Digital Future Society

Future.Government.

Design criteria for algorithmic systems: Guidelines for the public sector

A programme of

GOBIERNO DE ESPAÑA
VICEPRESIDENCIA TERCERA DEL GOBIERNO
MINISTERIO DE ASUNTOS ECONÓMICOS Y TRANSFORMACIÓN DIGITAL
SECRETARÍA DE ESTADO DE DIGITALIZACIÓN E INTELIGENCIA ARTIFICIAL
red.es
MOBILE WORLD CAPITAL BARCELONA

Digital Future Society

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 under key themes including public innovation, digital trust, citizen empowerment and equitable growth.

Visit digitalfuturesociety.com to learn more

A programme of

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

red.es

MOBILE WORLD CAPITAL™ BARCELONA

Published
December 2019

Disclaimer
The information and views set out in this report do not necessarily reflect the official opinion of Mobile World Capital Barcelona 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.

This report was produced in collaboration with:

Digital Future Society

Contents

Introduction

Algorithmic technology can fundamentally influence what a government does and how it operates. The integration of algorithmic systems across a wide range of public sector organizations is mainly used to support a variety of decision-making mechanisms, changing those structures into automated or algorithmic decision-making systems (ADMS). The shift towards automation has invited a lot of questions and expectations from the public around the uses of automation and its social and ethical impact.

Government plays an important safeguarding role to ensure human (and humane) intervention in ADMS will lead to fair and just outcomes for society. While much of the technological innovation is driven by the private sector, the application of ADMS in the public sector tends to have a significant impact on society and individuals. The use of ADMS by governments shapes our lives and opportunities; as such, the public sector faces much higher expectations to act in a fair and lawful manner in order to protect the interests of its citizens. Public institutions thus face a particular responsibility with regard to implementing ADMS.

But how can this responsibility be fulfilled? To ensure a transparent dialogue between the public and its institutions takes place, **governments need to develop a trustworthy, fair, and accountable approach to using ADMS.** It is therefore crucial for governments to keep abreast of the legal and ethical frameworks surrounding the development and use of this technology. Algorithmic technology and ADMS should always be designed in ways that respects human rights, democratic values and diversity. People's privacy and security must be protected and bias has to be mitigated, which depends on the data being of good quality and used in an ethical manner.

There have been a significant number of policy papers and consultations to improve the design of ADMS in keeping with ethical and social parameters. The Barcelona Declaration for a Proper Development and Usage of Artificial Intelligence in Europe[1] (Declaration) offers a useful set of guidelines that emphasise the importance of "prudence, reliability, accountability, responsibility, constrained autonomy and the human rule" in the design and usage of AI. Building on these ongoing discussions, this paper provides an overview of the application of ADMS in the public sector and explains which design criteria are needed to ensure a responsible implementation of ADMS throughout its life cycle.

The public sector includes all levels and institutions of government, including federal, state, regional and local level. By public sector we mainly refer to institutions of the executive (ministries, police forces, public administration), but also the judicial (e.g. courts) and legislative (e.g. parliaments) branches. The public sector is not limited to institutions of the state alone, but includes companies directly operated by the state or serving public functions (e.g. federal identity providers). We also address institutions directly funded by the public (e.g. libraries).

In the debate around ADMS, many different terms are used to describe the systems and developments. The term "algorithm" refers to a set of precise instructions or rules regarding actions to be taken in solving a predefined problem. In this paper we use the term ADMS or "algorithmic system". These are systems comprised of one or more algorithms used in a software to collect and analyze data as well as draw conclusions as part of a process designed to solve a pre-defined problem. Such systems are often used to inform, prepare or directly make decisions on the allocation of resources, information, positions or freedoms. This is why we often use the term algorithmic or automated decision-making systems. The scope of this description goes beyond the mere code of the software in use, by including the broader socio-technical context in which the software is embedded. For example, this involves a consideration of how results are interpreted, and how this informs the user of a system's decisions of which data is being used and how this data is being collected.

ADMS can be roughly divided into two groups: Systems which follow pre-programmed decision-making rules, and systems which involve some form of machine learning. The former are often relatively simple systems where the parameters to be applied and the calculations to be made are directly coded into the software and thus directly influenced by those designing these systems. The latter are such systems, which do not have a pre-defined problem-solving mechanism and rather use an analysis of training data or some other form of training mechanism to extract patterns and develop the ADMS. Such systems are also referred to as (narrow) artificial intelligence.

[1] B•Debate 2017

This report aims to provide a short overview of the field of ADMS and establish a set of guidelines regarding its application. In addition to listing the goals of ADMS usage in the public sector, the paper offers an overview of specific algorithmic systems currently in use across the world, with a particular focus on North America and Europe. **The paper also lays out recommendations on how to approach the implementation of ADMS in order to ensure the protection of existing rights and interests of citizens.** The Algo.Rules (Design Criteria for Algorithmic Systems)[2], developed by the Bertelsmann Stiftung jointly with the think tank iRights.Lab, have been selected as a benchmark to ensure responsible design of ADMS[3]. We give specific recommendations on how to apply these rules in the public sector in the form of guiding questions. To that purpose, three systems are chosen as examples to illustrate what this implementation would mean in practice. We close by giving a few general recommendations which the public sector should follow in order to operationalize fair, accountable and trustworthy use of ADMS.

[2] https://algorules.org/en/home/

[3] Bertelsmann Stiftung, iRights.Lab 2019

Goals and expectations of the use of ADMS in the public sector

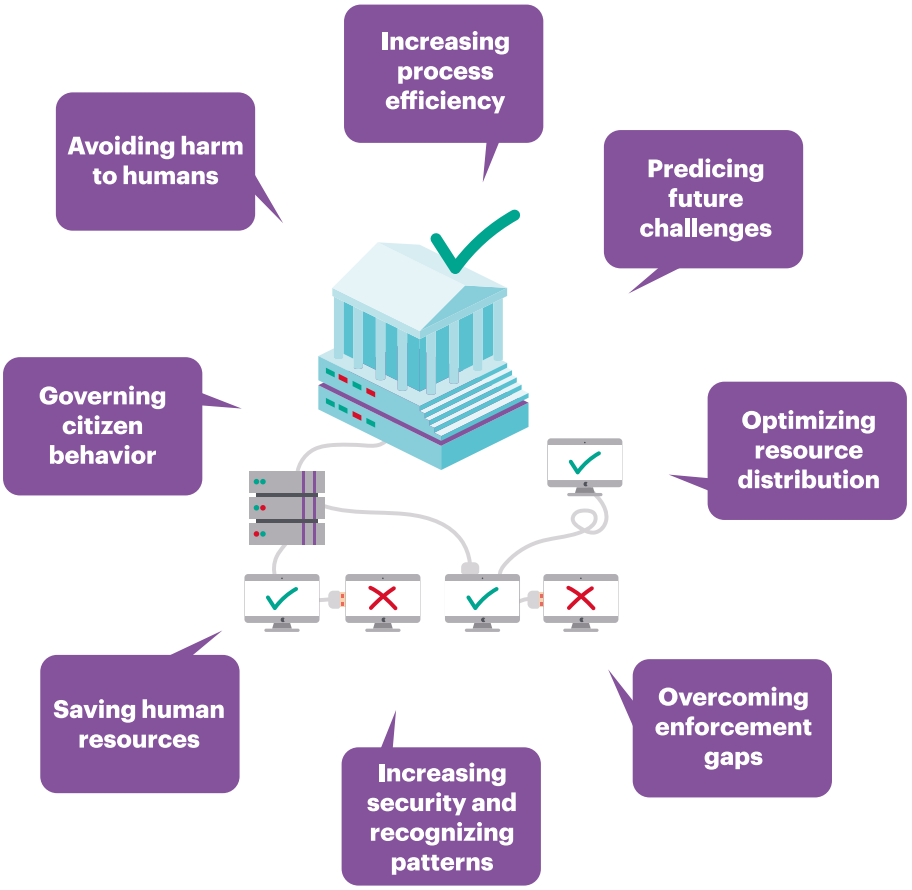


Figure 1: Goals for the use of ADMS in the public sector

ADMS has multiple uses for the public sector which are invariably linked to the expectations set by algorithms. In order to give useful recommendations on how ADMS should be designed in the public sector, we need to understand why they are being used. Generally, these expectations include: Increasing efficiency of a decision-making process; enabling management of complex systems with a lot of data; ensuring consistency of decision-making; increasing fairness through elimination of human bias, and supporting or relieving human labour by optimizing processes of public administration and decision-making. The main application of ADMS within the public sector—whose goals are partly overlapping and interdependent—are listed in greater detail below[4]. The aim of this paper is not to evaluate these overlapping goals for their public benefit or associated risks, but rather to offer a general overview of the goals and uses of ADMS within the public sector.

[4] See also Carahsoft, UiPath 2019, Molavi & Lahmann 2018

1 Increasing process efficiency

First and foremost, ADMS are used to optimize and enhance existing human-driven processes. One of the main incentives for the implementation of ADMS by governments looking to improve existing processes, is the time saved analyzing data and streamlining decision making processes. ADMS systems can be used to analyze applications for public benefit and decide on whether they are valid in a fraction of the time a human reviser would need. This ensures that more applications can be processed faster and in a shorter time frame.

2 Predicting future challenges

Long-term planning can be enhanced by the use of ADMS. ADMS can help to deal with situations of high uncertainty by analyzing existing data and making projections about future developments. Its usage has multiple public service benefits within the educational, public health, policing sector. For example, future unemployment could be significantly mitigated by making changes to the educational system based on predictive algorithmic calculations that anticipate shifts in the labour market. Predictive policing software is used in a similar way: By utilizing models based on the near-repeat-theory—which states that burglars tend to strike again near the location of a successful break-in—statistical analysis can identify areas where burglaries or car theft are likely to occur and act accordingly.

3 Optimizing resource distribution

In times of increased stress on public budgets, public sector resources, both financial and human, are limited. Technology, especially ADMS, allow for efficient use of resources available and limit misallocation of funds to ensure they are allocated to areas where they are most needed. Not only does this limit waste, but also it can ensure a fair distribution of resources on an empirical basis. In a similar vein, ADMS are being used to personalize services or information offered by the public sector so as to detect fraud. As an example, ADMS are used to analyze applications intended for public benefits in order to prevent individuals from receiving funds that they are not eligible for.

4 Overcoming law enforcement gaps

The public sector faces challenges around law enforcement. Law breaches in the digital area are not easy to track, and even more difficult to regulate. This includes breaches of privacy laws, copyright breaches and the discriminatory behavior of software are difficult to monitor and sanction. The lack of transparency, coupled with the imbalance of powers between big tech and public law enforcement and a lack of ideas of how to handle the situation has led to a rethinking of the role of automation in law enforcement. For example, ADMS are already being used to flag social media posts with inflammatory content like hate speech or cyberbullying. This could be extended to other online content production where intervention is required.

5 Increasing security and recognizing patterns

ADMS play an important role in automated surveillance programs. They autonomously identify relevant areas of communication data that require interception and greater analysis. By compiling and analyzing several related data sets, ADMS help to connect different pieces of information about a person or a subject, making patterns more visible and allowing for better analysis with more precise conclusions to be drawn. Increased security is a particularly relevant result. The implementation of ADMS in criminal justice systems thus presents a pertinent use case for its potential to increase security, particularly the ability to detect patterns of criminal behavior and assess whether a person is at risk of relapsing.

6 Saving human resources

ADMS can be used to automate tasks which would have usually been carried out by humans. The public sector can therefore use such systems to decrease the pressure on human resources. This can relieve public servants from tasks that would otherwise be time consuming, leaving them to focus on more important tasks which ADMS are not equipped to carry out, such as interactions with clients. Such applications can further bridge skills shortages increasingly present in the public sector. For example, the lack of personnel to deal with the management of migration, is one reason for the use of ADMS to speed the process of asylum applications.

7 Governing citizen behaviour

A controversial use of ADMS is citizen surveillance and data analysis, which aims to make governance easier. Collecting and permanently analyzing data on citizen behavior using ADMS, can help to nudge citizens to favorable behaviors such as social conformity, obedience to existing rules or the avoidance of unhealthy behavior. As an extreme example, China's social scoring system demonstrates how surveillance and actively communicated guidance tied to an incentive and punishment system, can be a powerful tool of governance. But of course, it shows how the use of ADMS for governing citizen behavior can also lead to unfavourable and oppressive outcomes for citizens.

8 Avoiding harm to humans

ADMS can be used to operate systems in situations where the execution by a human would be dangerous or in some way harmful. One example would be the use of ADMS to automate robotic systems to fight disasters like forest fires or chemical incidents. In such instances as these, the situation can be monitored and dealt with without the need for humans to expose themselves to dangerous situations.

ADMS: Use cases

In order to further develop an understanding of how ADMS is used in the public sector we present three illustrative use cases[5]. They are intended to give a general overview of the key uses of ADMS by the public sector and the role they play in achieving the above-mentioned goals. As very few public institutions have compiled lists of all ADMS they currently use, our research does not offer a comprehensive overview but goes some way to determining areas in which their deployment of automation is beneficial to the public sector. The cases listed below are predominantly drawn from North America and Europe due to data availability.

VeriPol (detection tool for false police reports)

VeriPol is a tool used to determine whether a report made to the police is false or worthy of further investigation. An ADMS automatically analyzes calls using natural language processing and machine learning components. The system is currently under development by the Spanish National Police. The goal is to increase efficiency in detecting trustworthy police reports, as false allegations are costly in terms of financial and human resources. False claims can then be penalized under articles 456 and 457 of the Spanish Penal Code[6].

SAVRY (Structured Assessment of Violence Risk in Youth)

The system SAVRY (other similar systems are J-SOAP-II, J-SORRAT-II) is used in forensic criminology and was developed for assessing the risk of violence in adolescents (aged 12-18). According to the suppliers of the system, SAVRY is composed of 24 items in three risk domains (Historical Risk Factors, Social/Contextual Risk Factors, and Individual/Clinical Factors), drawn from existing research and the professional literature on adolescence[7]. It works as an additional tool of assessment to human-made assessments of individuals. SAVRY can be used by juvenile and criminal courts and at nearly every juncture of the juvenile justice system in order to predict potential future behavior.

[5] See also Spielkamp (ed.) 2019, Digital Future Society 2019
[6] Lara Quijano-Sanchez et al. 2018
[7] Borum et al. (n.d.)

EMMA (virtual immigration assistant)

Since the end of 2015, the U.S. Citizenship and Immigration Services ("USCIS") launched a new virtual assistant named "EMMA"[8]. The system provides customers with quick help to allow for easier navigation in finding information from USCIS. Similar to Apple's SIRI, Emma can receive typed questions and guide users through USCIS' website to find answers. EMMA answers questions based on your own words and can provide immediate answers. It can therefore help to optimize the access to and distribution of information pertaining to immigration related queries.

LAWs (Lethal Autonomous Weapons)

ADMS are used in the military sector for multi-purpose applications. One such application is the use of ADMS for surveillance and pattern recognition, in order to detect potential targets in conflict zones and to launch an armed attack against the target, e.g. with missiles. These robotic weapons, referred to as LAWs, use drones or guided projectiles to execute decisions made by the ADMS. These systems have varying degrees of autonomy, from having a human in-the-loop at various stages of the process, to being fully autonomous. While obviously aimed at destroying targets, the automation and robotization of warfare is intended to minimize human risk.

RisCanvi (actuarial risk assessment tool)

RisCanvi is a tool for assessing the risk of violence in prisons, as well as the risk for recidivism of individuals[9]. The system is operated in Catalonia by the Prisons Directorate, Barcelona University and the Research and Training Institute of the Catalonian Justice Department. The tool is a statistical risk assessment system similar to LSI-R in Canada, COMPASS in the US and OaSys in the UK. It takes into account different risk factors, such as penitentiary history, personal and social background as well as an individual's psychological profile. Since 2010, the tool is applied to all inmates in prisons, both for cases involving violent crime and to predict the future behavior of inmates. Systems like RisCanvi and COMPASS are, for example, being used in trials to determine whether a criminal should be granted parole or to help determine the length of the sentence. These ADMS are not used in a fully automated manner, but are intended to inform and assist judicial experts in the decision-making process.

[8] USCIS 2019
[9] Andrés-Pueyo et al. 2017

Trelleborg Model (tool for social benefit assessment)

Since 2017, the Swedish municipality of Trelleborg uses ADMS to automate parts of its decision-making in the welfare sector[10]. Social benefit applications are automatically screened using an ADMS, in order to decide on their validity and to detect potential fraud. New applications for benefits are automatically checked and cross-referenced with other related databases; for example, data on housing support or from the tax agency. The system here is fully automated, with decisions on benefit applications being directly related to the ADMS results. The short-term goal of the project is to reduce the administrative burden on caseworkers and to give clients faster responses to applications[11]. In the longer term, the plan would be to reallocate saved resources to social work and improve the service of the social welfare institutions.

AMS (automated unemployment classification)

At the end of 2018, the Public Employment Service Austria (Arbeitsmarktservice – AMS) announced a plan to partially automate the process of allocating measures to reintegrate the unemployed into the labor market. The underlying statistical model that determines the factors relevant for the evaluation of reintegration[12], is publicly available, making the process transparent. The ADMS evaluates the chances of a job placement for unemployed persons, and then sorts the findings into three groups with high, medium and low reintegration chances. With a launch date of 2020, it remains to be evaluated whether the ADMS will have a direct impact on reintegration measures, such as the allocation of funds as per the individual's needs. The main aim is to optimize the allocation of these measures as well as lighten the burden of social service workers.

Taken together, the above examples of ADMS currently in use across the public sector, offer a useful overview of the diverse contexts in which these systems are used. Accordingly, the design of these automated systems must pay heed to the specific socio-political and cultural context they seek to serve. It is also crucial to consider the various risks introduced by the use of these systems and how they affect different kinds of groups or individuals.

[10] Ranerup & Henriksen 2019, Wisterberg 2018
[11] Nauwerck & Cajander 2019
[12] Holl et al. 2018

Impact relevance and analysis

In order to determine the requirements for process-oriented design criteria, an assessment of impact relevance needs to be carried out prior to the use of the ADMS.

There are currently different impact assessment methods for the public sector that have been designed in collaboration with private companies, academia and civil society. Examples include: the calculation of algorithmic processes' impact opportunities[13], the criticality pyramid for algorithmic systems[14], context-specific analysis of damage potential of algorithmic systems[15], the data ethics decision aid[16] or the Canadian Algorithmic Impact Assessment[17]. All of these systems do not analyze the code of the ADMS itself, but take into account its socio-technical context, including the application's environment, people and institutions involved with the decision-making process, the data used, as well as all other components relevant to the result and impact of the ADMS. The methods for assessing algorithmic impact differ, particularly as the calculation for its opportunities are well defined and quantifiable, whereas the criticality pyramid for algorithmic systems is not yet sufficiently defined to be used in practice.

For our own purposes, we summarize these methods into one model. In general, these impact assessments may consider a few or all of the following factors to determine the relevance of a specific ADMS:

- Direct impact on individuals and society, considering their reversibility, significance, durability and scope
- Indirect impact on individuals and society, considering their reversibility, significance, durability and scope
- Worst-case scenarios or effects of mistakes in the ADM process
- The environment of the ADMS, specifically the potential for alternative processes from the perspective of users and persons affected, as well as the variety of providers and systems in place or lack thereof
- The role and situation of those affected by the ADMS, e.g. their vulnerability towards discrimination and the competencies available.

Based on the impact relevance assessment, a risk-adaptive regulatory approach should be taken. In general, there are four groups or risk levels into which ADMS can be allocated:

1. Systems with no significant impact, therefore not needing any additional regulation or control mechanisms
2. Systems with some significant impact, requiring some form of regulation or control, ranging from ex-post reviews up to assessments prior to implementation
3. Systems with a high impact, requiring strict regulation and harsh control mechanisms
4. Systems with unacceptable potential consequences, which should not be put in place at all.

[13] Vieth & Wagner 2017
[14] Datenethikkommission 2019, p. 177
[15] Krafft & Zweig 2019
[16] Utrecht Data School 2019
[17] Treasury Board of Canada Secretariat 2019

To further sharpen the analysis and illustrate our recommendations on a practical level, we identify use cases from the ADM applications listed above. Taking into consideration the developed model for impact relevance of ADMS, and its deployment across diverse public sector industries, we identify three use cases. Below we establish their impact relevance by describing in more detail the ways in which these ADMS have an effect on individuals' lives and the functioning of society. The aim of our analysis is to highlight how mistakes in the design process can lead to massive consequences and the importance of implementing Algo. Rules at every stage of the process.

The first example is ADMS in the military sector. Here, ADMS are used to make split-second decisions of whether to use a certain weapon against a target or not. These systems are often discussed under the term lethal autonomous weapons (LAWs), where ADM is combined with systems of intelligence and surveillance. These systems in general have a very high impact relevance as their use directly affects the lives of people, and possible mistakes can have immediate and massive consequences.

The second example are the tools RisCavi and COMPASS, which are being used for risk assessments in prisons. They are predictive tools that use ADMS to prepare a decision rather than being fully-automated. Given that the risk scores produced by this system can have a significant impact on the lives of prisoners, the decisions made by these systems is used in trials to assist the judge in determining the length of sentence or the validity of parole. It offers a pertinent example of the use of ADMS as a complementary, rather than supplementary solution to decision-making in which careful adherence to the Algo.Rules must be taken.

The third example is the Trelleborg Model currently in use in Sweden, which highlights the use of ADMS in the social/ welfare sector. Here, an optimization system is being used to allocate social benefits to those in need in order to decide on the validity of applications while taking the burden from case workers. A study[18] found that the possibility of having more time to work with clients was seen as positive, but respondents were also aware that increased digitization might leave some clients more vulnerable, and that caseworkers' insights into their clients' needs may be reduced through an automated process. This case highlights the possibility of disenfranchising practices resulting from the use of an ADMS, as well as the risk of automated errors. An early 2019 the Swedish media outlet SVT reported[19] that a serious system error at the Employment Service means that more than one in seven decisions about canceled support may be incorrect. This has some significant impact on individuals' lives (risk level 2) as they depend on receiving benefits.

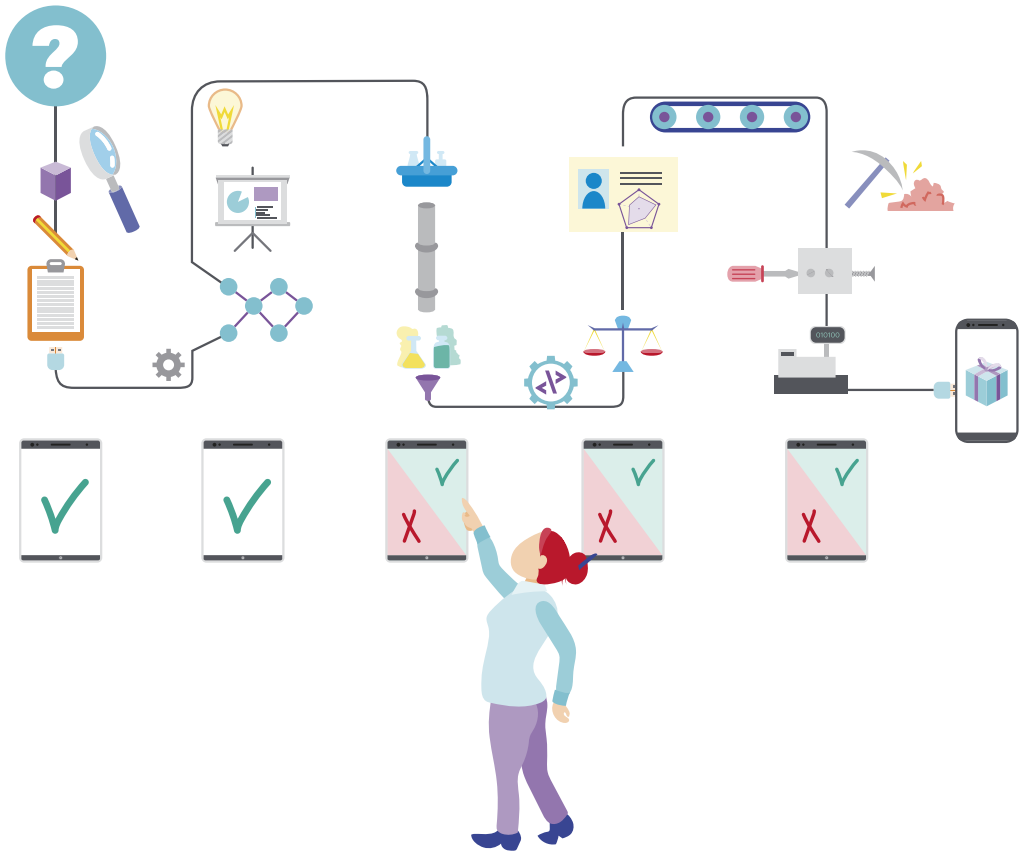
In the following chapter, we return to these cases and evaluate them against the Algo.Rules.

[18] Nauwerck & Cajander 2019
[19] SVT 2019

Guiding questions: Implementing Algo.Rules in public sector ADMS

The above examples pose significant challenges when thinking about human rights and basic ethical norms. Whilst the use of ADMS offers several advantages as described above, it also carries risks to social well-being and individual as well as collective rights.

To manage these challenges, a large number of different ethical guidelines have been developed. Collectively, this set of guidelines help to regulate the use of ADMS, and can be divided into two different but interconnected categories. Firstly, there are declarations which develop a set of normative principles, e.g. developing new moral norms for programmers such as respect for human autonomy, solidarity or respect for the freedom of others. These often build or expand on existing norms, derived from the ethics discourse or basic legal principles like human rights and redefining them for ADMS. They aim to define which ADMS are "good" and which are not.



Secondly, we include guidelines that lay out a set of formal principles. These focus on the development process of ADMS and define organizational and/or technical measures which should be taken into account "by design". Such principles, namely transparency, accountability and security, are not goals by themselves, but are means to an end that ensure the protection of existing norms. Therefore, they do not define new values but rather list the criteria to be follow so as to realize certain outcomes.

The Algo.Rules offer the most comprehensive set of formal design criteria, which are worth exploring in greater detail[20]. They were developed by the Bertelsmann Stiftung in cooperation with the think tank iRights.Lab and a large pool of about 450 contributors. It is a catalogue of formal criteria for enabling the socially beneficial design and oversight of algorithmic systems that aims to provide the basis for ethical considerations, and the implementation and enforcement of legal frameworks. The nine Algo.Rules focus on algorithmic systems that have a significant impact on society or individual lives. These criteria should be integrated at the start of any automated system, from the design stage through to the development stage. We focus on the Algo.Rules because they consist of design principles of a purely formal character. Leaving out normative discussions and values, the Algo.Rules offer relatively specific and process-oriented recommendations which can be easily expanded and adapted to different sectors. Therefore, they seem suitable to develop specific recommendations for the public sector to help it develop and use ADMS for the social good.

There are several other guidelines developed that are useful to consider. One of them is the Barcelona Declaration for a Proper Development and Usage of Artificial Intelligence in Europe, which was developed by a group of researchers during a workshop held in Barcelona in 2017 and further developed through additional meetings, workshops, and AI schools. Additional studies have analyzed a number of other guidelines[21]. The existing guidelines comprise normative and/or other formal principles and recommendations aimed to harness the potentials of ADMS while minimizing social risks.

We believe that a focus on the Algo.Rules should not overlook other guidelines. However, an analysis of these guidelines shows that there is significant overlap between them in terms of similar traits and principles and a general consensus on most of the underlying principles. It is worth mentioning that other guidelines were considered during the development of the Algo.Rules[22]. However, we believe the Algo.Rules offer more specific criteria then many other guidelines and have a clear advantage when applying them in practice due to their focus on process-oriented criteria.

Building on the Algo.Rules, we have developed guiding questions and specific recommendations for their implementation, working from the "what" to the "how"[23] by drawing on the piloting process of the EU Guidelines for Trustworthy AI[24]. Below, you can see the Algo.Rules listed on the left column. On the right column, we have developed a list of guiding questions, highlighting challenges and aspects specific to the public sector. These are intended to help public institutions identify potential gaps in their current procedures and competencies, and find ways of how to counteract them. They are not intended to provide a fine-grained solution, but rather invite a consideration of how to implement measures for the Algo.Rules and other process-oriented criteria for ADMS. The specific requirements for measures may depend on the legal framework already in place as well as the specific context for each and every ADMS application.

[20] For lists of other guidelines, see Hagendorff 2019, Jobin et al. 2019
[21] Hagendorff 2019, Jobin et al. 2019
[22] Rohde 2018
[23] Morley et al. 2019
[24] High-Level Expert Group on AI 2019

Guiding questions for the use of ADMS in the public sector

1. Strengthen competency

The function and potential effects of an algorithmic system must be understood.

Those who develop, operate and/or make decisions regarding the use of algorithmic systems, must have the necessary expertise and appropriate-to-scale understanding of how the technology functions and its potential effects. Sharing individual and institutional knowledge as well as promoting interdisciplinary exchange across task areas are just as crucial as ensuring appropriate skills development. These approaches should be integrated into the education, training and onboarding of new employees. In addition, interdisciplinary exchange should be an ongoing endeavor that remains open to those who are interested and/or affected.

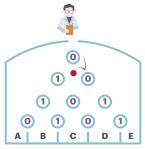
- Is the public sector aware of the competencies it is currently lacking? Has an up-to-date assessment of competencies been carried out?
- Are public servants, especially those dealing with the management of ADM, sufficiently aware of its effects?
- Are competencies in all relevant dimensions (technical, design and application, societal, legal) covered and further developed in the public sector?
- Is there sufficient area for discussion and interdisciplinary reflection, e.g. through workshops or training?
- Are good and bad practices as well as other learnings shared across the public sector, including in institutions?
- Are public servants actively seeking to gain knowledge about a specific ADMS when talking to ADMS providers?
- Are there measures in place to actively improve competencies among stakeholders, specifically those directly affected by the ADMS as well as the broader public?
- Are the public servants informed about the human bias that informs automation design and are they trained to use countermeasures to neutralize it?

2. Define responsibilities

A natural or legal person must always be held responsible for the effects involved with the use of an algorithmic system.

Accountability must be clearly assigned. The accountable person must be aware of the responsibilities associated with their tasks. This also applies to responsibilities that are shared by several people or organizations. The allocation of responsibility must be fully documented and transparent for internal and external parties. Responsibility may not be transferred to the algorithmic system itself, users or people who are affected by the system.

- Are roles and responsibilities in the public sector clearly defined regarding the ADM application and all its components, taking into account competencies?
- Joint controllership: Are responsibilities discussed and defined in relations with ADMS providers and other organizations involved in the development and/or application of the ADMS? Are they being made aware of their responsibilities, e.g. providing additional information, asking questions to the public sector, conducting impact assessments? Have the terms and duration of responsibilities been discussed and defined?
- Are the responsibilities clearly and officially assigned and documented? Are they set and agreed upon, e.g. as part of the contract?
- Are the assignments transparent to the public?



3. Document goals and anticipated impact

The objectives and expected impact of the use of an algorithmic system must be documented and assessed prior to implementation.

The objectives of an algorithmic system must be clearly defined and information regarding its use must be documented. This includes the underlying data and calculation models. Before an algorithmic system is put to use, an impact assessment should be conducted and documented. Particularly in the case of machine-learning systems and in dynamic areas of application that are subject to frequent change, an impact assessment should be repeated at regular intervals. The risk of discrimination and other consequences affecting individuals and the common good must be taken into consideration. The objectives considered, their underlying values and the use of algorithmic systems must be documented.

- Have the goals of a specific ADMS been defined and reflected within the institution as well as with the ADMS provider? Have unintended consequences been identified?
- Are the goals and anticipated impacts regularly assessed and, if necessary, adjusted? Are corresponding consequences for the impact of the ADMS taken into account?
- Have the relevant stakeholders and persons affected been identified for the implementation of a specific ADMS?
- Has an impact assessment been carried out, taking into account different scenarios, the perspectives of stakeholders and those affected?
- Have the goals and the impact assessment been reflected within the institution, with relevant stakeholders as well as with the ADMS provider and other project partners?
- Has the impact assessment been made public?
- Has the architecture of the ADMS been identified, including models and data used? Has an assessment been carried out regarding the type, purpose and use duration of data being used?

4. Guarantee security

The security of an algorithmic system must be tested before and during its implementation.

The reliability and robustness of an algorithmic system as well as its underlying data with respect to attacks, access and manipulation must be guaranteed. Security must be built into the architecture of the algorithmic system (security by design). The system must be tested in a protected environment prior to implementation. Security precautions must be documented.

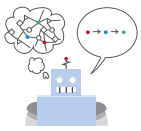
- Has a security risk assessment been carried out?
- Has the implementation of the ADM application as open source software been considered?
- Is software in use compliant with state-of-the-art security standards?
- Are regular updates carried out?
- Is the data in use for the ADMS sufficiently secured?
- Is the security of the system tested before implementation?
- Has a post-vulnerability plan been drafted?

5. Provide labeling

The use of an algorithmic system must be identified as such.

People interacting with algorithmic systems must be able to identify that a decision or prediction is based on an algorithm. This is particularly important in cases where the system imitates a human being in how it interacts (e.g., through language or appearance).

- Have the requirements of the users and people affected regarding labeling and potential risks of deception been identified?
- Is a comprehensive label being used to communicate to the public that an algorithmic system is part of the decision-making process?
- Is the label accompanied by an explanation of the general functioning of the algorithmic system, including the goals, mode, role to the user as well as used data (Algo.Rule 6)?
- Is the label easily accessible and visible? Is the language being used as part of the label inclusive and easy to understand (Algo.Rule 6)?

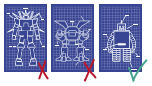


6. Ensure intelligibility

The decision-making processes within an algorithmic system must always be comprehensible.

In order to question and review decisions resulting from an algorithmic system, people must be able to understand both direct and indirect effects of the system as well as how decisions are reached. Information about the data and models on which the system is based, its architecture and potential effects, must be published in easily understood terms. In addition, it is important to check whether an objective can be achieved without a significant loss in quality through the use of a less complex algorithmic system that involves an easier to understand mode of operation.

- Have the requirements for intelligibility for different stakeholders (esp. users of the ADMS within the public administration and the persons affected by the ADM decision) been determined?
- Is a clear explanation of the functions of the system included in the label?
- Is a specific and individual explanation of a single ADM result available or are systems in place to provide an explanation if needed?
- Is that explanation made intelligible for the relevant stakeholders? Have potential challenges to ensure intelligibility, e.g. dyslexia, been taken into account?
- Is the explanation easily accessible and actively communicated?



7. Safeguard manageability

An algorithmic system must be manageable throughout the lifetime of its use.

In order for an algorithmic system to remain adaptable, everyone involved in its development and implementation must maintain joint control over the system. This involves ensuring broad oversight of the entire system, even when tasks are distributed across various departments within an organization and among several individuals. The complexity of a system's operations must never exceed the capacity of human oversight and a person's capacity to make changes to the system. This applies in particular to machine-learning systems. If this manageability cannot be guaranteed, the algorithmic system in question should not be used.

- Is the ADMS, its technical functions and features as well as its effects on society, monitorable?
- Is the ADMS adjustable in case of malfunctions or other reasons that would make adjustments necessary?
- Is there sufficient access and competence within the public institution present and/or is there a long-term cooperation with the ADMS provider in place to ensure long-term manageability?
- Is the life-cycle perspective of the ADMS taken into account?
- Is the ADMS stoppable? Is there a back-up plan in place if the ADMS needs to be turned off? Are the potential effects of (temporarily) stopping the ADMS determined and evaluated?
- Are changes to the system documented?

8. Monitor impact

The effects of an algorithmic system must be reviewed on a regular basis.

An algorithmic system must be subject to active monitoring in order to determine whether the targeted objectives are actually achieved, and the use of the system does not violate existing legislation. Taking the appropriate technological precautions, external bodies should be able to conduct an independent, comprehensive and effective audit of an algorithmic system without compromising legitimate concerns regarding business confidentiality. Should a negative impact be determined, the cause must be identified and the algorithmic system adapted accordingly.

- Are the defined goals for the ADMS, its architecture and functioning regularly reviewed?
- Is there a permanent monitoring of the impact of the ADM application on users and other people affected in place?
- Are external actors, such as civil society organizations, as well as competent regulatory authorities allowed and empowered to monitor the impact of the ADM application through audits?
- Is external feedback and critique welcomed?
- Is the accuracy and quality of data guaranteed?
- Is the impact of the ADM application reevaluated in case of changes of goals, architecture of the ADMS or a change of circumstance or environment?

9. Establish complaint mechanisms

If an algorithmic system results in a questionable decision or a decision that affects an individual's rights, it must be possible to request an explanation and file a complaint.

The person or organization using an algorithmic system must provide an easily accessible means of contact. First, those affected must be able to request appropriate and detailed information regarding a specific decision and the considerations that have fed into it. This should be an option also for organizations acting in their legitimate interest and for situations in which an organization acts on the behalf of an individual. Second, there must be an easily accessible and effective way to lodge a complaint. Complaints and actions taken must also be documented.

- Is there a contact point, e.g. an ombudsperson, selected by and accessible to the public, especially the persons affected?
- Is the contact point for complaints sufficiently staffed and authorized to effectively follow up on complaints, including across different pubic institutions?
- Is there an internal ombudsperson in place to allow for whistleblowing and other forms of internal complaint procedures?
- Are qualified organizations, e.g. civil society, enabled to file complaints on behalf of citizens?

Use cases of Algo.Rules implementation

The Algo.Rules and the accompanying guiding questions, aim to help the public sector arrive at a responsible design of ADMS. In order to illustrate how such an evaluation is undertaken and what the Algo.Rules mean in practice, we connect them to the use cases identified in chapter 4 and give illustrative recommendations.

LAWs

The first use case is the dual use of ADMS in the military sector, specifically their use for LAWs (Lethal Autonomous Systems). Before starting with the evaluation of these systems using the Algo.Rules, it is worth referring to the initial impact relevance assessment described in chapter 4, as it provides important guidance. LAWs in general have a very high impact relevance, as their use directly affects the lives of people; any mistake can therefore have immediate and detrimental consequences. The impact relevance and the associated risks are so high that we can justly categorize such systems as risk level 4, which means these ADMS are deemed unacceptable. These systems would fall into the red section (risk too high for use) of the graph in chapter 4. In that case the public sector should refrain from using ADMS in this manner at all[25].

If such systems fall within risk level 3, they would still require harsh scrutiny and significant control mechanisms, for which Algo.Rules would provide protective guidelines. Relevant consequences under the Algo.Rules would include the following:

- There is a need to strictly ensure the manageability of the ADMS with clear human oversight (Algo.Rule 7). The system must not become too complex to be understood and changed by developers. Therefore, it should be a relatively simple system without machine learning components. Due to the extremely high impact on individual lives that launching a weapon through automated design would incur, the process would require substantial and permanent human oversight and intervention. These persons would need to have the proper qualifications to question the recommendations made by the LAW system. If this manageability cannot be guaranteed, the algorithmic system in question should not be used.
- Permanent and publicly transparent review and monitoring mechanisms assessing the impact of these uses of ADMS—both on those affected but also on the people operating these systems (Algo.Rule 8)—need to be in place. This would include detailed reports on the uses of LAWs, an overall analysis of the criteria which influence the recommendation made by the ADMS, and sociological and psychological analysis on the operators. The latter must always ensure that the oversight and ultimate decision made by persons is effective and not merely window-dressing.
- Finally, a system with direct impact that results in high-risk consequences would need to follow the highest security standards in order to avoid manipulation or hacking of the systems (Algo.Rule 4).

[25] UNODA 2017

RisCanvi

The RisCanvi system currently in use in Catalonia to calculate risk assessments for convicted felons or defendants, gives insight into the potential of ADMS use within the justice sector. It is comparable to other systems implemented elsewhere, such as LSI-R in Canada, COMPASS in the US and OaSys in the UK.

- This use case shows why increasing competency among those applying ADMS like RisCanvi is critical. For humans using ADMS to inform their decision-making across multiple system uses, there is always the risk of automation bias. Such a bias occurs when a human decision maker either disregards or does not search for contradictory information in light of a computer-generated solution which is assumed to be correct. The risk of course, is that operators are likely to turn over decision processes to automation as much as possible, making teams of people, as well as individuals, susceptible to automation bias[26]. They tend to blindly "trust" the recommendations made by an ADMS because it is based on data and therefore seems objective. Algo.Rule 1 would therefore call for building awareness among civil servants of the potential bias underling these systems, so as to empower them to make non-partisan judgement. It is important to train users to adopt a healthy skepticism towards the results of the system and be enabled to question its decision, potentially carrying out their own data analysis to check for mistakes.
- It is imperative for system users to question the results of ADMS, by training them on how the decision-making is designed and implemented by the system. This leads to ensuring intelligibility of the ADM process (Algo.Rule 6). With RisCanvi, people applying the system should be informed about which factors played the decisive role for the determined risk of recidivism, and on which data this decision was based.
- Automation bias might lead to ADMS unthinkingly using faulty data sets (Algo.Rule 3). In such cases where the potential for bias has already been proven, the design process of the ADMS would require a thorough assessment of the data being used to train the systems, as well as the data used during its application. For example, it should be imperative to assess early on whether certain types of data might be unsuitable or generally unreliable for a quantitative analysis of the risk of recidivism (e.g. psychological data).
- Furthermore, regular impact monitoring is extremely relevant in this case, as there is potential for discriminatory effects and other unintended consequences (Algo.Rule 8). While RisCanvi itself has not yet been subject to public scrutiny in that regard, an analysis of the similar system COMPASS showed that the the use of ADMS systematically discriminated against people of color[27]. Such an analysis should be detected and addressed early on, and not—as in the case of COMPASS—have to be carried out by journalists or civil society but instead by the public sector.

[26] Cummings 2012

[27] Angwin et al. 2016

Trelleborg Model

For the social and welfare sector, the Trelleborg Model was chosen as a use case. Here, an ADMS automatically checks applications for benefits and cross-checks them against databases such as for housing support or with data from the tax agency. Several issues can be identified regarding the observance of the Algo.Rules:

- Firstly, there seems to be a disregard for Algo.Rule 5, as no form of labeling was provided. Neither citizens nor benefit applicants were informed about the use of an automation process. Labeling, however, is essential in order to inform the public, especially those directly affected by the ADMS, about its use.
- Since such a label was missing, there was also no other additional information given which would have ensured intelligibility of the ADMS to the public. In such a case, Algo. Rule 6 would have required at the very least a general explanation of how the system functions and the data it uses to check benefit applications. Furthermore, the public institutions in which the system is deployed, should have provided clear explanation of why an application is denied, to ensure intelligibility of each individual benefit application processing. A person denied a benefit application should then have been informed about why this application was denied. Greater clarity around the decision-making process would have enabled people affected to effectively file a complaint.
- There were additional issues with the complaint mechanism (Algo.Rule 9), especially since the system made a significant number of mistakes[28]. Until today, it seems that neither the persons affected nor the public institutions involved understood when and why the ADMS lead to mistakes. Mistakes were found and then processed manually after several complaints were filed. However, the actual cause of the malfunctions was not identified. Therefore, the public institutions were unable to actively check decisions already made by the ADMS and people affected were not given sufficient clarity as to why such errors occurred. It is possible that the institutions using the Trelleborg Model could have established effective complaint mechanisms beforehand by hiring an ombudsperson competent with the ADMS. This would have ensured that the failings of the systems earlier would have been identified early on, and such a person would have provided a contact point for persons affected.
- There seemed to be issues regarding the definition of responsibilities (Algo.Rule 2). A study[29] found that public servants were aware that increased digitization might leave some clients more vulnerable and that caseworkers' insights into their clients' situation might be reduced when more of the process was automated. Case workers where therefore not able to review different benefit applications in a responsible manner. This also led to case workers leaving their workplace. Clearly, responsibilities were not sufficiently discussed within the institution.

[28] Wills 2019

[29] Nauwerck & Cajander 2019

Conclusions and recommendations

This study offers an overview of the uses of ADMS in the public sector. We show how process-oriented design criteria can help to implement mechanisms that protect existing norms and regulations through the design process.

The concept of process-oriented design criteria, and specifically the Algo.Rules, has proven to be helpful when assessing ADMS applications and finding potential blind spots. The implementation of such criteria, from the development stage through to the design of an ADMS, is crucial to ensure those systems serve society. This is especially the case for the use of ADMS in the public sector, as those applications generally tend to have a higher societal relevance and impact.

To ensure the implementation of these criteria into the design of ADMS, public sector institutions could take the following steps to kick-off the process:

1 Start early

The implementation of design criteria needs to start at the very beginning of any ADMS project. There is no point in time too early to consider the wider societal impacts of a planned application.

2 Build awareness

It is crucial for institutions to be aware of the societal impacts of technology on any given community it is working for or within. It is therefore important to expand awareness of the impact of automation and algorithmic technologies by encouraging mutual dialogue between technology specialists and the public sector. Ensuring system users are aware of the relevance of the ADMS, as well as developing tools to manage its potential negative impacts on society, need to be considered by all.

3 Develop a critical mindset

Building on public awareness, it is vital to establish a realistic and critical approach towards understanding the social and political impact of ADMS. This should not translate as cynicism or opposition to the use of ADM, but rather nurture a realistic awareness that those systems are not perfect and mistakes may happen.

4 Have an overview

At the beginning of the project it is crucial to establish an overview of the potential ethical and social impact of the ADMS planned. For that, the social relevance and criticality, i.e. the impact relevance, should be evaluated.

5 Identify a first step

Based on the impact relevance analysis, identify which area or question is most relevant to your project and where you can easily implement measures to realize design criteria. Develop a roadmap for your specific application, taking into account existing public sector regulation and tools at your disposal.

6 Exchange ideas

The benefit of ADMS for the public sector is ultimately to serve society. Therefore, the public sector needs to exchange ideas with the private sector and academia, but especially with civil society and groups representing people affected by the use of ADMS. This exchange can also help bridge competency gaps and find avenues of cooperation.

Having developed a roadmap and identified the project-specific challenges, the implementation of design criteria will always be context-specific. Their design depends on: the impact relevance of the ADMS; existing regulation like GDPR; Anti Discriminations Laws, and other socio-technical imperatives. It also depends on the technical possibilities and tools available: As an example, tools that ensure the intelligibility of machine learning systems such as Explainable AI, are still in development and research is constantly carried out to develop new methods. Once these methods become available, requirements for the design of ADMS will adjust accordingly, since new tools can help to increase the intelligibility of ADMS. It is crucial for the evolution of ADMS, that the implementation of design criteria—in line with the Algo.Rules—needs to be dynamic and adjust to the current state-of-the-art. This will help leverage the potentials of ADMS while minimizing societal risks.

References

Andrés-Pueyo, A., Arbach-Lucioni, K. and Redondo, S. (2017). The RisCanvi. In: Handbook of Recidivism Risk/Needs Assessment Tools. John Wiley & Sons. [online] Available at: https://doi.org/10.1002/9781119184256.ch13

Angwin, J., Larson, J., Mattu, S. and Kirchner, L. (2016). Machine Bias. ProPublica. [online] Available at: https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

Bertelsmann Stiftung and iRights.Lab. (2019). Algo.Rules: Rules for the Design of Algorithmic Systems. [online] Available at: https://algorules.org/en/home/

B•Debate. (2017). Barcelona Declaration for the Proper Development and Usage of Artificial Intelligence in Europe. [online] Available at: https://www.bdebate.org/sites/default/files/barcelona-declaration_v7-1-eng.pdf

Borum, R., Bartel, P. and Forth, A. (n.d.). Structured Assessment of Violence Risk in Youth (SAVRY). [online] Available at: https://www.parinc.com/Products/Pkey/390#details

Carahsoft & UiPath. (2019). Unlocking RPA's Potential in Government. [PDF] Available at: https://www.uipath.com/hubfs/resources/documents/PDFs/WP-Unlocking%20RPAs%20Potential%20in%20Government.pdf

Cummings, M. (2012). Automation Bias in Intelligent Time Critical Decision Support Systems. Cambridge : Massachusetts Institute of Technology.

Datenethikkommission. (2019). Gutachten der Datenethikkommission. [PDF] Available at: https://www.bmi.bund.de/SharedDocs/downloads/DE/publikationen/themen/it-digitalpolitik/gutachten-datenethikkommission.pdf?__blob=publicationFile&v=6

Digital Future Society. (2019). Where emerging tech meets government. [PDF] Available at: https://digitalfuturesociety.com/app/uploads/2019/08/062719_Where_emerging_tech_meets_government_DFS_mwcapital_DIGITAL.pdf

Hagendorff, T. (2019). The Ethics of AI Ethics: An Evaluation of Guidelines. [online] Available at: http://arxiv.org/abs/1903.03425

High-Level Expert Group on AI. (2019). Ethics guidelines for trustworthy AI. [online] Available at: https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai

Holl, J., Kernbeiß, G. and Wagner-Pinter, M. (2018). Das AMS-Arbeitsmarktchancen-Modell. [PDF] Available at: http://www.forschungsnetzwerk.at/downloadpub/arbeitsmarktchancen_methode_%20dokumentation.pdf

Jobin, A., Ienca, M. and Vayena, E. (2019). Artificial Intelligence: the global landscape of ethics guidelines. Nat Mach Intell 1. [online] Available at: https://doi.org/10.1038/s42256-019-0088-2

Krafft, T. and Zweig, K. (2019). Transparenz und Nachvollziehbarkeit algorithmenbasierter Entscheidungsprozesse: Ein Regulierungsvorschlag aus sozioinformatischer Perspektive. [PDF] Available at: https://www.vzbv.de/sites/default/files/downloads/2019/05/02/19-01-22_zweig_krafft_transparenz_adm-neu.pdf

Molavi, R. and Lahmann, H. (2018). Zukunft E-Government: Vorschläge für eine bürgerfreundliche und sichere Digitalisierung der Verwaltung. [PDF] Available at: https://irights-lab.de/wp-content/uploads/2018/12/bdr_studie_E_Government_RGB_V11_ES.pdf

Morley, J., Floridi, L., Kinsey, L. and Elhalal, A. (2019). From What to How: An Initial Review of Publicly Available AI Ethics Tools, Methods and Research to Translate Principles into Practices. [online] Available at: http://arxiv.org/abs/1905.06876

Nauwerck, G. and Cajander, Å. (2019). Automatic for the People: Implementing Robotic Process Automation in Social Work. [PDF] Available at: https://dl.eusset.eu/bitstream/20.500.12015/3289/1/ecscw2019_p04.pdf

Quijano-Sanchez, L., Liberatore, F. and Camacho-Collados, Miguel. (2018). Applications of Data Science in Policing: VeriPol an Investigation Support Tool. [online] Available at: https://bulletin.cepol.europa.eu/index.php/bulletin/article/view/328

Ranerup, A. and Henriksen, H. (2019). Value positions viewed through the lens of automated decision-making: The case of social services. Government Information Quarterly 36. [online] Available at: https://doi.org/10.1016/j.giq.2019.05.004

Rohde, N. (2018). Gütekriterien für algorithmische Prozesse: Eine Stärken- und Schwächenanalyse ausgewählter Forderungskataloge. [PDF] Available at: https://www.bertelsmann-stiftung.de/fileadmin/files/BSt/Publikationen/GrauePublikationen/GuetekriterienalgorithmischeProzesse29062018.pdf

Spielkamp, M. (ed.). (2019). Automating Society. [online] Available at: https://algorithmwatch.org/en/automating-society/

SVT. (2019). Data errors may have created tens of thousands of incorrect decisions at the Employment Service. [online] Available at: https://www.svt.se/nyheter/inrikes/svt-avslojar-stort-datafel-hos-arbetsformedlingen-tusentals-kan-ha-forlorat-ersattning

Treasury Board of Canada Secretariat. (2019). Algorithmic Impact Assessment (AIA). [online] Available at: https://www.canada.ca/en/government/system/digital-government/modern-emerging-technologies/responsible-use-ai/algorithmic-impact-assessment.html

United Nations Office on Disarmament Affairs (UNODA). (2017). Perspectives on Lethal Autonomous Weapon Systems. [PDF] Available at: https://www.unog.ch/80256EDD006B8954/s)/6866E44ADB996042C12581D400630B9A/$file/op30.pdf

USCIS (2018). Meet Emma, Our Virtual Assistant. [online] Available at: https://www.uscis.gov/emma

Utrecht Data School. (2019). Data Ethics Decision Aid (DEDA). [online] Available at: https://dataschool.nl/deda/?lang=en

Vieth, K. and Wagner, B. (2017). Calculated participation: How algorithmic processes impact opportunities to be part of society. [PDF] Available at: https://ethicsofalgorithms.org/wp-content/uploads/sites/10/2018/02/Teilhabe-ausgerechnet-ENG_final.pdf

Wills, T. (2019). Sweden: Rogue algorithm stops welfare payments for up to 70,000 unemployed. AlgorithmWatch. [online] Available at: https://algorithmwatch.org/en/rogue-algorithm-in-sweden-stops-welfare-payments/

Wisterberg, E. (2018). Titta! Kommunens robot avgör vem som får socialbidrag... . Breakit. [online] Available at: https://www.breakit.se/artikel/10770

Acknowledgements

Digital Future Society is a transnational initiative of Mobile World Capital Barcelona and Tercera Vicepresidencia del Gobierno de España - Ministerio de Asuntos Económicos y Transformación Digital

This report has been made possible thanks to the work of the research team at iRights.Lab, comprised of the following members:

- **Ramak Molavi Vasse'i (author)** – Legal Policy Advisor
- **Michael Puntschuh (author)** – Policy Advisor
- **Wiebke Glässer (editor)** – Head of Operations
- **Philipp Otto (editor**) – Director


Digital
Future Society