automated decision-making tools. To understand this complexity requires studying how humans behave and interact with machines, and moreover, acknowledging the organisational, legal and sociocultural environment. Within this context, there are human factors to take into account, such as the workload of the human operator, their motivation, confidence and trust in the automated tool. While, on the other hand, the performance of the system itself, which could range from its transparency, effectiveness as a tool, etc. should also be considered (Ananny and Crawford 2018; Kemper and Kolkman 2019; Zhang et al. 2020; Lee and See 2004).

For the BODEGA project[6], a research group that analysed the decision-making context in the EU's automated border control gates (later analysed in the case study section), the human factors framework (Figure 1.) proved to be a useful model to visualise the human factors in border guard's work and their interrelatedness. The framework describes the environment where border guards operate and defines the factors that contribute to the system performance, all which can be extrapolated to other decision-making contexts. These range from broad environments such as: (1) **the social and cultural environment**, which describes the values, norms and public opinion; (2) **the legal environment**, which conditions the legal implications of the system, for example what type of automation or data can be processed; (3) **the organisational environment**, heavily influenced by the two previous environments, and that includes the organisational culture and structure; and finally, (4) **the operational environment**, which describes the physical and spatial environment where the decision-making takes place.

6. For more information, see: https://bodega-project.eu/.

Figure 2. **Border Control Human Factors Framework**

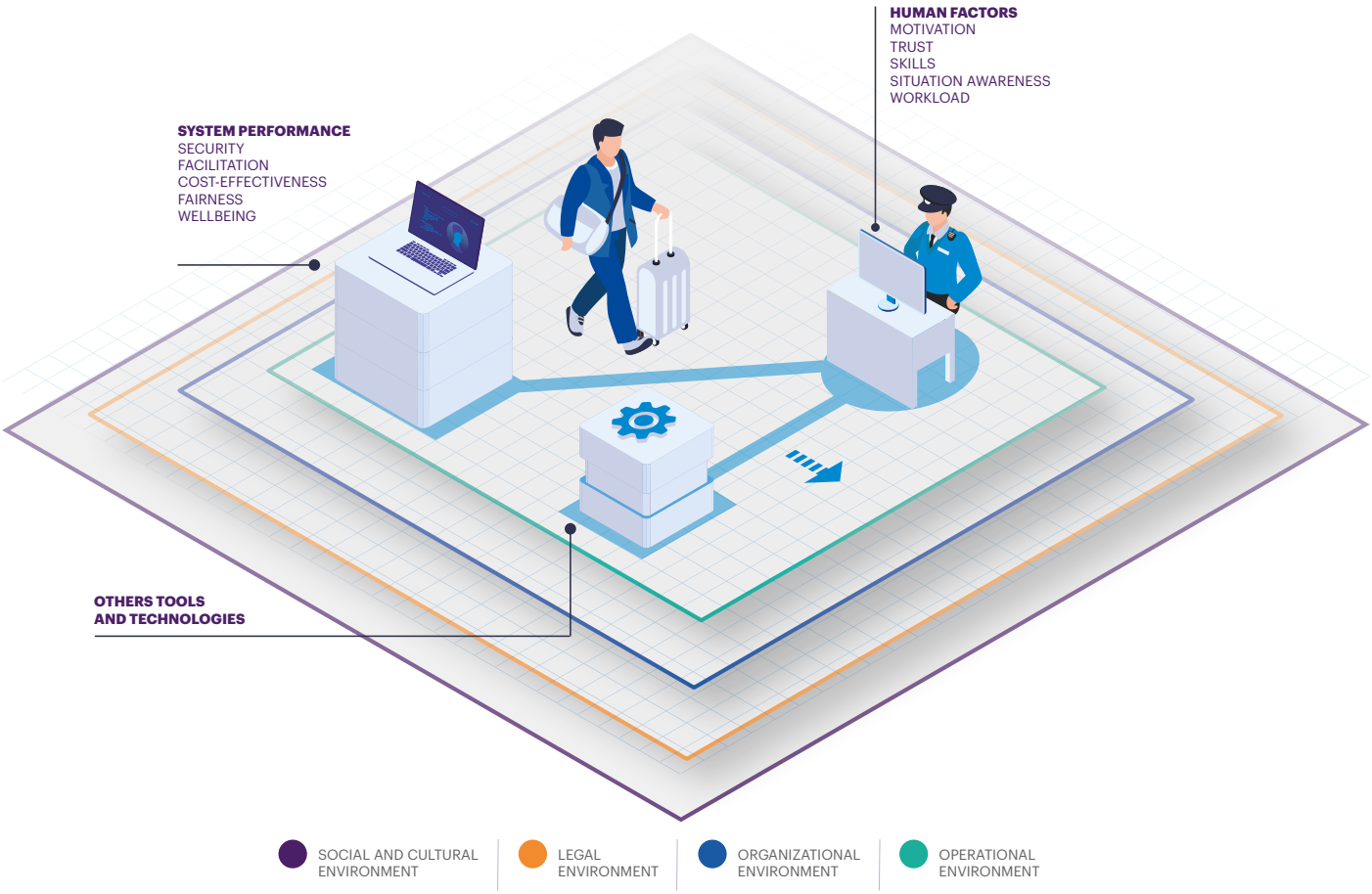


Image source: Kulju et al. 2019.

This Figure is taken from the study of border controls and illustrates how the decision context (environments) and the human factors are entangled during the process of decision-making tasks. In addition, the interaction with the system and other tools and technologies enables or blocks the successful implementation of the task. In sum, in this scenario, the passenger goes through a process influenced by several factors.

Going into deeper detail, **the broader organisational environment** involves the organisations that develop the software and those that use them — they can be the same actors, or as the case studies later show, these tools may be acquired or developed by public sector administrations. These organisations such as public institutions and the private sector, heavily influence, among other things, which procedures are automated, what data is used for training algorithms, how the systems are designed, and the different levels of liability

found within the system. In particular, they define how and where human oversight is placed as well as selecting the operators who are in charge of making those decisions. This is highly important in terms of human oversight because neither humans nor algorithms are free of contextual bias or exposure to biased decisions, and a system should foresee associated potential risks.

Regarding the operational environment, human decision-makers are influenced by interaction with technological devices and their associated interfaces. This environment is heavily influenced by the human factors mentioned previously and impact the decision-makers performance which can be positive, neutral, or negative depending on the situation. For example, for human operators to be able to give a discrete and justified response, they might have a different level of expertise in the context of use, from long-experienced professionals to operators without expertise (Myers-West et al. 2019). A user's ability to use an ADMS successfully can depend on how the system provides relevant information as well as their expertise in understanding data and visual communication.

Therefore, at the moment the decision is made, humans are interacting with algorithms through the interface of a machine while also being influenced by the human factors described above. For scholars in Human-Computer Interaction (HCI), it is imperative to consider these interactions within the context the interaction takes place (Suchman 2007). For that reason, researchers have been trying to create frameworks to convey all the complex variables at play during a human-machine interaction. Things like communication abilities, technological skills, personal variables, attitudes, or motivations (among others) can define the expected behaviour/interaction with a system (Cranor 2008).

"When the human element is introduced into decision support system design, entirely new layers of social and ethical issues emerge but are not always recognised as such" (Cummings 2006).

4. The complexities of human oversight

As mentioned in the previous section, there are various constraints that influence the decision-making context. These factors of varying scale, impact the way human operators oversee automation. For the regulation, a meaningful oversight is when operators exercise their agency while being aware of the system's (and their own) biases or limitations. This would mean human operators are able to prevent harms if they can understand when an algorithm errs, understand why an algorithm has made a decision and account for the potential biases of the system. Therefore, in theory, for human oversight to be effective, the system design should also consider the limitations and biases of human operators. This section lists a series of limitations that may present a risk to effective human oversight — therefore creating a false sense of security. This non-exhaustive list is important to consider as later, they will prove useful for analysing the case studies.

1. Humans have limited capacity to interpret and process complex information in short periods of time or when under pressure

To make a judgement, perception and understanding of information is vital. And although humans are equipped to do so, making judgements is challenging for humans (Kahneman 2011). In the context of ADWMS, it involves understanding abstract concepts, reading graphics, and contextualising information. However, when faced with work and information overload, human performance will drop (Balfe et al. 2018). Humans are also less capable of processing more than three or four pieces of information simultaneously, while computers can make hundreds of simultaneous calculations.

Algorithms deal with significant amounts of information and complex data relations better than humans. However, despite their processing capabilities algorithms lack context (e.g., an understanding of the meaning of the information being processed). Also, many algorithms such as deep neural networks are seen as hard to understand by the human mind, known as black boxes (Rudin 2019). However, not all types of algorithms fall into this definition. For example, most machine learning techniques used for ADMS, like logistic regressions or decision trees, are easy to understand.

All this means that, depending on the decision-maker's previous knowledge and training and the type of algorithm implemented, the algorithm output may be harder or easier to understand and contextualise in limited timeframes. When humans are not adequately trained to understand and process context-sensitive information, the information can end up confusing or misinforming them, particularly in contexts of high societal risk (Scurich et al. 2012; Batastini et al. 2019).

2. Humans have difficulty understanding the algorithm's role in the course of a decision

For humans to override algorithmic decisions, they should be able to identify when and where such decision-making systems produce errors as well as the algorithm's influence in causing the error. Algorithms can play different roles, they can either support the decision-maker in providing information to aid in the process, or they can reach a final decision with a decision-maker's input (these different typologies of systems will be discussed in further detail, in the following section). Consequently, humans might not be conscious of the particular role the algorithms have played in a particular decision. In a study exploring risk assessments in human decision-making processes, researchers Ben Green and Yiling Chen found that incorporating risk assessments in human prediction, such as those used to predict criminal re-offending, is a very challenging task that requires more expertise than expected. According to the study, the participants were unable to accurately assess their performance nor that of the algorithm. The study found that participants were most successful when they followed the risk assessment — which is problematic given the inherent biases of risk assessments in criminal justice systems. Green and Chen argue that there is little evidence that proves that risk assessments along with the decision-maker's own judgment lead to better decisions — which attests to the limited understanding of whether and how risk assessments aid in decision-making processes (Green and Chen 2019a).

3. Humans are unlikely to question suggestions made by algorithms (over-reliance)

Instead of considering a suggestion made by an algorithm, human decision-makers could be 'rubber-stamping' — giving approval without consideration. Ben Wagner calls this a "quasi-automation" process not contemplated by the current and proposed regulatory frameworks (Wagner 2019). Furthermore, Ben Green and Yiling Chen have identified several issues that could lead to rubber-stamping (Green and Chen 2019a). For example, *automation bias*, where people tend to over-rely on automated suggestions, or *affirmative bias*, where human individuals tend to agree with an automated decision that coincides with their values and beliefs.

Effective human oversight is especially challenging to implement given the constantly evolving context that algorithms need to adapt to if they are to remain accurate and relevant. In general, algorithms are trained with particular data and deployed into a particular digital system. Algorithms may be re-trained with new data, but they often remain the same for weeks, months, or years depending on the type of system and requirements.

Algorithms should be constantly updated to reflect all relevant social and cultural changes. However, if the data that feeds those updates is biased and is coupled with human oversight that fails to recognise these biases, or over relies on the automated decision, it will reinforce prejudices. This practice might even reintroduce new biases that were previously mitigated. Consequently, new algorithms are developed, and new biases go unnoticed.

The above is known as a feedback effect, and it could lead to unexpected or undesirable biases such as self-fulfilling prophecies (i.e., a prediction that comes true because you made it and acted as if it were true), or predictions that feed a morphing of the social situation (Barocas et al. 2019). This means interacting with automated decisions without proper analysis could reduce the benefits of human discretion, or in a worst-case scenario, cause new harms.

"Because of the inherent complexity of socio-technical systems, decision support systems that integrate higher levels of automation can possibly allow users to perceive the computer as a legitimate authority, diminish moral agency, and shift accountability to the computer, thus creating a moral buffering effect" (Cummings 2006).

4. Humans may erroneously value human judgement over algorithmic recommendation (under-reliance)

As mentioned previously, effective human-algorithm interaction depends on the level of training, the human decision-maker's experience working with the algorithm, and the specific professional domain in which the decision is made. As operators may over-rely on the automated decision, at the opposite end of the spectrum, under-reliance describes the phenomena in which the human operator disagrees or deviates from the algorithmic recommendation. This may be due to a series of reasons, depending on the operator's training, perceptions, workload, etc.

In some contexts, where the final decision requires an experienced and conscientious understanding of the process, evaluators might not take full advantage of the benefits of predictive accuracy, overriding the algorithm's decision even when it has made valid predictions (McCallum et al. 2017). This happens mainly because human evaluators do not understand the reasoning behind the automated decision-making process as they do not have access to data and therefore cannot assess the prediction process. Human decision-makers may also exhibit algorithm aversion by discontinuing the use of a particular algorithm, following a mistake, even if, on average, the algorithm is more accurate than they are (Dietvorst et al. 2014; Burton et al. 2020; De-Arteaga et al. 2020). Consequently, the human in charge stops trusting the system's outcome.

A similar case occurs when, even with clear information and a correct suggestion, experienced decision-makers may be more inclined to deviate from algorithmic recommendations, relying instead on their cognitive processes (Green and Chen 2020). Instead of mistrusting the system, like in the previous point, humans can deviate from an algorithm's recommendation because their objectives might not align with the algorithm's optimised purpose, or because the context may create incentives for the human decision-maker to deviate from the recommendation (Green 2020; Stevenson and Doleac 2019).

5. Defining different typologies of human-algorithm interaction

The border guard framework, discussed earlier, illustrates how human-computer interaction goes beyond the simple operation of a machine, and can be approached from different scales or environments — whether they be organisational, legal, etc. (as illustrated in Figure 1.).

However, as not all automated systems are alike, there are specific complexities that should be addressed from a public sector perspective. To further explain what automation looks like for ADMS in a public sector decision-making context, this policy brief will make reference to three different automation typologies (see Figure 2.). These typologies, defined by researchers, Reuben Binns and Michael Veale, are especially useful to help explain how automation supports human decision-making.

As the researchers explain, **these typologies are simplified versions of the different roles that automation can play, and, in reality, ADMS can include more than one of these typologies**. Nonetheless, for the sake of clarity, this document will analyse them separately. The following roles are:

- **Summarising:** the system consolidates human interventions/data from one or more decision-makers that leads to an automated decision.
- **Supporting:** the system provides information to the human decision-maker with the human then considering the system's "advice".
- **Triaging:** the system automatically processes cases unless these are flagged for human review.

Each typology can be divided into two key moments. These key moments are the *upstream* process of automation — where data is collected, systematised, and processed — and the *downstream* process, which is during the deployment and monitoring stages, after the automated outcome.

The upstream could be understood as the process of information gathering, which is fed into the algorithm, while the downstream process involves everything that takes place once the algorithm has provided the output.

Using the three structures mentioned above, we can better understand how a human can intervene in the process.

When a system is **Summarising**, for example, human decision-makers contribute to the upstream of the system, by **providing assessment or evaluation that is turned into structured data**. One such example of summarisation, which will be explored later, is the evaluation of inmates. Assessments encoded into numerical scores produced by caseworkers

undergo an automated process that results in risk scores assigned to inmates. Systems that provide information (**Supporting**) requires human intervention in the downstream, that is, **humans receive an automated score and consider it along with additional information to make a final decision**.

Triaging provides automated decisions in which humans can intervene after the initial automated analysis. If the system detects that the case at hand is suspicious or in need of human review, they then require human input in the downstream. A type of system that falls into this category, is, for example, anomaly detection systems. The *Udbetaling Danmark (UDK)* (Payout Denmark) uses this type of system to detect error or fraud in welfare payments. While lastly, systems that are **no intervention** means humans cannot provide any input in the automated process.

Figure 3. **Typologies of automated decision-making and human intervention**

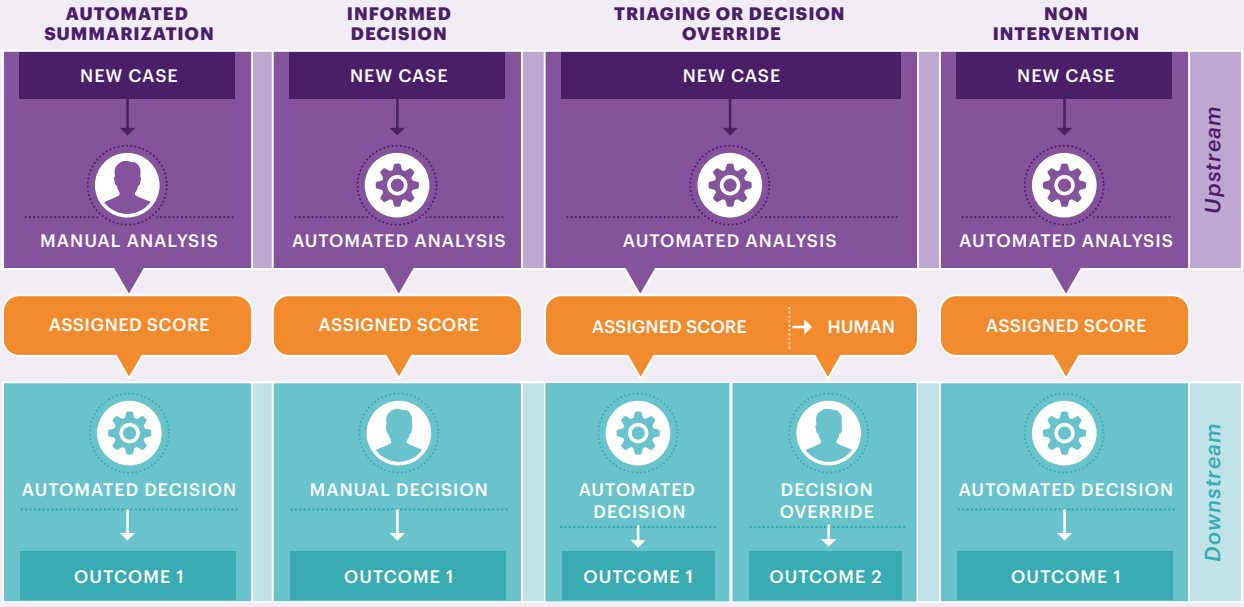


Image source: Binns y Veale 2021.

The case studies examined in the following section demonstrate how these different typologies of human-algorithm interactions overlap with each other in many systems and the consequences that that arise from their presence.

6. Case studies

Having explored human oversight, different typologies of human-ADMS interaction and examined the related complexities, this policy brief now seeks to show human oversight in practice through the analysis of three case studies. Given the paper's focus on European legislation, all cases have been selected from public sector applications of human oversight of ADMS in Europe:

1. ***Udbetaling Danmark (UDK)* (Payout Denmark)** — a data driven application to detect error and fraud in welfare payments (Triage).

2. **Frontex's Automated Border Control** — automation of border control in selected Schengen states (Triage).

3. **RisCanvi** — a risk assessment protocol that predicts violent recidivism among inmates in Catalonia, Spain (Summarising).

The case studies selected focus specifically on human-machine interaction and how human oversight has been planned for each tool. This section seeks to materialise the typologies presented in Binns and Veale's framework, giving the reader a snapshot of different cases and how each one presents challenges to human discretion.

Udbetaling Danmark (UDK)

Social welfare error and fraud detection

Context

Denmark has been recognised as a front runner in the digital transformation of the public sector, including the current trend of digitalising the welfare state and automating decisions relating to access to benefits. An e-government strategy set out in 2011 envisioned mandatory digitalisation to better serve citizens and businesses (Deloitte 2020). Since 2015, Denmark's digital self-service has required citizens to apply for public services and benefits online. In parallel to the digital transformation plan, the government established UDK to centralise payments previously carried out by different municipalities and help ease their digital transitions (Østergaard Madsen et al 2022). UDK took on the administration of housing, family disability and maternity leave payments. The reorganisation of the services cut administrative costs and transitioned in-person communication to a system of over-the-phone and digital points of contact. The UDK received 1,500 of the 2,000 administrators from across the different municipalities who were transferred to the organisation. The remaining group of 500 personnel handled the cases for those who were not ready to make the digital transition. UDK was not only in charge of automating payments, but also implemented *Den Fælles Dataenhed (DFD)* (Data Mining Unit) to detect error and fraud in the system by cross referencing data and analysis (Ibid.). The ultimate goal of the DFD is to detect fraud and error at the earliest stage possible. According to UDK, this is to avoid situations in which beneficiaries are required to pay back benefits that were incorrectly allotted (Deloitte and the Lisbon 2020).

How does it work?

The focus in this case study, is UDK's data-driven application of error and fraud detection. Before the implementation of the DFD, detection of potential fraud was done on the basis of citizen tips or the experience of investigators who were responsible for detecting potential fraud cases. Since 2015, the DFD has focused its efforts on learning from these cases and detecting irregularities to prevent the system from granting benefits to those who are not entitled to them. However, using machine learning techniques, the unit first focused on targeted inspection and detecting irregularities in existing benefit payments in order to develop an accurate system under the unit's mandate. The system's ultimate goal that is at an incipient stage, is for the algorithm to prevent error and fraud, by detecting irregularities in new applications and catching complex cases that are not usually detected by caseworkers.

What type of human oversight does it entail?

The complex task of fraud detection can be split between simple 'routine' and more complex cases (Østergaard Madsen et al 2022). In the first category, algorithms are trained to look out for usual cases of fraud. The system highlights a case, with a caseworker deciding after, in the downstream, whether it is fraudulent or not. In an example posed by the DFD, a simple case could consist of using data to detect whether a beneficiary is eligible for single parent child benefits. The system in this case, may point out "suspicious cases" by gathering data on those who receive the benefit, looking at the partner's address, the size of the residence, etc. Here a caseworker is required to investigate further. In the second category, for more complex cases, machine learning techniques are used to detect data outliers, and do so by cross referencing data, from other public registries (national income registry, health contact data, labour market and recruitment) to reduce the number of false positives and ultimately to detect new cases of fraud that have been traditionally overlooked by caseworkers. For example, by cross referencing health and residence data, the system may identify fraud with sick leaves. However, when and how caseworkers are included in the decision-making process is not clear. Designed without human supervision in mind, triage systems can produce solely automated decisions. In the UDK case, there are cases that go unsupervised as human overseers do not monitor all cases that are processed (Eiriksson Arent 2019).

Discussion

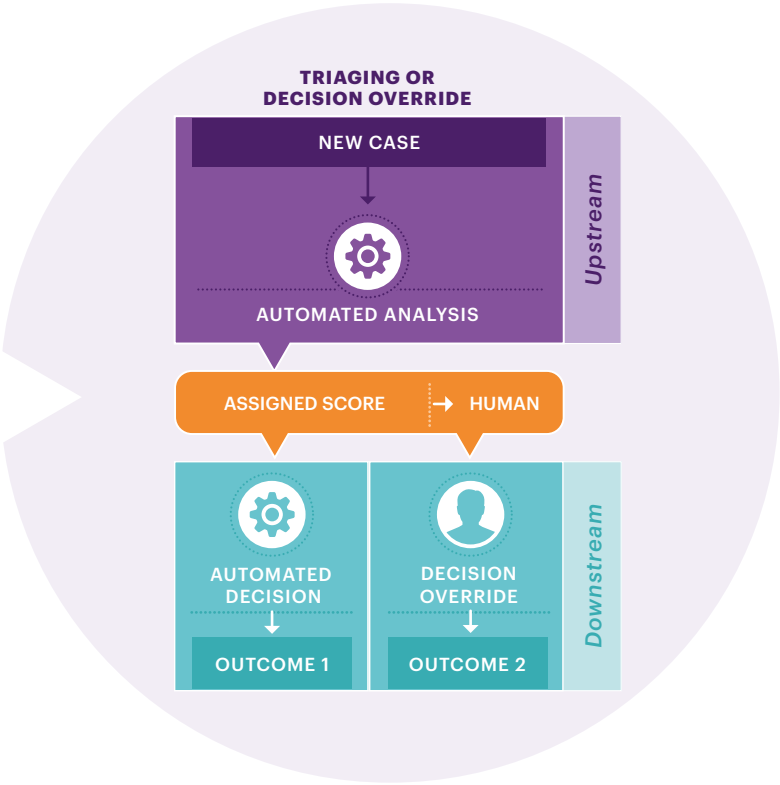
According to Denmark's National Audit Office, establishing UDK has helped to achieve 40 million EUR of annual savings due to the reduction of full-time personnel (Østergaard Madsen et al 2022). Digital welfare programs have often been criticised for putting more weight on the efficiency gains of automation, rather than the broad range of potential harms these systems can have, from infringement of privacy to misplaced accusations of fraud.

There are more questions than answers, however, when discussing whether UDK's implementation of human oversight effectively minimises harms. With regards to the allocation of benefits, the organisation has been called out for administrative errors. In 2019 UDK provided the Danish Tax Authority with erroneous information, sending emails to 111,000 households requiring them to repay taxes back to the authorities (Kayser Bril 2020).

Figure 4. **Human Oversight typology for Udbetaling Danmark**

Image source: Digital Future Society.

In this case, we consider that the welfare distribution ADMS may sometimes present supervision in the downstream, as municipalities receive a priority list for further examination. For other cases, all the data is currently processed automatically, and the decision is also automated. Nevertheless, in flagged cases, caseworkers use the automated decision to inform their decisions on suspicious or mistaken cases.



Opaque data handling

In 2019 the Danish Data Protection Authority (DPA), called out UDK for collecting data on relatives of beneficiaries, declaring it an infringement of GDPR. This happened despite UDK claiming the action was justified as the ultimate goal of the data collection was to determine welfare fraud. That same year, Denmark established the Data Ethics Council to investigate the ethics behind data sharing in the public sector. The case was reopened in 2020 as the DPA questioned the scope of the data collection. UDK committed to remove all improperly collected data (Eiriksson Arent 2019).

UDK leaves citizens in the dark with regards to how their data will be used, the algorithms that are working and how the system classifies them. According to Danish think tank Justitia, this includes not only the applicants but also their cohabitants, spouses, or other household members. Furthermore, although there are doubts and criticisms relating to the scope of UDK's data collection and how it handles said data, Justitia states that it is not possible to make an adequate assessment of the organisation, including whether its monitoring of citizens meets legal requirements (Eiriksson Arent 2019).

Caseworker's experience called into question

Caseworkers are involved at different points during the detection of fraud and error. Once a citizen is flagged as a possible case of fraud or error the information is sent to the respective municipalities so they can filter and analyse the case manually.

However, the Justitia report hints that it might be intentional that there is no human oversight during the previous stage as UDK's methods for detecting fraud imply a broad scope of data collection meaning human oversight during this stage of automation would violate citizens' privacy. It is difficult to judge whether humans are kept out of the loop to solve a bad praxis, but it is worth considering (Eiriksson Arent 2019).

Frontex

Automated Border Control (ABC)

Context

Frontex, also known as the European Border and Coast Guard Agency, has flagged increasing passenger traffic at international borders, and further expected rises to come in the near future, as an unprecedented challenge. According to the agency, an EU border guard has on average 12 seconds to evaluate the traveller in front of them. Furthermore, due to ongoing administrative and political challenges such as the ongoing migration crisis or the longstanding presence of terrorism threats, ensuring free movement between countries, both inside and outside of the Schengen area, is one of the European Union's most pressing issues. Therefore, the limited amount of time border guards have, increasing traffic at international checkpoints and the high stakes at play are driving the implementation of border gate technology and leading to the adoption of Automated Border Control (ABC) (Fergusson 2014).

Following the institution's rationale, many Schengen countries have installed smart gates to ease passenger traffic and improve the performance of border security. The first country to do so was Portugal in 2008 and although EU border controls employ a low level of automation which depends heavily on operators, demand is increasing. As of 2019, records showed ABC gates to be operating at more than 50 airports (Noori 2022).

Reflecting the interest in the automation of border control, the EU Commission initiated a pilot project called the ABC4EU (Automated Border Control for Europe) that ran from 2014 to 2016 with the objective of harmonising ABC gates processing third-country nationals entering the EU. Harmonisation consisted of, in general terms, updating the current ABC gate systems to make them more flexible and to encourage use. The pilot also sought to assess the impact of ABC gates, evaluating the automation process and identifying potential obstacles (European Commission 2022).

How does it work?

At present, ABC consists of semi-automated electronic gates — including document readers, two physical barriers, and biometric scanners. The upstream is automated, as travellers with an electronic passport (ePassport) can pass through an ABC portal, where e-gates scan documents and perform database queries. Once the traveller's face is scanned, the system compares their facial image with biometric data stored in the ePassport. These previously manual steps are intended to relieve border guards from repetitive tasks and direct their attention to screening and interrogating travellers flagged by the system — human oversight is focused on the downstream.

What type of human oversight does it entail?

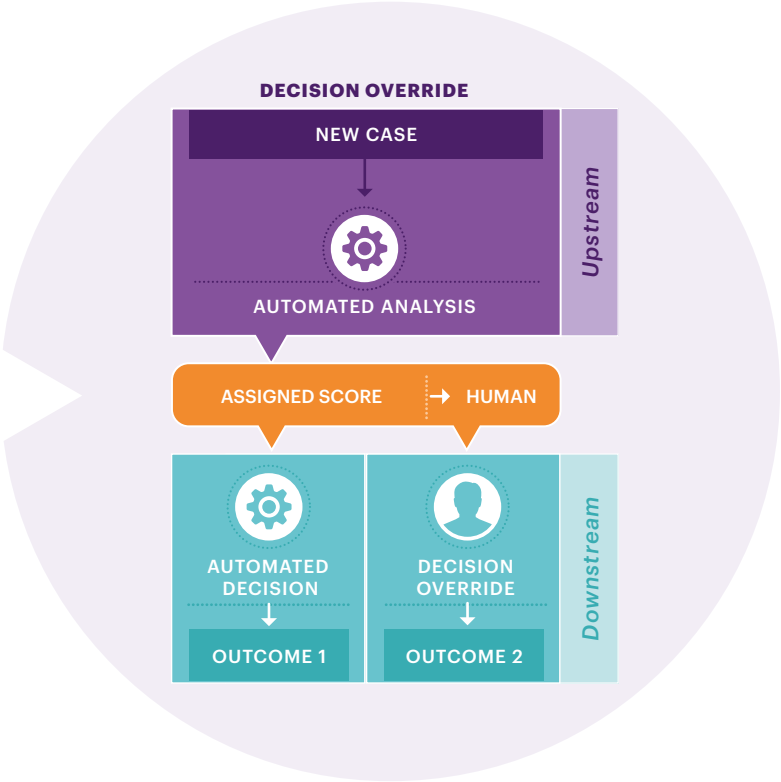
Considering the previously mentioned typologies of human oversight, the ABC system exhibits a decision override system that sees operators only taking action following a system breakdown or anomaly detection. The BODEGA project, an EU-funded research project, interviewed border guards to explore the impact of ABC on their work. During the study,

border guards expressed their mistrust of the system's abilities to replace them. Due to the automation, the border guards' role changed from an active role, controlling travellers, to a passive one, supervising the automated check (e.g., restarting the system when hardware problems arise or manually checking documents when particular cases are not correctly detected).

Figure 5. **Human Oversight typology for ABC**

Image source: Digital Future Society.

The process of automation is focused on the upstream (by scanning passport and passenger faces to match them to the databases). For the downstream, those cases that the system flags the passenger as suspicious or detected by the human supervisor will be treated manually. Otherwise, an automated process will make the decision.



Discussion

As mentioned earlier, the pressure for ABC to partially replace border guards stems from an increase in travellers, with implementation supposedly freeing the guards to deal with more complex tasks. However, automation is also seen as a way to make border control less biased and error prone. Consequently, experts argue that smart borders place a high level of mistrust on border guards, questioning their competence and capacity to verify identity, effectively turning them into security "problems" (Noori 2020). Often, however, the guards see it differently, not trusting the automated system to do their job properly. Therefore, implementation of ABC gates has created a general mistrust of the automated system and, in the process, the role of the border guards themselves. Having to deal with system malfunctions caused guards to mistrust ABC further, not believing it to be capable of replacing their role. The attitudes of the guards, which results in both mistrust of the system and under-reliance, hampers the overall effectiveness of automated border control (Noori 2020). In this scenario, oversight is meaningful only when an error is detected, and a manual process is taken instead. Given that errors are hard to understand and address, many blind spots regarding border policy and security remain.

Furthermore, Frontex foresees that future ABC gates will be designed for operation to be intuitive and to require border guards to have little technical knowledge (European Border and Coast Guard Agency 2021). Currently, border guards that have dealt with e-gates do not have access to specific knowledge about how the system works, and more specifically, what leads to system failures. Therefore, one risk is algorithm aversion, causing border guards to inadequately address system errors on their own — a certain level of training or personnel with a good understanding of AI techniques is required to mitigate this risk (European Border and Coast Guard Agency 2021).

RisCanvi

Criminal risk assessment

Context

A risk assessment instrument (RAI) is a type of algorithmic tool that aims to predict a defendant's future risk for misconduct, commonly used for pre-trial judicial decisions. Assessments were based on professional judgment before the use of RAIs in criminal justice systems became commonplace in the 1970s. As RAIs provide structured, evidence-based predictions they were introduced to reduce discretion and increase objectivity. Nonetheless, the promised objectivity of these assessments is still open to debate (Heilbrun et al. 1999). Recent innovations like the incorporation of computer-based algorithms over the last decade have further increased these concerns due to different studies arguing that an algorithm-based RAI can outperform human prediction (Tan et al. 2018; Green and Chen 2019b). Not only are there concerns about whether these machines are as accurate and fair as human operators, they are also controversial as these types of tool have exhibited biases against race and gender, as in the case of the UK's Offender Assessment System (OASys) (Angwin et al. 2016). OASys, comparable to RisCanvi, was found to generate different predictions for race, gender and age (Big Brother Watch 2020).

RisCanvi (named from the Catalan words for risk and change) is a RAI used in Catalonia, Spain. Created in 2009 to help criminologists and social workers improve the treatment of inmates, the tool is based on several instances of clinical analysis. The Department of Justice in Catalonia commissioned *Grup d'Estudis Avançats en Violència* (Group of Advanced Studies of Violence) of *Universitat de Barcelona* (Barcelona University) to create RisCanvi.

How does it work?

RisCanvi consists of 43 professionally assessed risk factors, based on an inmate's record and personal interviews. These risk factors can be related to the inmate's attitude and personality, and personal and clinical history as well as their response to treatments. Such factors include drug and alcohol abuse, history of mental illness and being a victim of violence. The algorithm uses these factors to assess the risk of five different outcomes: 1. self-directed violence, 2. violence directed towards other inmates or staff, 3. recidivism 4. violent recidivism 5. breaching parole. A team of multidisciplinary professionals collects data regarding each factor alongside clinical history, observations, and interviews. They then input the information into

the protocol which assigns the inmate a risk score. The algorithm's outcome is only a three-level classification meaning risk can be low, medium or high.

The final evaluation is then assessed by the team to determine the type of treatment the inmate receives. Inmates are evaluated at least every six months. The predicted levels are also used in reports sent to prosecutors and judges to consider a conditional release.

What type of human oversight does it entail?

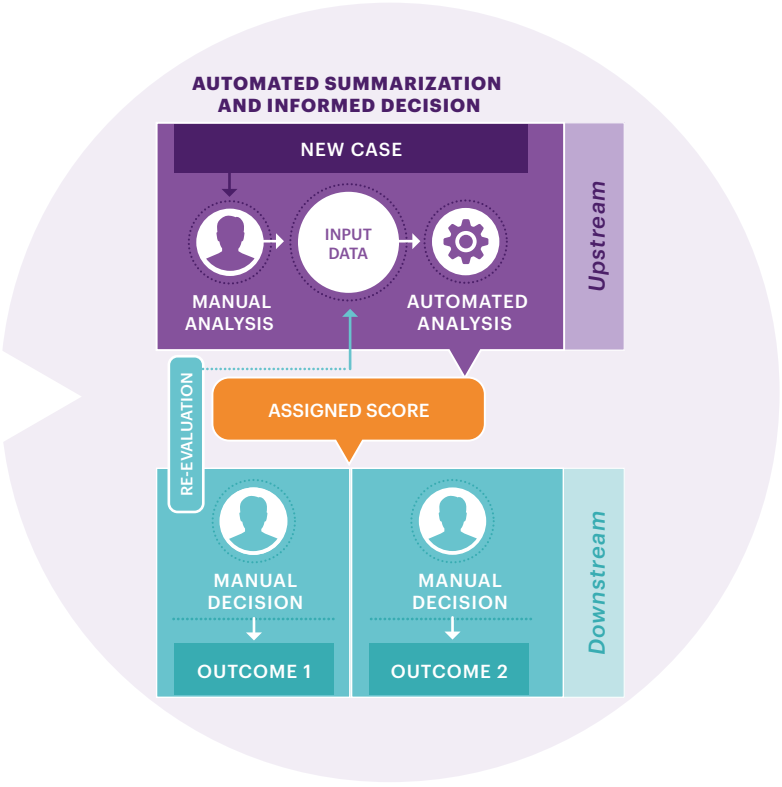
This tool incorporates human oversight at two points of the protocol, both in the upstream and downstream stages. First, caseworkers collect and process evidence to provide input data for the algorithm. Then a senior professional supervises a discussion of each factor between different professionals and either validates or adjusts the outcome of their discussion.

The algorithm's output and the professional recommendations then have two outcomes. The report is sent to prosecutors and judges to inform them of the type of sentence to deliver. At the same time, social workers and psychologists use this tool to follow-up on inmates' correctional treatments and re-evaluate their status. This means the decision, is not in fact a decision, and therefore not fully automated, as professionals interact with a form of summarised information before making an informed decision.

Figure 6. **Human Oversight typology for Riscanvi**

Image source: Digital Future Society.

In the Upstream, data is introduced manually in a first assessment made by caseworkers. Afterwards, an automated analysis assigns a score. This score and other additional information are used by two types of users with different goals, caseworkers/ social workers and prosecutors/ judges. Caseworkers can re-evaluate inmates and override the score if they consider it necessary. In both cases, the final decision is always subject to the discretion of a human.



Discussion

RisCanvi has evolved significantly since its inception, as it was initially meant to be a guide for prison management rather than for criminal sentencing. However, at present it is considered for both decision-making processes, which has led many to question the impact the algorithm has on actual decision-making. As mentioned previously, RAIs are supposed to facilitate debate on the impact of the risk level, however it has been found that judges and prosecutors often overlook report recommendations, basing their decision solely on the algorithm's recommendations (Saura and Aragó 2021).

Rubber-stamping was not considered a risk when the algorithm was created, as the designers expected in-depth analysis of cases to take place. The teams who input and analyse the data are trained to use RisCanvi, therefore they should have enough experience using the tool to know what they can expect from the algorithm. This knowledge could play in favour of overriding predictions in case of unexpected outputs. Furthermore, caseworkers should use the tool to improve their decisions by contrasting their conclusions with the algorithm's outcome. Despite this, some have questioned the low rate at which these multidisciplinary teams are adjusting the algorithm's findings.

Furthermore, there have been few external audits of the system, which causes concerns for transparency and potential bias against some underrepresented populations (Planas Bou 2021). Bias can arise from input data which is primarily based on caseworker interviews with inmates. If this initial information is biased, it can create a feedback effect in which the system's biased data will be used in future algorithm updates, further reinforcing such biases. This has been the case for RAIs like COMPAS (see OASys case cited above) leaving entire social groups vulnerable to algorithmic discrimination because of historical data.

One positive lesson is that interpersonal interaction between professionals is critical for successful use of the system. Even though judges may omit caseworkers' reports favouring the algorithm, caseworkers perceive it positive practice to translate their professional perspective to other professionals. Having this in mind, trust plays a key role in improving the human-algorithm collaboration, allowing for a more structured exchange of information and parity between different roles.

Key takeaways

The previous section introduced three case studies that illustrate the broad depth of situations where ADMS are present and most importantly, highlights the complexity that lies beyond the automation of processes.

On the whole, the case studies have shown how, in theory, using an algorithm for regular tasks could free human operators from intense mental tasks — tasks that humans tend to have a hard time performing — and enable them to further investigate cases that require more attention. At the same time, automated systems could augment human ability to identify and correct mistakes, avoiding automation bias (overreliance) or algorithm aversion (under-reliance) (De-Arteaga et al. 2020).

Suppose all algorithms were explainable and transparent. In that case, humans could consider and interpret the information provided and act accordingly, yet this is rarely the case. Risk communication literature makes it clear that contextual information and expertise are essential for making informed decisions (Heilburn 1999). However, many unexpected situations can arise for which the algorithm has not been trained, causing failure, and humans should be ready to act upon these situations.

One solution might be a collaborative process between individuals that could improve their perception and increase agreement with algorithms, like in the interaction between professionals presented in the RisCanvi case (Van Berkel et al. 2019). However, collaborative processes require more time and resources and could reduce the efficiency promised by ADMS. Nevertheless, scenarios where consequences represent high risks for society and time is not a determinant variable (like in the ABC case) could be considered. Collaborative approaches are an often-overlooked opportunity in other cases that could help reduce bias and errors.

There is still an open debate on whether automation alters human liability for a final decision. Much of the regulations and assumptions stem from the idea that human oversight would bring accountable results (Wagner 2019). The ABC and UDK cases demonstrate that most of the systems are not designed with human liability in mind (e.g., when humans can only mitigate the errors, when they are only responsible for the final decision or to provide the algorithm with new information), taking away discretionary power from the human overseer. The relationship between data protection and data privacy concerns relating to the algorithm training process in the UDK case shows that there are still many flaws in how an oversight process can be put in place.

Furthermore, when an algorithmic system is designed for oversight, ADMS can be used in unintended ways. As seen in the RisCanvi case, there are different system 'users' — judges, for example — that use the automated result as part of their own decision-making processes, which can be more complex to analyse. Such systems should be articulated with the different contexts mentioned in Figure 1., including the human factors, the organisational values and practices, and the societal and cultural assumptions. A system should also consider different degrees of liability in mind, both for the system, the human operator and the institution responsible for its implementation. They should be able to act accordingly to their particular level of liability.

Another trend identified in the case studies is when staff are treated as weak links in the system, or their agency is diminished. In the ABC case, treating the staff this way put the entire system at risk because they had no agency to act. In the UDK case, staff were excluded from the loop and used as a secondary safeguard after the harm was done. If an operator acts with sufficient information, training and experience it would not only mitigate the algorithm's disparate impact but also increase trust and enhance the experience of citizens (Chouldechova 2017).

High-risk AI systems should contemplate the role of users (or supervisors), which should be defined alongside algorithm capabilities and with the required expertise in mind. Looking beyond human oversight is to consider humans with an active role in the complexity of the system.

7. Policy Recommendations

In most situations algorithms are typically deployed to aid human decision-makers rather than act autonomously (Green and Chen 2020). Developing human oversight as a solution for algorithmic bias and disparate harm should be considered carefully.

Many government decisions require balancing accurate predictions against other social goals (e.g., equitable distribution of resources, fair treatment of citizens, etc.). The following general policy recommendations address the multiple trade-offs and complexities explained in this policy brief, but some recommendations will be more or less true depending on the particular case/context.

Define the minimum human involvement

The first recommendation for implementing algorithms with human oversight is to consider how the supervision will be meaningful towards error mitigation. The GDPR and AI Act highlight this characteristic, but it is usually ambiguous whether a contribution is meaningful or not (Green 2021). Human involvement can be superficial or can provide a false sense of security. An important task is to define the system and analyse trade-offs relating to the opportunities and challenges that come with having human oversight in the system. For example, the design of the ABC did not involve the guards in the implementation of the system and garnered mistrust in the border guards, offering only a false sense of security.

An implementation that does not use valuable human resources is inefficient and offers a way for flaws to enter the process. Automated systems should not be adopted solely to justify staff reductions or facilitate the use of untrained staff. This could harm system operation and oversight and lead to the rubber-stamping of algorithm decisions. Human operators should have the ability to contest and mitigate potential threats as well as being able to agree with the system. Understanding these valuable actions would enhance human performance (Almada 2019). In a double-loop decision-making model, an “action is evaluated in terms of the degree it helps participants generate valid and useful information” (Argyris 1976 p. 368).

Beware of automation context-dependency

Following the AI Act, it is context dependant whether decisions can be easily automated, or whether automation poses a high risk and requires safety measures such as human oversight. But interactions with algorithms and data structures do not occur in isolation (Seaver 2019). Like any technological system, an algorithmic system is interconnected with other objects, processes and individuals (Jung et al. 2008). A correct evaluation of the context, the infrastructure and the people involved should be made for all implementations including algorithms mediating or defining decision-making, as shown in Figure 1. However, a prior evaluation and an impact assessment are not enough.

To analyse the affects these algorithms can have, a risk management plan and a quality assurance process should also be prescribed. In addition, any test and evaluation should take place both in-lab (without the possibility of external risks) and in-context (where system errors and behaviour failures can be spotted). The results of both evaluations could lead to algorithms being held from deployment, the typology of human oversight changed (whether to a higher degree of involvement or in a different moment of intervention), or a judgement that the interaction between both is not possible and that the entire system should be modified. A poor context analysis could increase system risk and fail to anticipate errors that could appear following deployment of the system.

Choose open over closed systems

When it comes to software, closed systems represent rigorously protected intellectual property and do not allow for any open exchange or cooperative development of code. This is in contrast to open-source software that allows external developers to revise code and suggest changes that will improve the software. Many of the issues described in this brief stem from algorithms being part of closed systems. Like in the ABC case, users cannot see how the algorithms work or how they relate to the other technologies they are using to accomplish their objectives. As the systems are interconnected, the algorithm-based solution should also be transparent to ensure a better interaction.

An open system can be tested and explained, developers and users can correctly identify errors. Users being able to identify system errors could promote mitigation strategies and alert them to unexpected situations. This would increase reliability and trust. Open systems also enable more opportunities for proper care for and maintenance of technologies to take place, reducing friction and associated costs. On the contrary, a closed system is not explainable, cannot be tested and adopted; it prevents users from highlighting potential problems and leads to algorithm aversion and under-reliance.

Define a governance scheme and degrees of liability

Any system should have a governance scheme to define the problem it will address and manage how decisions will be made. With current trends requiring public registers for all ADMS deployed in the public sector, these should also include data, code and interfaces that different communities can access and scrutinise. This would increase public trust and agreement on how to treat disparate impacts in particular case assessments (Van Berkel et al. 2019). Given that ADMS is deployed by public services looking to increase efficiency and objectivity, it is desirable to pay attention to how liability will be shared. As mentioned earlier, a transfer of responsibility to technology providers has taken place and reintroducing human oversight without defining liability would pose a problem. The case studies also showed that algorithms can appear to have higher authority than the people supervising them, opening the door to rubber-stamping which could allow users to avoid having to take responsibility for their decisions. The effect of a moral buffer — distancing from the decision — could be problematic for the accountability goals public institutions have. Instead, a transparent process facilitating the identification and sharing of standard practices, including scrutiny by civil society would enable human oversight to be effectively introduced with sufficient authority and support.

Train and promote knowledge sharing among developers and operators

Regulation should require appropriate training and resources be provided to support the staff who operate and oversee AI systems. In addition, previous experience and acquired knowledge should be documented correctly, which will help distribute good practices and mitigation strategies, and also strengthen confidence and reliability between human operators. Furthermore, sharing what developers know about the system and what the operators know about the task will help to catalogue errors and bugs, enabling the system and algorithm to be continuously updated. Denying access to knowledge promotes a less dynamic organisation, causing more delays in problem-solving and miscommunication between teams.

Define a whistle-blower procedure

A human-machine system can be perfectly reliable, but there is always a chance to cause harm, which those involved could miss. In some cases, institutional complaints about failures, bad practices, or biased decisions can put the work and life of human supervisors at risk. Given that, in many cases, algorithms are considered more objective and authoritative compared to the discrete choices made by humans, supervisors can feel vulnerable making a decision that goes against an algorithm. It is crucial to develop whistle-blower mechanisms and other anti-retaliation safeguards that protect workers when they override an algorithm's decision, challenge any automated decision or denounce a system failure. Otherwise, they might choose to ignore algorithm decisions to avoid problems with the authorities and safeguard their jobs.

8. Conclusion

This policy brief has explored and highlighted the complexities that come with the implementation of human oversight, to facilitate government use of ADMS in different scenarios. Unfortunately, consideration of human oversight in current and proposed regulation is not sufficient. The policy recommendations outlined in this brief shed some light on the existing trade-offs, opportunities and challenges that relate to the issue.

Public administrations and governments should consider the level of human oversight to implement for high-risk AI systems according to the specific context and the capabilities of systems and staff. At the same time, operators should be trained to understand the trade-offs that exist in using such systems and be offered the opportunity to learn about and understand how the systems will work as well as having active roles in their design process.

The three cases presented represent algorithm systems with complex organisational frameworks. The precise operations and potential ramifications of the algorithm's impact could be obscure, requiring sufficient mechanisms of transparency and explanation. Furthermore, these systems are hard to understand even for experts and professionals. For that reason, a governance scheme should open the door to scrutiny of the system and offer the possibility to confidently denounce any error or harm caused while these systems are being used.

Implementing human oversight in ADMS beyond the objective of simple mitigation can bring many benefits to society. Humans can contribute to safer and more compliant systems, while computational capacities and automation have the potential to greatly enrich society. Human oversight will not come easily, and it must be implemented with significant care. Actions such as considering and following the recommendations laid out in the policy brief, need to be taken to avoid all potential harms.

References

Almada, M. (2019). Human intervention in automated decision-making: Toward the construction of contestable systems. Proceedings of the 17th International Conference on Artificial Intelligence and Law, ICAIL 2019, pp. 2–11. [online] Available at: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3264189 (Accessed: May 24, 2022)

Ananny, M. and Crawford, K. (2018). Seeing without knowing: Limitations of the transparency ideal and its application to algorithmic accountability. New Media and Society, 20(3), pp. 973–989. [online] Available at: https://journals.sagepub.com/doi/10.1177/1461444816676645 (Accessed: May 24, 2022)

Angwin, J. et al. (2016). Machine bias: There's software used across the country to predict future criminals. And it's biased against blacks. Available at: [online] Available at: https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing (Accessed: May 24, 2022)

Argyris, C. (1976). Single-loop and Double-Loop Models in Research on Decision Making. Administrative Science Quarterly, 21, pp. 363–377.

Balfe, N., Sharples, S. and Wilson, J.R. (2018). Understanding Is Key: An Analysis of Factors Pertaining to Trust in a Real-World Automation System. Human Factors, 60(4), pp. 477–495. [online] Available at: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5958411/ (Accessed: May 24, 2022)

Barocas, S. (2014). Big Data's Disparate Impact. California Law Review, 104(671), pp. 671–732. [PDF] Available at: https://www.californialawreview.org/wp-content/uploads/2016/06/2Barocas-Selbst.pdf (Accessed: May 24, 2022)

Barocas, S., Hardt, M. and Narayanan, A. (2019). Fairness and Machine Learning: Limitations and Opportunities. Fairness and Machine Learning: Limitation and Oppotunities [Preprint]. Available at: https://fairmlbook.org (Accessed: May 24, 2022)

Batastini, A.B. et al. (2019). Does the Format of the Message Affect What Is Heard? A Two-Part Study on the Communication of Violence Risk Assessment Data. Journal of Forensic Psychology Research and Practice, 19(1), pp. 44–71. [online] Available at: https://www.tandfonline.com/doi/abs/10.1080/24732850.2018.1538474 (Accessed: May 24, 2022)

Van Berkel, N. *et al*. (2019). Crowdsourcing perceptions of fair predictors for machine learning: A recidivism case study. Proceedings of the ACM on Human-Computer Interaction, 3(CSCW). [online] Available at: https://dl.acm.org/doi/abs/10.1145/3359130 (Accessed: May 24, 2022)

Big Brother Watch. (2020). Big Brother Watch briefing on Algorithmic Decision-Making in the Criminal Justice System. [PDF] Available at: https://bigbrotherwatch.org.uk/wp-content/uploads/2020/02/Big-Brother-Watch-Briefing-on-Algorithmic-Decision-Making-in-the-Criminal-Justice-System-February-2020.pdf (Accessed: 22-9-2022)

Binns, R. and Veale, M. (2021). Is That Your Final Decision? Multi-Stage Profiling, Selective Effects, and Article 22 of the GDPR. International Data Privacy Law, 11(4), pp. 319–332. [online] Available at: https://academic.oup.com/idpl/article/11/4/319/6403925 (Accessed: May 24, 2022)

Birhane, A. (2021). The impossibility of automating ambiguity. Artificial Life, 27(1), pp. 44–61. [online] Available at: https://direct.mit.edu/artl/article-abstract/27/1/44/101872/The-Impossibility-of-Automating-Ambiguity (Accessed: May 24, 2022)

Brkan, M. (2017). AI-supported decision-making under the general data protection regulation. Proceedings of the International Conference on Artificial Intelligence and Law, pp. 3–8. [online] Available at: https://dl.acm.org/doi/10.1145/3086512.3086513 (Accessed: May 24, 2022)

Burton, J.W., Stein, M.K. and Jensen, T.B. (2020). A systematic review of algorithm aversion in augmented decision making. Journal of Behavioral Decision Making, 33(2), pp. 220–239. [online] Available at: https://onlinelibrary.wiley.com/doi/abs/10.1002/bdm.2155 (Accessed: May 24, 2022)

Campolo, A. and Crawford, K. (2020). Enchanted Determinism: Power without Responsibility in Artificial Intelligence. Engaging Science, Technology, and Society, 6, p. 1. [online] Available at: https://www.researchgate.net/publication/338486570_Enchanted_Determinism_Power_without_Responsibility_in_Artificial_Intelligence (Accessed: May 24, 2022)

Chouldechova, A. (2017). Fair Prediction with Disparate Impact: A Study of Bias in Recidivism Prediction Instruments. Big Data, 5(2), pp. 153–163. [PDF] Available at: https://arxiv.org/pdf/1610.07524.pdf (Accessed: May 24, 2022)

Cranor, L.F. (2008). A framework for reasoning about the human in the loop. in Proceedings of the 1st Conference on Usability, Psychology, and Security. San Francisco, California: USENIX Association, p. 15. [online] Available at: https://dl.acm.org/doi/10.5555/1387649.1387650 (Accessed: May 24, 2022)

Cummings, M. L. (2006). Automation and Accountability in Decision Support System Interface Design. The Journal of Technology Studies, 32(1). [PDF] Available at: https://scholar.lib.vt.edu/ejournals/JOTS/v32/v32n1/pdf/cummings.pdf (Accessed: May 24, 2022)

De-Arteaga, M., Fogliato, R. and Chouldechova, A. (2020). A Case for Humans-in-the-Loop: Decisions in the Presence of Erroneous Algorithmic Scores. in Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems. New York, NY, USA: ACM, pp. 1–12. [PDF] Available at: https://arxiv.org/pdf/2002.08035.pdf (Accessed: May 24, 2022)

Deloitte and Lisbon Council (2020). Study on public sector data strategies, policies, and governance. European Commission. [PDF] Available at: https://joinup.ec.europa.eu/sites/default/files/custom-page/attachment/2020-06/DIGIT%20-%20D01%20-%20Study%20on%20public%20sector%20data%20strategies%2C%20policies%20and%20governance%20v3annexes.pdf (Accessed: May 24, 2022)

Dietvorst, B.J., Simmons, J.P. and Massey, C. (2014). Algorithm Aversion: People Erroneously Avoid Algorithms after Seeing Them Err. SSRN Electronic Journal, 143(6), pp. 1–13. [online] Available at: https://repository.upenn.edu/cgi/viewcontent.cgi?article=1392&context=fnce_papers (Accessed: May 24, 2022)

Digital Future Society (2020) Exploring gender-responsive designs in digital welfare. [online] Available at: https://digitalfuturesociety.com/report/exploring-gender-responsive-designs-in-digital-welfare/ (Accessed: June 22, 2022)

Digital Future Society (2021). Governing algorithms: perils and powers of AI in the public sector, p. 36. [online] Available at: https://digitalfuturesociety.com/report/governing-algorithms/ (Accessed: May 24, 2022)

Dourish, P. (2001). Where the Action Is: The Foundations of Embodied Interaction. Where the action is the foundations of embodied interaction. Cambridge, MA: MIT Press. [online] Available at: https://direct.mit.edu/books/book/3875/Where-the-Action-IsThe-Foundations-of-Embodied (Accessed: May 24, 2022)

Eiriksson Arent, B. (2019). Analyse : Udbetaling Danmarks Systematiske. Justitia og forfatteren. [PDF] Available at: https://justitia-int.org/analyse-udbetaling-danmarks-systematiske-overvaagning/ (Accessed: May 24, 2022)

European Border and Coast Guard Agency (2021). Artificial Intelligence-Based Capabilities for the European Border and Coast Guard. [PDF] Available at: https://frontex.europa.eu/assets/Publications/Research/Frontex_AI_Research_Study_2020_final_report.pdf (Accessed: May 24, 2022)

European Commission (2020). White Paper On Artificial Intelligence - A European approach to excellence and trust EN. [PDF] Available at: https://ec.europa.eu/info/sites/default/files/commission-white-paper-artificial-intelligence-feb2020_en.pdf (Accessed: May 24, 2022)

European Commission (2021). Proposal for a regulation of the european parliament and of the council laying down harmonised rules on artificial intelligence (artificial intelligence act) and amending certain union legislative acts com/2021/206 final. European Commission, 0106, pp. 1–108. [online] Available at: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A52021PC0206 (Accessed: May 24, 2022)

European Commission (2022). ABC Gates for Europe. Cordis EU research results. [online] Available at: https://cordis.europa.eu/project/id/312797 (Accessed: September 22, 2022)

European Parliament and Council of the European Union (2016). EU general data protection regulation. [online] Available at: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A32016R0679 (Accessed: May 24, 2022)

Fergusson, J. (2014). 12 Seconds to decide. In search of excellence: Frontex and the principle of best practice. [PDF] Available at: https://frontex.europa.eu/assets/Publications/General/12_seconds_to_decide.pdf (Accessed: May 24, 2022)

Goodman, B.W. (2016). Economic Models of (Algorithmic) Discrimination. 29th Conference on Neural Information Processing Systems [Preprint], (Nips). Available at: [PDF] Available at: http://www.mlandthelaw.org/papers/goodman2.pdf (Accessed: May 24, 2022)

Green, B. (2020). The false promise of risk assessments: Epistemic reform and the limits of fairness. FAT* 2020 - Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, pp. 594–606. [online] Available at: https://dl.acm.org/doi/pdf/10.1145/3351095.3372869 (Accessed: May 24, 2022)

Green, B. (2021). The Flaws of Policies Requiring Human Oversight of Government Algorithms. SSRN Electronic Journal, pp. 1–42. doi:10.2139/ssrn.3921216. [PDF] Available at: https://arxiv.org/ftp/arxiv/papers/2109/2109.05067.pdf#:~:text=4.1%20Flaw%201%3A%20Human%20Oversight,provide%20reliable%20oversight%20of%20algorithms (Accessed: May 24, 2022)

Green, B. and Chen, Y. (2019a). Disparate Interactions. in Proceedings of the Conference on Fairness, Accountability, and Transparency. New York, NY, USA: ACM, pp. 90–99. [online] Available at: https://dl.acm.org/doi/proceedings/10.1145/3287560 (Accessed: May 24, 2022)

Green, B. and Chen, Y. (2019b). The principles and limits of algorithm-in-the-loop decision making. Proceedings of the ACM on Human-Computer Interaction, 3(CSCW). [online] Available at: https://dl.acm.org/doi/10.1145/3287560.3287563 (Accessed: May 24, 2022)

Green, B. and Chen, Y. (2020). Algorithmic risk assessments can alter human decision-making processes in high-stakes government contexts. [online] Available at: https://scholar.harvard.edu/files/bgreen/files/19-cscw.pdf (Accessed: May 24, 2022)

Heilbrun, K. et al. (1999). Violence risk communication: Implications for research, policy, and practice. Health, Risk and Society, 1(1), pp. 91–105. [online] Available at: https://www.tandfonline.com/doi/abs/10.1080/13698579908407009?-journalCode=chrs20 (Accessed: May 24, 2022)

Jung, H. et al. (2008). Toward a framework for ecologies of artifacts: how are digital artifacts interconnected within a personal life? Proceedings of the 5th Nordic conference on Human-computer interaction: building bridges, pp. 201–210. [online] Available at: https://dl.acm.org/doi/10.1145/1463160.1463182 (Accessed: May 24, 2022)

Kahneman, D. (2011). Thinking, Fast and Slow. Farrar, Straus and Giroux. [online] Available at: https://books.google.es/books?id=ZuKTvERuPG8C (Accessed: May 24, 2022)

Kayser-Bril, N. (2020). In a quest to optimize welfare management, Denmark built a surveillance behemoth. [online] Available at: https://automatingsociety.algorithmwatch.org/report2020/denmark/denmark-story/ (Accessed: September 22, 2022)

Kemper, J. and Kolkman, D. (2019). Transparent to whom? No algorithmic accountability without a critical audience. Information Communication and Society, 22(14), pp. 2081–2096. [online] Available at: https://www.researchgate.net/publication/325827444_Transparent_to_whom_No_algorithmic_accountability_without_a_critical_audience (Accessed: May 24, 2022)

Kulju, M. et al. (2019). A framework for understanding human factors issues in border control automation. IFIP Advances in Information and Communication Technology. Springer International Publishing. [online] Available at: https://hal.archives-ouvertes.fr/hal-02264619/ (Accessed: May 24, 2022)

Lee, J.D. and See, K.A. (2004). Trust in Automation: Designing for Appropriate Reliance. Human Factors: The Journal of the Human Factors and Ergonomics Society, 46(1), pp. 50–80. [PDF] Available at: https://user.engineering.uiowa.edu/~csl/publications/pdf/leesee04.pdf (Accessed: May 24, 2022)

McCallum, K.E., Boccaccini, M.T. and Bryson, C.N. (2017). The Influence of Risk Assessment Instrument Scores on Evaluators' Risk Opinions and Sexual Offender Containment Recommendations. Criminal Justice and Behavior, 44(9), pp. 1213–1235. [online] Available at: https://journals.sagepub.com/doi/abs/10.1177/0093854817707232 (Accessed: May 24, 2022)

Miron, M. (2018). Interpretability in AI and its relation to fairness, transparency, reliability and trust. [online] Available at: https://ec.europa.eu/jrc/communities/en/community/humaint/article/interpretability-ai-and-its-relation-fairness-transparency-reliability-and (Accessed: 14 February 2022)

Misuraca, G. and Noordt, C. van (2020). Overview of the use and impact of AI in public services in the EU. Luxembourg: Publications Office of the European Union. [online] Available at: https://joinup.ec.europa.eu/collection/elise-european-location-interoperability-solutions-e-government/document/report-ai-watch-artificial-intelligence-public-services-overview-use-and-impact-ai-public-services (Accessed: May 24, 2022)

Myers-West, S., Whittaker, M. and Crawford, K. (2019). Discriminating systems. AI Now Institute. [PDF] Available at: https://ainowinstitute.org/discriminatingsystems.pdf (Accessed: May 24, 2022)

Noori, S. (2022). Suspicious Infrastructures: Automating Border Control and the Multiplication of Mistrust through Biometric E-Gates. Geopolitics, 00(00), pp. 1–23. [PDF] Available at: https://www.tandfonline.com/doi/pdf/10.1080/14650045.2021.1952183?needAccess=true (Accessed: May 24, 2022)

Østergaard Madsen, C. Lindgren, I., Melin, U. (2022). The accidental caseworker - How digital selfservice influences citizens' administrative burden. Government Information Quaterly. [PDF] Available at: https://www.sciencedirect.com/science/article/pii/S0740624X21000897 (Accessed: July 20, 2022)

Redden, J., Dencik, L. and Warne, H. (2020). Datafied child welfare services: unpacking politics, economics and power. Policy Studies, 41(5), pp. 507–526. [PDF] Available at: https://www.tandfonline.com/doi/pdf/10.1080/01442872.2020.1724928?needAccess=true (Accessed: May 24, 2022)

Planas Bou, C. (2021). *Catalunya usa un algoritmo para ayudar a decidir a qué presos concede la libertad condicional* (Catalonia uses an algorithm to help decide which prisoners receive parole). El Periodico. [online] Available at: https://www.elperiodico.com/es/sociedad/20211117/catalunya-algoritmo-decidir-presos-concede-12859785 (Accessed: July 12, 2022)

Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nature Machine Intelligence, 1(5), pp. 206–215. [PDF] Available at: https://arxiv.org/pdf/1811.10154.pdf (Accessed: May 24, 2022)

Saura, G. y Aragó, L. (2021) Un algoritmo impreciso condiciona la libertad de los presos. La Vanguardia. [online] Available at: https://www.lavanguardia.com/vida/20211206/7888727/algoritmo-sirve-denegar-permisos-presos-pese-fallos.html (Accessed: September 22, 2022)

Scurich, N., Monahan, J. and John, R.S. (2012). Innumeracy and unpacking: Bridging the nomothetic/idiographic divide in violence risk assessment. Law and Human Behavior, 36(6), pp. 548–554. [online] Available at: https://www.researchgate.net/publication/224869265_Innumeracy_and_Unpacking_Bridging_the_NomotheticIdiographic_Divide_in_Violence_Risk_Assessment (Accessed: May 24, 2022)

Seaver, N. (2019). Knowing Algorithms. digitalSTS, pp. 412–422. [PDF] Available at: https://digitalsts.net/wp-content/uploads/2019/03/26_Knowing-Algorithms.pdf (Accessed: May 24, 2022)

Stevenson, M. and Doleac, J.L. (2019). Algorithmic Risk Assessment in the Hands of Humans. SSRN Electronic Journal [Preprint]. [PDF] Available at: https://docs.iza.org/dp12853.pdf (Accessed: May 24, 2022)

Suchman, L.A. (2007). Human-Machine Reconfigurations: Plans and Situated Actions. 2nd Edition. New York: Cambridge University Press.

Tan, S. et al. (2018). Investigating Human + Machine Complementarity for Recidivism Predictions. [online] Available at: http://arxiv.org/abs/1808.09123 (Accessed: May 24, 2022)

Van Berkel, N., Goncalves, J., Hettiachchi, D.,Wijenayake, S., Kelly, R. M. y Kostakos, V. (2019). Crowdsourcing Perceptions of Fair Predictorsfor Machine Learning: A Recidivism Case Study. Proceedings of the ACM on Human-ComputerInteraction, 3(CSCW). [PDF] Available at: https://dl.acm.org/doi/abs/10.1145/3359130 (Accessed: 13-9-2022)

Wagner, B. (2019). Liable, but Not in Control? Ensuring Meaningful Human Agency in Automated Decision-Making Systems. Policy and Internet, 11(1), pp. 104–122. [PDF] Available at: https://onlinelibrary.wiley.com/doi/epdf/10.1002/poi3.198 (Accessed: May 24, 2022)

Zhang, Y. et al. (2020). Effect of confidence and explanation on accuracy and trust calibration in AI-assisted decision making. FAT* 2020 - Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, pp. 295–305. [online] Available at: https://arxiv.org/pdf/2001.02114.pdf (Accessed: May 24, 2022)

Zuiderwijk, A., Chen, Y.C. and Salem, F. (2021). Implications of the use of artificial intelligence in public governance: A systematic literature review and a research agenda. Government Information Quarterly, (May 2020), p. 101577. [online] Available at: https://www.researchgate.net/publication/350333329_Implications_of_the_use_of_artificial_intelligence_in_public_governance_A_systematic_literature_review_and_a_research_agenda (Accessed: May 24, 2022)

Acknowledgements

Lead author

Manuel Portela is a postdoctoral researcher at the Web Science and Social Computing Research Group (Universitat Pompeu Fabra). His background is in interaction design and human geography. His research focuses on algorithmic fairness and how the interaction between humans and machines impacts the use of AI in decision-making scenarios. Before his academic career, he led digital transformation projects at the local administration of Buenos Aires (Argentina).

Co-author

Tanya Álvarez leads the Digital Future Society Think tank research on digital divides and digitalisation of the public sector. She advocates for an interdisciplinary perspective of how technology impacts society. She has a degree in art history from Swarthmore College and a master's degree in cultural heritage management from the University of Barcelona.

Think Tank team

Thank you to the following Digital Future Society Think Tank colleagues for their input and support in the production of this report:

- **Carina Lopes**, Head of the Digital Future Society Think Tank.
- **Olivia Blanchard**, Researcher, Digital Future Society Think Tank.

Citations

Please cite this report as:

- Digital Future Society. 2022.Towards meaningful oversight of automated decision-making systems. Barcelona, Spain.

Contact details

To contact the Digital Future Society Think Tank team, please email:
thinktank@digitalfuturesociety.com


Digital
Future Society