Understanding algorithmic decision-making: Opportunities and challenges

STUDY

Panel for the Future of Science and Technology

EPRS | European Parliamentary Research Service

Scientific Foresight Unit (STOA)

PE 624.261 – March 2019

EN

Understanding algorithmic decision-making: Opportunities and challenges

While algorithms are hardly a recent invention, they are nevertheless increasingly involved in systems used to support decision-making. These systems, known as 'ADS' (algorithmic decision systems), often rely on the analysis of large amounts of personal data to infer correlations or, more generally, to derive information deemed useful to make decisions. Human intervention in the decision-making may vary, and may even be completely out of the loop in entirely automated systems. In many situations, the impact of the decision on people can be significant, such as access to credit, employment, medical treatment, or judicial sentences, among other things. Entrusting ADS to make or to influence such decisions raises a variety of ethical, political, legal, or technical issues, where great care must be taken to analyse and address them correctly. If they are neglected, the expected benefits of these systems may be negated by a variety of different risks for individuals (discrimination, unfair practices, loss of autonomy, etc.), the economy (unfair practices, limited access to markets, etc.), and society as a whole (manipulation, threat to democracy, etc.).

This study reviews the opportunities and risks related to the use of ADS. It presents policy options to reduce the risks and explain their limitations. We sketch some options to overcome these limitations to be able to benefit from the tremendous possibilities of ADS while limiting the risks related to their use. Beyond providing an up-to-date and systematic review of the situation, the study gives a precise definition of a number of key terms and an analysis of their differences to help clarify the debate. The main focus of the study is the technical aspects of ADS. However, to broaden the discussion, other legal, ethical and social dimensions are considered.

AUTHORS

This study has been written by Claude Castelluccia and Daniel Le Métayer (Institut national de recherche en informatique et en automatique - Inria) at the request of the Panel for the Future of Science and Technology (STOA) and managed by the Scientific Foresight Unit within the Directorate-General for Parliamentary Research Services (DG EPRS) of the Secretariat of the European Parliament.

ADMINISTRATOR RESPONSIBLE

Mihalis Kritikos, Scientific Foresight Unit (STOA)

To contact the publisher, please e-mail STOA@ep.europa.eu

Acknowledgments

The authors would like to thank all who have helped, in any way whatever, in this study, in particular Irene Maxwell and Clément Hénin for their careful reading and useful comments on an earlier draft of this report.

LINGUISTIC VERSION

Original: EN

Manuscript completed in March 2019.

PE 624.261
ISBN: 978-92-846-3506-1
doi: 10.2861/536131
QA-06-18-337-EN-N

http://www.europarl.europa.eu/stoa (STOA website)
http://www.eprs.ep.parl.union.eu (intranet)
http://www.europarl.europa.eu/thinktank (internet)
http://epthinktank.eu (blog)

Executive Summary

Scope of the study: While algorithms are hardly a recent invention, they are nevertheless increasingly involved in systems used to support decision-making. Known as 'ADS' (algorithmic decision systems), ADS often rely on the analysis of large amounts of personal data to infer correlations or, more generally, to derive information deemed useful to make decisions. Human intervention in the decision-making may vary, and may even be completely out of the loop in entirely automated systems. In many situations, the impact of the decision on people can be significant, such as on access to credit, employment, medical treatment, judicial sentences, among other things. Entrusting ADS to make or to influence such decisions raises a variety of different ethical, political, legal, or technical issues, where great care must be taken to analyse and address them correctly. If they are neglected, the expected benefits of these systems may be negated by the variety of risks for individuals (discrimination, unfair practices, loss of autonomy, etc.), the economy (unfair practices, limited access to markets, etc.), and society as a whole (manipulation, threat to democracy, etc.).

This study reviews the opportunities and risks related to the use of ADS. It presents existing options to reduce these risks and explain their limitations. We sketch some options to benefit from the tremendous possibilities of ADS while limiting the risks related to their use. Beyond providing an up-to-date and systematic review of the situation, the study gives a precise definition of a number of key terms and an analysis of their differences to help clarify the debate. The main focus of the study is the technical aspects of ADS. However, to broaden the discussion, other legal, ethical and social dimensions are considered.

ADS opportunities and risks: The study discusses the benefits and risks related to the use of ADS for three categories of stakeholders: individuals, the private sector and the public sector. Risks may be intentional (e.g. to optimise the interests of the operator of the ADS), accidental (side-effects of the purpose of the ADS, with no such intent from the designer), or the consequences of ADS errors or inaccuracies (e.g. people wrongly included in blacklists or 'no fly' lists due to homonyms or inaccurate inferences).

Opportunities and risks of ADS for individuals: ADS may undermine the fundamental principles of equality, privacy, dignity, autonomy and free will, and may also pose risks related to health, quality of life and physical integrity. That ADS can lead to discrimination has been extensively documented in many areas, such as the judicial system, credit scoring, targeted advertising and employment. Discrimination may result from different types of biases arising from the training data, technical constraints, or societal or individual biases. However, the risk of discrimination related to the use of ADS should be compared with the risk of discrimination **without the use of ADS**. Humans have their own sources of bias that can affect their decisions and, in some cases, these could be detected or avoided using ADS.

The deployment of ADS may also pose a threat to privacy and data protection in many different ways. The first is related to the massive collection of personal data required to train the algorithms. Even when no external attack has been carried out, the mere suspicion that one's personal data is being collected and possibly analysed can have a detrimental impact on people. For example, several studies have provided evidence of the chilling effect resulting from fear of online surveillance. Altogether, large-scale surveillance and scoring could narrow the range of possibilities and choices available to individuals, affecting their capacity for self-development and fulfilment. Scoring also raises the fear that humans are increasingly treated as numbers, and reduced to their digital profile. Reducing the complexity of human personality to a number can be seen as a form of alienation and an offence to human dignity.

The opacity or lack of transparency of ADS is another primary source of risk for individuals. It opens the door to all kinds of manipulation and makes it difficult to challenge a decision based on the result of an ADS. This is in contradiction with defence rights and the principle of adversarial proceedings with right to all evidence and observations in most legal systems. The use of ADS in court also raises far-reaching questions about the reliance on predictive scores to make legal decisions, in particular for sentencing.

Opportunities and risks for the public sector: ADS are currently being used by state and public agencies to provide new services or improve existing ones in areas such as energy, education, healthcare, transportation, the judicial system and security. Examples of applications of ADS in this context are predictive policing, smart metering, video protection and school enrolment. They can also contribute to improving the quality of healthcare, education and job skill training. They are increasingly used in cyber-defence to protect infrastructures and to support soldiers in the battlefield. ADS, or smart technologies in general, such as mobility management tools, or water and energy management systems, can improve city management efficiency. They can also help make administrative decisions more efficient, transparent and accountable, provided however that they themselves are transparent and accountable.

ADS also create new 'security vulnerabilities' that can be exploited by people with malicious intent. Since ADS play a pivotal role in the workings of society, for example in nuclear power stations, smart grids, hospitals and cars, hackers able to compromise these systems have the capacity to cause major damage. Furthermore, ADS such as those used for predictive policing, may become overwhelming and oppressive. ADS can be (mis)used by states to control people, for example by identifying political opponents. More generally, interest groups or states may be tempted to use these technologies to control and influence citizen behaviour. These technologies can also be used to distort information to damage the integrity of democratic discourse and the reputation of the government or political leaders.

Opportunities and risks for the private sector: The opportunities presented by ADS for the private sector are endless, but there are also numerous risks. Any task that is repetitive, pressured by time, or that could benefit from the analysis of high volumes of data, is a prime target for ADS. Such tasks concern low-skilled as well as highly-skilled personnel, for example in sectors such as banking, insurance or justice. Certain types of jobs will change enormously or be eliminated, whilst new ones will appear. The expression the 'fourth industrial revolution' describes this dramatic change.

Desiderata for algorithms: We define the key properties required to reduce the risks related to ADS by making a distinction between properties that apply to any algorithmic system, such as safety, security or privacy, and properties specific to ADS. The latter include intrinsic and extrinsic requirements.

Intrinsic requirements, such as fairness, absence of bias or non-discrimination, can be expressed as properties of the algorithm itself in its application context. We equate 'fairness' with 'absence of undesirable bias' and we characterise 'discrimination' as a particular form of unfairness related to the use of specific types of data (such as ethnic origin, political opinions, gender, etc.).

As far as extrinsic requirements are concerned, we define 'understandability' as the possibility to provide understandable information about the link between the input and the output of the ADS.

The two main forms of understandability considered are transparency and explainability:

- **Transparency** is defined as the availability of the ADS code with its design documentation, parameters and the learning dataset when the ADS relies on machine learning (ML). Transparency does not necessarily mean availability to the public. It also encompasses cases in which the code is disclosed only to specific actors, for example for audit or certification.

- **Explainability** is defined as the availability of explanations about the ADS. In contrast to transparency, explainability requires the delivery of information beyond the ADS itself. Explanations can be of different types (operational, logical or causal); they can be either global (about the whole algorithm) or local (about specific results); and they can take different forms (decision trees, histograms, picture or text highlights, examples, counterexamples, etc.). The strengths and weaknesses of each explanation mode should be assessed in relation to the recipients of the explanation (e.g. professional or individual), their level of expertise, and their objectives (to challenge a decision, take actions to obtain a decision, verify compliance with legal obligations, etc.).

- **Accountability** is another key desideratum often put forward in the context of ADS. In accordance with previous work in this area, we see accountability as an overarching principle characterised by the obligation to justify one's actions and the risk of sanctions if justifications are inadequate. Accountability can therefore be seen as a requirement on a process (obligation to provide justification), which applies to both intrinsic and extrinsic requirements for ADS (each case corresponding to specific types of 'justification').

- **Technical issues and approaches:** The report includes a review of some of the technical issues and available solutions. **Safety:** is an important issue to consider, especially when ADS are embedded in physical systems whose failure may cause fatal damage. The study explores several types of accidents related to machine learning and presents relevant research and directions to protect against them. While many ADS failures can be addressed with ad-hoc solutions, there is a strong need to define a unified approach to prevent ADS from causing unintended harm. A minimum requirement should be to perform extensive testing and evaluation before any large-scale deployment. It is also important to provide accountability, including the possibility of independent audits and to ensure a form of human oversight.

- **Integrity and availability:** Increasingly, ADS will be used in critical contexts. It is therefore important to guarantee that they are secure against malicious adversaries. ADS should not jeopardise integrity and availability. Since most ADS rely heavily on machine learning algorithms, it is important to consider their security properties in the context of these algorithms. Adversaries can threaten the integrity or availability of ADS in different ways, i.e., by polluting training datasets with fake data, attacking the machine learning (ML) algorithm itself or exploiting the generated model (the ADS) at run-time. We argue that existing protection mechanisms remain preliminary and require more research.

- **Confidentiality and privacy:** An adversary may seek to compromise the confidentiality of an ADS. For example, they may try to extract information about the training data or retrieve the ADS model itself. These attacks raise privacy concerns as training data is likely to contain personal data. They may also undermine intellectual property since the ADS model and the training data may be proprietary and confidential to the owner. Different proposals have been made to address these privacy attacks. Some of them involve anonymising the training datasets and the generated models i.e. designing privacy-preserving ADS. Other proposals rely on the distribution of the learning phase, so that the training data does not leave the device which collects them. These privacy-preserving solutions are still in their infancy and require more work.

- **Fairness (absence of undesirable bias):** ADS are often based on machine learning algorithms that are trained using collected data. This process includes multiple potential sources of unfairness. Unfair treatment may result from the content of the training data, the way the data is labelled or the feature selection. As shown in this study, there are different definitions of fairness, and others will be proposed in the future. Research has shown however that many definitions of fairness are actually incompatible. Several research groups

have also started to work on the design of 'fair' ADS. This study introduces some of these new projects.

Explainability: Three main approaches can be followed to implement the requirements of explainability:

- **The black box approach:** this approach analyses the behaviour of the ADS without 'opening the hood', i.e. without any knowledge of its code. Explanations are constructed from observations of the relationships between the inputs and outputs of the system. This is the only possible approach when the operator or provider of the ADS is uncollaborative (does not agree to disclose the code). Examples of this category of approach include LIME (local interpretable model-agnostic explanations), Anchor, TREPAN, AdFischer and Sunlight.
- **The white box approach:** in contrast to the black box approach, this approach assumes that analysis of the ADS code is possible. An example of early work in this direction is the Elvira system for the graphical explanation of Bayesian networks. Other solutions based on neural networks have been proposed more recently.
- **The constructive approach:** in contrast to the first two approaches, which assume that the ADS already exists, the constructive approach is to design ADS taking explainability requirements into account ('explainability by design'). Two options are possible to achieve explainability by design: (1) relying on an algorithmic technique which, by design, meets the intelligibility requirements while providing sufficient accuracy, or (2) enhancing an accurate algorithm with explanation facilities so that it can generate, in addition to its nominal results (e. g. classification), a faithful and intelligible explanation for these results.

The explanations generated by these methods can take very different forms. A number of criteria have been proposed to evaluate their quality, including intelligibility, accuracy, precision, completeness and consistency. However, some of these criteria may be in tension with each other. For example, higher levels of accuracy and precision may reduce intelligibility. In addition, their evaluation is a difficult (and often partly subjective) task.

Legal instruments: Technical solutions are necessary but cannot solve all the issues raised by ADS by themselves. They must be associated with other types of measures and in particular legal requirements for transparency, explainability or accountability. In fact, various existing laws already apply to ADS and can, to a greater or lesser extent, address some of the requirements identified above. In this report, we first discuss the situation in Europe with the new General Data Protection Regulation (GDPR). In particular we analyse the highly-debated provisions of the GDPR in terms of transparency or explainability, and discuss possible answers to questions such as: Is such a right really set forth in the GDPR and, if so, what does 'explanation' mean exactly, in this context? If this right is set forth in the GDPR, what are the conditions for its application and is it likely to be effective? We also discuss recent developments in European Union Member State France, before sketching proposals that originate in the United States of America. These proposals, which stem from the legal doctrine, emphasise due process and accountability as the most effective way to introduce a form of control over ADS.

Open questions and remaining challenges: This study presents some of the many challenges to be addressed to reduce the risks related to ADS, classified according to the following three perspectives: (1) ethical and political, (2) legal and social and (3) technical challenges.

Ethical and political:

ADS exacerbate existing problems or force us to rethink issues such as discrimination, but they also introduce new ethical questions that are very difficult to address. Examples of critical and complex questions raised by ADS include:

- Legitimate use of an ADS: in certain contexts, such as evidence-based sentencing or lethal weapons, their use has been heavily criticised, but establishing clear and firm boundaries between acceptable uses of ADS and situations in which they should be banned is far from straightforward.
- Beyond existing fairness criteria already identified in anti-discrimination laws, what types of treatment should be considered undesirable? Where should the line be drawn and in relation to which principles?
- How can online manipulation be characterised and distinguished from (acceptable) influence or 'nudging'?
- In which cases should transparency, explainability or other forms of accountability be required and in relation to which underlying principles? Should certain types of ADS be forbidden when an acceptable level of transparency, explainability or accountability cannot be achieved (for example in court, or to support medical diagnosis)?
- What choices of design should be made for autonomous vehicles when a 'life or death' decision has to be taken? Should ethical behaviours be encoded in the system and, if so, what should they be and who should decide upon the choice of 'ethical behaviour'?

The study sketches some proposals to address these issues in a principled way.

Legal and social:

Ethical and political debates are prerequisites for further action. Assuming that a fairly broad agreement is reached on some of the issues discussed above, the next step is to decide upon the most appropriate instruments to implement that agreement. In law, we discuss different types of regulation (state regulation, self-regulation or co-regulation, hard law or soft law, general or sectorial regulation) and different modes of enforcement (regulatory agencies, dedicated oversight bodies, etc.). We also discuss different options for certification.

Technical:

The technical instruments presented in this study are useful to meet the identified desiderata, but are still in their infancy, with a number of challenges that need be addressed. Some of these challenges are 'conceptual', such as defining the best types of explanations depending on the different recipients, their level of expertise and objectives. Other challenges are 'operational', such as the implementation of explainability by design, fairness by design or privacy by design. These properties should be taken into consideration from the beginning of the conception of an ADS, as already required by the GDPR for data protection. However, this phase requires a strong level of technical expertise that cannot be expected from all ADS developers. Providing guidance and assistance to designers and developers to help them implement these principles remains an open challenge.

Options: Based on existing studies and the present analysis, we put forward the options listed below. These options are mostly organisational or procedural (in the general sense of the term), rather than substantive, as positions on this matter should rather result from public debate than be issued by expert groups. We do however provide guidance on criteria and issues that should be carefully considered before the adoption of ADS. We distinguish five complementary types of options for ADS:

1. **Development and dissemination of knowledge:** ADS raise complex questions that are not entirely understood by experts, not to mention users, or the people affected by them. It is therefore of prime importance to develop interdisciplinary research in ADS. More research is needed, for example, on ADS security, safety, privacy, fairness or explainability. In addition,

philosophers, experts in ethics, social scientists, lawyers, computer scientists and AI experts should work together to develop further conceptual tools to analyse ethical issues raised by ADS. A key condition to facilitate this research is the possibility to provide the research community with access, under specific conditions and the strictest confidentiality, to datasets held not only by public entities but also by private companies. This access right is justified by the fact that such large amounts of data can be considered 'data of public interest'. For the same reason, it should be made clear that reverse engineering for the purpose of analysing, explaining or detecting biases in ADS should be considered lawful and should not be limited by trade secret, or more generally by intellectual property rights laws.

2. **Public debate about the benefits and risks:** Considering that ADS can have a major impact on society, they must be subject to public debate. Several conditions have to be met to ensure the quality of this debate. It must involve all stakeholders, opinions and interests, which means experts of all disciplines, policy-makers, professionals, NGOs and the general public. It must be conducted in a rigorous fashion and without overshadowing any of the key issues, including the preliminary question of the legitimacy of the use of an ADS in the context being examined.

3. **Adapt legislation to enhance accountability:** Different types of legal instruments can be used to enhance the accountability of ADS. Considering that technology and its use evolve very quickly in this area, it is wise to avoid hasty legislation that could create more problems than it solves. New regulation should be enacted only when the matter has been properly understood, the recommended public debate has taken place and it is established that existing laws are insufficient to address the identified issues. It may be the case that certain sectors require further regulation or clarification on the application of existing laws. As far as enforcement is concerned, we believe that a clear distinction should be made between (1) ethical committees, with the mission to stimulate discussion, to conduct debates and publish recommendations; and (2) operational bodies, such as accreditation bodies, certification agencies and oversight agencies who together, provide a framework for the monitoring, certification and oversight of specific ADS. Oversight agencies should also have the power to sanction operators of non-compliant ADS. Ethical committees can operate at a general (cross-sector) level, while operational bodies should be sectoral because different application areas raise different issues and have different histories, cultures, sets of practices and regulations.

4. **Development of tools to enhance accountability:** Most ADS designers and developers are not experts in privacy, security, fairness or explainability. It is therefore important to provide tools and methodologies to help them reconcile the tensions that exist between accuracy, cost and explainability/fairness/privacy. Recommendation guides are not enough. Tools and methodologies that consider the entire development cycle of ADS should be developed and disseminated. Similarly, frameworks, composed of metrics, methodologies and tools that assess the impact of an ADS and test its desired properties should be developed. These frameworks could be used by designers to test their ADS, and by third-party entities, such as certification authorities, to validate them. As far as users are concerned, better explanation facilities are required, in particular, more interactive interfaces and dialogue models.

5. **Effective validation and monitoring measures:** The GDPR introduces an obligation for data controllers to conduct data protection impact assessments (DPIA) and encourages certification mechanisms. Considering the high stakes involved in ADS, there is no reason why they should not be subject to the same types of precautions. We recommend in particular that: (1) ADS should not be deployed without a prior algorithmic impact assessment (AIA) unless it is clear they have no significant impact on individuals lives; and (2) the certification of ADS should be encouraged and even mandatory in certain sectors.

Conducting an AIA is not an easy task and models and tools should be proposed to make it easier. The report presents some key issues which should be considered in an AIA: (1) legitimacy of the ADS, including the legitimacy of its purpose, techniques and parameters; (2) qualities of the ADS; and (3) integration of the ADS within the human environment. It should also be clear that AIA should not only focus on the risks of **using** an ADS: they should also assess the risks of **not using** an ADS. In other words, AIA should consider both benefits and risks. Finally, certifications and labels, if properly implemented, can be a way to enhance trust in ADS and to verify that they comply with certain rules (such as the absence of bias or discrimination). We believe that certification requirements and obligations should be sectoral. Indeed, the needs and the risks vary greatly from one type of application to another and sectoral supervisory authorities or agencies are in a better position to define reference evaluation criteria and to control their application. For the deployment of ADS, certification can be on either a voluntary basis (as encouraged by the GDPR), or mandatory in certain areas such as justice and healthcare.

Conclusion: In the conclusion to this study, we revisit the study's objectives and put them into perspective. We argue that transparency should not be seen as the ultimate solution for users or people affected by the decisions of an ADS since source code is illegible to non-experts. Transparency mainly benefits independent experts, NGOs, evaluation bodies or data protection authorities (DPA), to audit and certify ADS for example. 'Explainability' is shown to have different meanings and the needs vary considerably according to the audience. Designers, developers, users or affected people do not need the same level and type of explanation. It is also important to note that the requirements for explainability vary from one ADS to another, according to the potential impact of the decisions made and whether the decision-making process is fully automated. Although transparency and explainability are essential to reduce the risks related to ADS, we argue that accountability is the most important requirement as far as the protection of individuals is concerned. In fact, transparency and explainability may allow for the discovery of deficiencies, but do not provide absolute guarantees for the reliability, security or fairness of an ADS. Accountability can be achieved via complementary means such as AIAs, auditing and certification. The main virtue of accountability is to put the onus on the providers or operators of the ADS to demonstrate they meet expected requirements. It cannot provide an absolute guarantee either, but if certification is rigorous and audits are conducted on a regular basis, potential issues can be discovered and corrective measures taken. In addition, if sanctions are significant enough, an accountability approach provides strong incentives for ADS providers to carefully design their system. In this perspective, oversight agencies and supervisory authorities should play a central role and it is critical that they have all the means necessary to carry out their tasks. These means go beyond funding and expertise. They should include the right to access and analyse the details of the ADS, including their source code and, if necessary, the training data.

Finally, we believe that if appropriate accountability measures are taken, in certain situations ADS have the potential to improve transparency and reduce unfairness and discrimination. Another benefit of using ADS, and one that can already be observed, is the fact that they put decisions at the front and centre of public debate. Decisions that, up to now, had been taken far out of citizens' sight.

Table of contents

1. Introduction

While algorithms are hardly a recent invention, they are nevertheless increasingly involved in systems used to support decision making. Known as 'ADS' (**algorithmic decision systems**), these systems often rely on the analysis of large amounts of personal data to infer correlations or, more generally, to derive information deemed useful to make decisions. Human intervention in the decision-making may vary, and may even be completely out of the loop in entirely automated systems. In many situations, the impact of the decision on people can be significant, such as: access to credit, employment, medical treatment, judicial sentences, etc. Entrusting ADS to make or to influence such decisions raises a variety of different ethical, political, legal, or technical issues, where great care must be taken to analyse and address them correctly. If they are neglected, the expected benefits of these systems may be counterbalanced by the variety of risks for individuals (discrimination, unfair practices, loss of autonomy, etc.), the economy (unfair practices, limited access to markets, etc.) and society as a whole (manipulation, threat to democracy, etc.).

Different requirements such as transparency, explainability, data protection and accountability are often presented as ways to limit these risks but they are generally ill-defined, seldom required by law, and difficult to implement.

1.1. Objectives

This study reviews the opportunities and risks related to the use of ADS. We present existing options to reduce the risks and explain their limitations. We sketch some options to overcome these limitations to be able to benefit from the tremendous possibilities of ADS while limiting the risks related to their use. Beyond providing an up-to-date and systematic review of the situation, the report gives a precise definition of a number of key terms and an analysis of their differences. This helps clarify the debate. The main focus of the report is the technical aspects of ADS. However, to broaden the discussion, other legal, ethical and social dimensions are considered.

1.2. Methodology and resources

The methodology that was followed in preparing this document is based on traditional literature review including:

- All types of scientific literature for technical aspects. This includes articles published in peer-reviewed scientific journals or conference proceedings, surveys, books, science magazines and also papers published as reports or pre-print papers in scientific repositories.
- Reports, recommendations or studies published by (or for) governmental agencies, ethical committees, data protection authorities (DPA), NGOs or think tanks.
- General literature, including newspapers, magazines and web sites for information on the actual use of ADS, their benefits, risks and social acceptance.

A distinctive feature of the domain covered in this study is that it is not only rapidly evolving on the technical side but also in terms of its deployment and impact on society. This influenced the choice to analyse a wide variety of sources.

In addition, key issues such as explainability and, to a less extent, fairness have not received enough attention from the research community in the past. Interest in these topics in different research communities (including AI, computer science and law) is increasing dramatically, due to the development of ADS. As a result, a large part of the work on this topic is rather recent and has often not yet appeared in peer-reviewed journals. As an illustration, the first edition of the XAI (eXplainable

Artificial Intelligence) Workshop co-located with the flagship artificial conference, IJCAI,[1] took place in 2017, the annual Workshop on Human Interpretability in machine learning (WHI) was initiated in 2016, and the FAT/ML workshop on 'Fairness, Accountability and Transparency in Machine Learning' was launched in 2014.

On the social and political side, a plethora of reports, recommendations and guidelines have been published by various committees and agencies. Most of these reports rightly alert citizens and policy-makers about the potential risks posed by artificial intelligence, but they generally focus on the societal aspects and do not discuss the technical dimensions.

One of the goals of this study is to try to bridge precisely this gap by:

1. First studying the actual or future uses of ADS and the associated opportunities and risks.
2. Analysing and defining in a precise manner the main requirements of ADS that could reduce these risks.
3. Studying the technical and legal approaches to meet the aforementioned requirements.
4. Analysing the limitations of these approaches and providing policy options to address them.

1.3. Document structure

Chapter 2 defines the scope and objectives of the study. After introducing some key definitions, we provide examples of ADS and categorise them according to three classes, which correspond to different objectives and stakes. Chapter 3 analyses the opportunities and risks related to the use of algorithms. We consider in succession the opportunities and risks for individuals (Section 3.1), the public sector (Section 3.2) and the private sector (Section 3.3). Chapter 4 defines the desired properties of ADS to reduce the risks identified in Chapter 3. Many terms, such as transparency, explainability, interpretability and accountability, are often used with different meanings in this context. For the sake of clarity, we give a precise definition of the notions considered in the study and compare them with their previous use in the literature. Chapter 5 reviews the technical issues and solutions available to meet the desiderata presented in Chapter 4. Sections 5.1, 5.2 and 5.3 focus respectively on safety, security and privacy, while Sections 5.4 and 5.5 are devoted to fairness and explainability respectively. Chapter 6 discusses more briefly the legal instruments to enhance explainability, privacy and accountability. The instruments presented in Chapter 5 and Chapter 6 are useful but far from sufficient to address all the challenges raised by ADS. ADS raise substantive issues that are not yet fully understood and that need to be thoroughly analysed and debated. Chapter 7 analyses the main existing challenges from different (and complementary) perspectives: ethical, political, legal, social and technical. Finally, Chapter 8 presents our proposed options to address these challenges and concludes the study.

[1] International Joint Conference on Artificial Intelligence.

2. Scope and objectives

This study focuses on the use of algorithmic systems to support decision-making. In practice, this use can occur in different situations, with different types of impact. For example, decisions may or may not be automatic, whether and how the system is used may be decided by the affected persons or imposed upon them, they may or may not be aware of the existence of the system, etc.

In general, we distinguish three types of stakeholders: the **designers** of the algorithmic system, the **operators** or **users** (professionals or individuals) and the **affected persons**. In certain situations, different roles can be played by the same person (for example, the users of recommendation systems are also the affected persons).

Technically speaking, decision-making algorithms also vary: they can rely on 'standard' algorithms or on machine learning and they may involve a different models such as decision trees, Bayesian networks, neural networks, etc.

Decision-making algorithms are increasingly used in areas such as access to information, e-commerce, recommendation systems, employment, health, justice, policing, banking and insurance. They can provide great benefits for individuals and for organisations, both in the public and the private sectors. For example, they can lead to better informed decisions, to the discovery of previously unknown correlations, to better patient treatment etc. However, they also give rise to a variety of risks, such as discrimination, unfairness, manipulation or privacy breaches.

The objective of this report is to assess the actual and potential extent of not only the current use of algorithms in decision-making and their respective risks and opportunities, but also their future use. The report equally assesses the potential solutions to overcome these risks. The report emphasises the need to scrutinise the use of algorithms for decision-making and whether algorithmic decision-making can be done in a transparent and accountable way.

Whilst the main focus of the report is on the technical aspects, to broaden the discussion, legal, ethical and social dimensions are considered.

2.1. Some definitions

Algorithmic systems refer to a wide range of applications and techniques. We hereby consider the following key concepts and definitions.

Algorithm: An algorithm is an unambiguous procedure to solve a problem or a class of problems. It is typically composed of a set of instructions or rules that take some input data and return outputs. As an example, a sorting algorithm can take a list of numbers and proceed iteratively, first extracting the largest element of the list, then the largest element of the rest of the list, and so on, until the list is empty. Algorithms can be combined to develop more complex systems, such as web services or autonomous cars. An algorithm can be hand-coded, by a programmer, or generated automatically from data, as in machine learning.

Algorithmic decision system (ADS): In the study, we focus on a specific type of algorithm aimed at supporting decision-making. We use the generic expression 'algorithmic decision system' (ADS) to stress the fact that these algorithms should be studied in a general setting that includes their parameters, context of use and, if they rely on machine learning, their training data. ADS, whether based on machine learning or not, usually rely on the analysis of a variety of data. They may assume varying degrees of human involvement. Semi-automatic ADS assist humans in making decisions. For example, ADS can assist doctors in identifying diseases in a clinical setting, where data is

complex and sparse, and help them to make diagnoses.[2] ADS can also be used to take fully automated decisions, as in automated metro systems. Very often, they are used to make predictions or to estimate risks. A distinction is sometimes drawn between predictive and prescriptive ADS, but the frontier between the two categories is often fuzzy.

Artificial intelligence (AI): Although there is a lack of a precise, universally accepted definition of artificial intelligence (AI), it is usually conceived as the capacity for machines to resemble human intellectual abilities. Narrow, or weak, AI is designed to perform a specific task, such as facial recognition or product recommendation. General, or strong, AI aims at outperforming humans across multiple domains.[3]

Machine learning (ML): There are multiple definitions of machine learning. Andrew Ng defines it as 'the science of getting computers to act without being explicitly programmed'.[4] Machine learning is an AI component that provides systems with the ability to automatically learn over time, generally from large quantities of data. The learning process is based on observations or data, such as examples, in order to identify patterns in data and make better predictions. An ML algorithm can therefore be seen as an algorithm that, from data, generates another algorithm, usually referred to as a model. For example, the Amazon recommendation algorithm uses customers' profiles to learn which products are likely to be of interest to them. When users visit the Amazon site, the recommendation model built by the system uses their profiles to produce personalised recommendations. Machine learning is usually classified into three types:

1. **Supervised learning** relies on labelled data to train a model. This model is then used to predict for a given piece of data, that was not part of the training data, the corresponding label. It can be used to predict a continuous value (e.g. a score), through regression, or a discrete value (e.g. a word associated with a picture) through classification.

2. **Unsupervised learning** does not require labelled data. It automatically identifies patterns and structures from the training data, for example through clustering.

3. **Reinforced learning** relies on the exploitation of the feedback on success and failure received from its environment. In other words, it takes actions in an environment so as to maximise a reward function.

2.2. Some examples of ADS

There are many different ways to categorise ADS. In this report, we propose the following three classes, which correspond to the different objectives of ADS and what is at stake in using them:

- **ADS that aim at improving general knowledge or technology:** ADS in this class use algorithms to generate new knowledge, generally through the analysis of complex phenomena. Algorithms are crucial in this context since they can be used to analyse very large datasets to extract knowledge. They can, for example, help improve climate forecasts,[5] detect

[2] Zhe Li, Chong Wang, Mei Han, Yuan Xue, Wei Wei, Lia-Jia Li, Li Fei-Fei; Thoracic disease identification and localization with limited supervision. arXiv:1711.06373 [cs.CV] ; 2017 ; https://arxiv.org/abs/1711.06373.

[3] Peter Stone, Rodney Brooks, Erik Brynjolfsson, Ryan Calo, Oren Etzioni, Greg Hager, Julia Hirschberg, Shivaram Kalyanakrishnan, Ece Kamar, Sarit Kraus, Kevin Leyton-Brown, David Parkes, William Press, AnnaLee Saxenian, Julie Shah, Milind Tambe, Astro Teller; Artificial Intelligence and Life in 2030. One Hundred Year Study on Artificial Intelligence: Report of the 2015-2016 Study Panel; Stanford University; Doc: http://ai100.stanford.edu/2016-report.

[4] Daniel Faggella; What is Machine Learning?; 2017; https://www.techemergence.com/what-is-machine-learning/.

[5] Nicola Jones; How machine learning could help to improve climate forecasts; Nature (548); 2017; https://www.nature.com/news/how-machine-learning-could-help-to-improve-climate-forecasts-1.22503.

diseases[6] or discover new viruses.[7] These ADS are used to make decisions which have a global impact (or an impact on society) rather than on specific individuals.

- **ADS that aim at improving or developing new digital services:** Applications of this category are used to help make predictions, recommendations or decisions in various areas such as information, finance, planning, logistics, etc. These services aim at optimising one or several specific criteria, such as time, energy, cost, relevance of information, etc. For example, navigation services help users identify the 'optimal' route to their destination taking parameters such as the current traffic, cost and road conditions into account. New services, such as intermediary platforms, propose accommodation (AirBnB) or transportation alternatives (Uber) that did not exist a few years ago. Smart home applications are being deployed to improve comfort and optimise energy consumption. Similarly, quantified-self[8] or medical applications are proposed to help users improve their health (e.g. by monitoring their physical activities or eating habits). These services use a lot of data and complex algorithms or models. They may address individuals but also private and public services. For example, new services are being deployed to improve logistics (optimal product placement in stores, optimal road constructions, or the frequency of refuse collection), finance (real-time auctions) or security (automated detection of vulnerabilities in computer systems). ADS can be also used to 'optimise' existing services. In this context, decisions that were so far taken by humans are now performed with the assistance of, or directly by, ADS (for example in task allocation, recruitment or customer relationship management).

  COMPAS, or correctional offender management profiling for alternative sanctions, is an ADS, used by some US jurisdictions, that predicts a defendant's risk of committing further crimes. It works through a proprietary algorithm that considers some of the answers to a questionnaire.

- **ADS integrated within cyber physical systems:** Within this context, ADS are used to provide autonomy to physical objects by limiting human supervision. Examples are autonomous cars, robots or weapons. Autonomous cars are being experimented with all over the world. Algorithms should replace, or at least assist, users in the way they operate vehicles and should make decisions on behalf of 'drivers'. The goals are essentially to make roads safer and optimise connection times. Similarly, autonomous robots are being developed to help or replace humans in performing difficult physical tasks at work or in the home. Examples include robots used in factory chains, domestic robots that provide services to humans, or robots on the battlefield. A variety of autonomous weapons are under development to assist soldiers in action and to limit collateral damage.

Another way to look at ADS is to consider their users. ADS can be used by individuals, or by private or public organisations. Figure 1 presents some examples of ADS according to the two dimensions of objectives and users.

[6] Andre Esteva, Brett Kuprel, Roberto A. Novoa, Justin Ko, Sasan M. Swetter, Helen M. Blau, , Sebastian Thrun; Dermatologist-level classification of skin cancer with deep neural networks; Nature (2017).

[7] Amy Maxmen; Machine learning spots treasure trove of elusive viruses; Nature News; 2018; https://www.nature.com/articles/d41586-018-03358-3

[8] https://en.wikipedia.org/wiki/Quantified_self

Figure 1 – Examples of applications of ADS by objectives and types of users.

Users / Objectives	**Individuals**	**Private sector**	**Public sector**
Improvement of general Knowledge	N/A	Drugs discovery	Climate Weather forecast Environment Healthcare
Digital services	Quantified-self Finance Note taking Smart home Recommendations	Risk scoring Payment systems Targeting Personalised services	Predictive justice Predictive policing Hazard prediction Infrastructure development planning
Physical systems	Autonomous Cars Home Robots Security Personal assistants in the home	Autonomous robots	Autonomous weapons Defence Transport Smart cities Smart grids

3. Opportunities and risks related to the use of algorithms

In this chapter, we discuss the benefits and risks related to the use of ADS across the three categories of stakeholders referred to in the previous chapter: individuals, the private and the public sector. Note that risks may be intentional (e.g. to optimise the interests of the operator of the ADS), accidental (side-effects of the purpose of the ADS, without intent by the designer), or consequences of errors or inaccuracies by the ADS (e.g. people wrongly included in blacklists or 'no fly' lists due to homonyms or inaccurate inferences).

3.1. Opportunities and risks for individuals

The first category of stakeholders affected by the use of ADS are individuals, who may benefit from the use of ADS, but may also face a variety of undesirable consequences.

We distinguish three categories of opportunities and risks for individuals:

- Opportunities and risks related to the principle of equality.
- Opportunities and risks related to the principles of privacy, dignity, autonomy and free will.
- Opportunities and risks related to health, quality of life, wellbeing and physical integrity.

3.1.1. Opportunities and risks related to the principle of equality

Discrimination in a legal sense: Discrimination is often put forward as one of the primary risks related to the use of ADS. Considering that ADS are used to classify, rank, rate or produce any kind of useful result to inform the decision process, they are bound to discriminate, in the technical sense of making distinctions between people based on certain features. However, certain types of discrimination are undesirable, and even prohibited by law. Even though specific rules vary between countries, most regulations identify specific factors that must not have any impact on certain decisions. For example, Directive 2000/78/EC[9] lays down:

> 'a general framework for combating discrimination on the grounds of religion or belief, disability, age or sexual orientation as regards employment and occupation, with a view to putting into effect in the Member States the principle of equal treatment.'

In a similar vein, the Convention for the Protection of Human Rights and Fundamental Freedoms[10] provides that:

> 'the enjoyment of the rights and freedoms set forth in this Convention shall be secured without discrimination on any ground such as sex, race, colour, language, religion, political or other opinion, national or social origin, association with a national minority, property, birth or other status.'

The fact that ADS can lead to discrimination has been documented in many areas, such as the justice system, targeted advertisements and employment. It should be noted that these discriminations do not necessarily arise from deliberate choices: they may result from different types of bias, for example bias in training data (in which case, the algorithm reproduces and systematises already existing discriminations), societal or individual bias (e.g. of designers or programmers of the ADS),

[9] Council Directive 2000/78/EC of 27 November 2000 establishing a general framework for equal treatment in employment and occupation; https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex:32000L0078.

[10] Convention for the Protection of Human Rights and Fundamental Freedoms as amended by Protocols No. 11 and No. 14; https://www.coe.int/en/web/conventions/full-list/-/conventions/rms/0900001680063765.

or bias arising from technical constraints[11] (e.g. limitations of computers or difficulty to formalise the non-formal).

Credit scoring is one of the domains most studied, because the use of ADS in this context can have significant impact on individuals' lives. For example, the US National Consumer Law Center released a report in June 2007, referring to several studies on the disparate impact effect of the use of credit scoring. A striking case was the figures from the Missouri Department of Insurance[12] showing significantly worse insurance scores for residents of high-minority ZIP codes, even after eliminating other factors such as income, education, or unemployment. Lisa Rice and Deidre Swesnik also report many examples of discrimination against communities of colour resulting from the use of credit-scoring systems in the US.[13] One may argue that credit scores perpetuate a long history of discrimination in the loan sector. However, the disparate impact of credit scoring goes far beyond this sector because credit scores such as the FICO score are increasingly used in different types of context, such as employment, insurance or rental accommodation. In addition, as stated by Lisa Rice and Deidre Swesnik, credit-scoring mechanisms are not necessarily fair to borrowers in the sense that they take features that are not related to them as individuals but to their environment into account. This issue is bound to become more acute as the variety of factors that can be used to assess risk scores will increase with the growing amount of information available on the web or collected by internet trackers. For example, Facebook has filed a patent that could be used by banks to decide to deny a loan to an individual if the average credit ranking of their friends is below a given threshold.[14] In the same spirit, sentiment analysis based on social network or quantified-self information could be used by insurers to personalise pricing. For example, according to a Swiss re-insurer, Twitter data could be a more reliable predictor of heart disease than traditional health and socioeconomic measures.[15]

Credit scoring is one of the most studied domains because the use of ADS in this context can have a strong impact on individuals' lives. A report by the National Consumer Law Center shows significantly worse insurance scores for residents of high-minority ZIP codes in Missouri, even after eliminating other factors, such as income, education or unemployment.

Discriminatory practices in online services are attracting increasing attention from the computer science community. For example, using their **Sunlight** system, Mathias Lecuyer and his colleagues have shown with statistical confidence that Google services used protected attributes such as race, religious affiliation or health to generate targeted advertisements.[16] In the same spirit, Amit Datta and his co-authors have developed a tool, called **AdFisher**, which has been used to provide evidence of discrimination based on gender in employment ads: simulated males receive ads for positions with large salaries more frequently than simulated females with the same profile.[17]

---

11 Batya Friedman, Helen Nissenbaum; Bias in computer systems; ACM Transactions on Information Systems; (14, 3); 1996.

12 Birny Birnbaum; Credit scoring and insurance: costing consumers billions and perpetuating the racial divide; National Consumer Law Center; 2007.

13 Lisa Rice, Deidre Swesnik; Discriminatory effects of credit scoring on communities of color; Suffolk University Law Review; (46), 2013.

14 Robinson Meyer; Could a bank deny your loan based on your Facebook friends?; The Atlantic; 2015.

15 Brenna Hughes Neghaiwi; In insurance big data could lower rates for optimistic tweeters; Reuters; 2016.

16 As stated by the authors however, the `system cannot assign intention of either advertisers or Google for the targeting': Mathias Lecuyer, Riley Spahn, Yannis Spiliopolous, Augustin Chaintreau, Roxana Geambasu, Daniel Hsu ; Sunlight: fine-grained targeting detection at scale with statistical confidence; 22nd ACM SIGSAC Conference on Computer and Communications Security (CCS); ACM; 2015.

17 Amit Datta, Michael Carl Tschantz, Anupam Datta; Automated experiments on ad privacy settings; Privacy Enhancing Technologies (PET); 2015.

The use of certain ADS can also lead to discrimination against underprivileged or minority neighbourhoods. For example, some geo-navigational applications are designed to avoid 'unsafe neighbourhoods', which could lead to a form of redlining and 'reinforce existing harmful and negative stereotypes about poor communities and communities of colour'.[18] The same criticism has been raised about predictive policing systems increasingly used by police forces in the USA. The goal of these systems is to predict places where crimes are most likely to happen in the future based on input data such as the location and timing of previously reported crimes. Leaving the quality of their predictions aside, these systems may just produce self-fulfilling prophecies, as more controls lead to more reported crimes, and reinforce disproportionate and discriminatory policing practices.

Discrimination in justice: Another area that has raised much concern is the increasing reliance on ADS in the criminal justice system. A widely-publicised case is the COMPAS[19] system used to assess individual risk levels of recidivism, violence or failure to appear. COMPAS scores can be used at different stages of the criminal justice system, e.g. to decide whether to release or detain a defendant before their trial or whether to grant parole to an offender. A study conducted by ProPublica[20] led to the conclusion that:

> 'black defendants were far more likely than white defendants to be incorrectly judged to be at a higher risk of recidivism, while white defendants were more likely than black defendants to be incorrectly flagged as low risk.'

However, ProPublica's analysis has received a number of criticisms both from the academic community and from Northpointe Inc., the company that developed COMPAS.[21] In particular, Alexandra Chouldechova[22] shows that the difference in false positive and false negative rates identified by ProPublica results from a difference in the proportion of individuals who reoffend across the groups (prevalence). Chouldochava even shows that different fairness criteria cannot be satisfied simultaneously. Sam Cobett-Davies and his co-authors argue along the same lines, showing that the error rate balance (used by ProPublica) is not compatible with predictive parity (used by Northpointe). Predictive parity is defined by the fact that, above a given threshold, the likelihood of recidivism among high-risk offenders is the same regardless of group membership. This measure ensures that scores mean essentially the same thing regardless of race, which is a reasonable expectation from a non-discriminatory ADS.

One conclusion that can be drawn from the COMPAS debate is that several definitions of discrimination are possible, which at first sight, may appear equally legitimate. The same comment can be made for other domains of application of ADS.[23] For example, in the credit and mortgage markets, different approaches to discrimination rely on rejection (disproportionate rate of rejection between groups), pricing (different costs for different groups) or default (probabilities of defaults in the different groups). A distinction can also be made between group based measures of discrimination (e.g. similar acceptance rates or similar levels of revenues for different groups)[24] or at

---

18 Joe Silver; Is your turn-by-turn application racist?; ACLU; 2013.

19 Correctional Offender Management Profiling for Alternative Sanctions.

20 Jeff Larson, Surya Mattu, Lauren Kirchner, Julia Angwin ; How we analyzed the COMPAS recidivism algorithm; ProPublica; 2016; https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm.

21 Now rebranded to equivant : http://www.equivant.com

22 Alexandra Chouldocheva; Fair prediction with disparate impact: a study of bias in recidivism prediction instruments; Big Data; Special issue on social and technical trade-offs; 2017.

23 Romei, A., Ruggieri, S.; A multidisciplinary survey on discrimination analysis; The Knowledge Engineering Review; (29,5); doi:10.1017/S0269888913000039; 2014.

24 For example, the US Equal Employment Opportunity Commission refers to an 80 percent rule as a measure of disparate impact: the success rate of the protected group should not be less than 80 percent of the success rate of the non-protected group.

the individual level (people with similar profiles should be treated equally, regardless of which group they belong to). The best that can be done from the technical side is to state these definitions clearly but the choice between them is not technical. It is a matter of political options or ethics.

Beyond legal discrimination: Beyond discriminatory practices, which are specific forms of treatment considered unfair for 'protected groups' and prohibited by regulation, ADS can also threaten or strengthen equality in different ways. For example, in certain situations, personalised pricing and price steering can be considered unfair. In other cases, they can be sources of new opportunities for the less favoured. Depending on the criteria used to personalise the process, they may be considered acceptable or not. For example, a study conducted by Aniko Hannak and her co-authors shows that travel web sites and retail sites personalise search results based on the operating system used by the customers.[25] Researchers have also shown that some online shops charge customers different prices depending on their location.[26] As another example, Uber users may experience big price differences due to only small changes in their location.[27] In contrast with discrimination, personalised pricing is not in itself illegal, even though most customers find it unfair, especially when the process is opaque. As an illustration, the revelation that Amazon was charging different prices for different customers based on demographic data created discontent in 2000 and led its CEO to officially deny such practices. However, some economists also point out that price personalisation can be beneficial not only for the economy but also for the less-favoured because sellers can offer some goods or services at a lower price than would be possible under a uniform pricing regime.[28]

One conclusion that can be drawn from the COMPAS debate is that several definitions of discrimination are possible, which, at first sight, may appear equally legitimate. The best that can be done from the technical side is to state these definitions clearly, but the choice between them is not technical. It is a matter of political options or ethics.

ADS to reduce or detect discrimination: When ADS are used to support human decision making, the risk of discrimination should also be compared with the risk of discrimination **without the use of ADS**. Human beings have many sources of bias that can affect their decisions. For example, a study conducted on more than one thousand judicial rulings by judges presiding over parole boards in Israel has shown that the ratio of favourable rulings drops from about 65 % to nearly zero during a session and goes back up to about 65 % after a break.[29] Another area in which minorities often suffer from discrimination is police activities, in particular 'stop-and-search' (investigatory pedestrian stop). Sharad Goel and her co-authors argue that the use of ADS by police forces could both reduce disparate racial impact and increase the efficiency of stop-and-search practices. In addition, it would make the police force more accountable.[30] More generally, one could then argue

[25] Aniko Hannak, Gary Soeller, David Lazer, Alan Mislove, Christo Wilson; Measuring price discrimination and steering on e-commerce web sites; ACM Internet Measurement Conference (IMC); 2014.

[26] Frederik J. Zuiderveen Borgesius; Online price discrimination and data protection law; Amsterdam Law School Legal Studies Research Paper; (2015-32); 2015.

[27] Le Chen, Alan Mislove, Christo Wilson; Peeking beneath the hood of Uber; ACM Internet Measurement Conference (IMC); 2015.

[28] Matthew A. Edwards; Price and prejudice: the case against consumer equality in the information age; Lewis & Clark Law Review; (10, 3); 2010.

[29] Shai Danzinger, Jonathan Levav, Liora Avnaim-Pesso; Extraneous factors in judicial decisions; PNAS; (108, 17); 2011. See also a criticism of this paper and the authors reply in PNAS; (108, 42).

[30] Sharad Goel, Maya Perelman, Ravi Shroff, David Alan Sklansky; Combatting police discrimination in the age of big data; New Criminal Law Review; (20, 2); 2017.

that a potential benefit of the use of ADS is to avoid certain types of human bias.[31] In addition, ADS may enhance traceability and therefore make it easier to detect bias.

The fact that a decision procedure is automated (or partly automated) may also encourage public discussion about the criteria used by the system, the underlying logic and the expectations of society in this respect, in particular in terms of fairness or non-discrimination. Several occurrences of this process have already been observed, not only in the field of justice with COMPAS, but also in education with the public debate raised by an algorithm called APB[32] in France. APB was used to decide upon the assignment of students to universities. Following this debate, APB has been discontinued and replaced by a system leaving more room for human intervention. A necessary condition for constructive public debate is the availability of a minimum amount of information about the algorithms. We discuss this issue in detail in Chapter 4.

To conclude this short review of discrimination and infringements on the principle of equality, it should be clear that it is a topic of great concern for ADS, but that there is no technological determinism in this area. On the one hand, ADS can be used to create new inequalities, or to amplify and hide discriminations; on the other hand, they can also make discriminatory or unfair practices more traceable and reduce them. In Chapter 5, we discuss the technical instruments that can be used to realise the second option.

ADS can be used to create new inequalities, or to amplify and hide discrimination; on the other hand, they can also make discriminatory or unfair practices more traceable and combat them.

3.1.2. Benefits and risks related to the principles of privacy, dignity, autonomy and free will

Privacy: In the previous sections, we reviewed the potential impact of ADS in terms of different persons or groups not being treated in the same way. We now turn our attention to their potential impact on individuals in absolute terms, in particular on their autonomy, free will and privacy. Are ADS bound to be a source of alienation and a threat to an individuals' autonomy, as many of their detractors claim, or could they also serve self-development and free will? Privacy and data protection are major issues in this respect since they are generally associated with individual autonomy.[33] For example, referring to the German Federal Constitutional Court's Census decision of 1983, Antoinette Rouvroy and Yves Poullet state that:

> 'the Court establishes a clear and direct link between the Data Protection regime and two basic values enshrined in the Constitution, interpreting legal data protection regimes as mere implementations of those fundamental constitutional rights. The first of those fundamental constitutional rights is the right to respect and protection of one's 'dignity' guaranteed by Article 1 of the Constitution and the second one is the right to 'self-development', enacted by Article 2 of the Constitution. The fact that the Court will refer directly to these principles without mentioning the already existing Data Protection Law is noticeable. In its view, the major data protection principles derive directly from these two

[31] Unless they are trained or programmed to use time as a key decision factor, algorithms should not be affected by breaks.

[32] APB stands for 'Affectation Post Bac'. It was replaced in 2018 by a new system called 'Parcoursup'.

[33] Even though, it is not the only aspect of privacy and its social dimension it should not be neglected: Julie Cohen; Privacy, autonomy and information; Configuring the Networked Self; 2012; http://www.juliecohen.com/attachments/File/CohenCNSCh5.pdf.

> Constitutional provisions that consecrate the value of autonomy (self-determination) and the incommensurability (dignity) of each person in the society.'[34]

The deployment of ADS may pose a threat to privacy and data protection in many different ways. The first is related to the massive collection of personal data required to train algorithms. Personal data can be the target of a variety of attacks initiated by different parties (data controllers themselves, their employees, cybercriminals, states, etc.) with varying impact on individuals (financial, psychological, physical, etc.). Several frameworks have been proposed for the systematic analysis of privacy risks[35] and to perform the data protection impact assessments required by the European General Data Protection Regulation.[36]

One of the areas where tremendous progress has been made in AI in the last decade and which can have strong impact on privacy is image recognition. These techniques can be used in many types of ADS, in particular to identify people through facial recognition. This can be applied to images published on the web but also potentially to pictures taken in public places. It can help police forces identify potential criminals, but its generalisation would represent a serious threat to privacy. As an illustration, facial recognition is already in use in Shenzen, China, to identify, fine and notify jaywalkers via instant messaging.[37] More generally, the integration of facial recognition within augmented reality glasses could lead to the end of anonymity. Furthermore, face recognition is just one amongst many other ways to identify people based on physical features. For example, considering that each human being has a unique way of walking, gait analysis can also be applied to the identification or authentication of people. Again, this technology can also be used for legitimate reasons and provide valuable support in certain areas such as health. For example, gait cycles can provide useful information about neurodegenerative diseases such as Parkinson's and Alzheimer's disease.[38]

Facial recognition can be applied to monitor people and create significant privacy threats. However, this technology can also be used for legitimate reasons and provide valuable support in certain areas such as health. For example, gait cycles can give information about neurodegenerative diseases such as Parkinson's and Alzheimer's disease.

Chilling effect and conformism: Even when no attack has been carried out, the mere knowledge or suspicion that personal data about people is being collected can have a detrimental impact on them. Several studies have provided evidence of the chilling effect resulting from fear of online surveillance. For example, for certain Wikipedia articles, a 19.5 % fall in view counts was observed after Edward Snowden's revelations in June 2013.[39] In addition, as stated by Jonathon Perrey, 'the graph still suggests more than an ephemeral chilling effect that dissipates quickly. Rather the data

---

[34] Antoinette Rouvroy, Yves Poullet; The right to informational self-determination and the value of self-development: reassessing the importance of privacy for democracy; Reinventing data protection; Serge Gutwirth et. al. (eds); Springer; 2009.

[35] Sourya Joyee De, Daniel Le Métayer; Privacy Risk Analysis; Synthesis Series; Morgan & Claypool Publishers; 2016. Sourya Joyee De, Daniel Le Métayer; PRIAM: A Privacy Risk Analysis Methodology; 11th International Workshop on Data Privacy Management (DPM); IEEE; 2016. Mina Deng, Kim Wuyts, Riccardo Scandariato, Bart Preneel, Wouter Joosen; A privacy threat analysis framework: supporting the elicitation and fulfilment of privacy requirements; Requirements Engineering; (16,1); 2011. David Wright, Paul De Hert; Privacy Impact Assessment; Springer ; 2012.

[36] Commission Nationale de l'Informatique et des Libertés (CNIL) ; Privacy Impact Assessment (PIA) Tools (templates and knowledge bases); 2018.

[37] http://www.scmp.com/tech/china-tech/article/2138960/jaywalkers-under-surveillance-shenzhen-soon-be-punished-text

[38] Omer Faruk Ince, Ibrahim Furkan Ince, Jang Sik Park; Gait analysis and identification based on joint information using RGB-depth camera; 14th International Conference on Electrical Engineering/Electronics, Computer, Telecommunications and Information Technology (ECTI-CON); 2017.

[39] Jonathon W. Penney; Chilling effects: online surveillance and Wikipedia use; Berkeley Technology Law Journal; (31,1); 2016; https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2769645.

suggests a lasting impact on total article views'. In the same spirit, an evolution towards a 'scored society'[40] would inevitably generate more conformity, everybody trying to comply with the explicit or implicit norm to obtain the benefits associated with good scores. For example, knowing (or suspecting) that banks analyse individual's social network links before deciding to grant or deny a loan, might tempt people to adapt their behaviour accordingly. In particular, they might decide to stop interacting with friends suspected as having a low score that would negatively impact their own score. Altogether, the impact of large-scale surveillance and scoring made possible by ADS would be to reduce the range of possibilities for individuals and therefore affect their capacity for self-development.

Reducing human beings to numbers: The discussion about scoring relates to a more general fear that humans are increasingly treated as numbers and reduced to their digital profiles. Many signs of the advent of the 'scored society' can already be observed, with the use of scores in insurance, banking, employment and many other areas. One of the most extreme illustrations of this trend is the 'social credit system' which is currently being experimented with in China and which will become mandatory in 2020.[41] Each Chinese citizen will be rated based on a wide variety of information, such as their credit history, shopping habits or interpersonal relationships. This score will affect their life in many ways, not only their ability to get a loan, but also to rent a car without leaving a deposit, to be entitled to faster check-in at hotels, or the fast-track application to get a Schengen visa, etc. In addition to the aforementioned risks of constant monitoring, this reduction of human personality to a single number could be seen as a form of alienation and an offence to human dignity. As stated by Luciano Floridi, 'Our dignity rests in being able to be the masters of our own journeys, and keep our identities and our choices open. Any technology or policy that tends to fix and mould such openness risks dehumanising us, not unlike Circe's guests, who are prevented from leaving her island.'[42]

As stated by Luciano Floridi, 'Our dignity rests in being able to be the masters of our own journeys, and keep our identities and our choices open. Any technology or policy that tends to fix and mould such openness risks dehumanising us, not unlike Circe's guests, who are prevented from leaving her island.'

Filter bubble effect: Another surreptitious effect of ADS is described as the **filter bubble** by Eli Pariser. According to Pariser, the personalisation of web searches hinders creativity and the ability to think, because it limits the diversity of content to which people are exposed. An often quoted example is the result of the Brexit vote in the UK, which was unthinkable for many anti-brexit voters based on the information they had seen during the campaign. The same comment has been made about the 2016 US presidential election. As Pariser states, 'If you only see posts from folks who are like you, you're going to be surprised when someone very unlike you wins the presidency'.[43] The filter bubble can obviously affect democratic life, which is discussed in Section 3.2, but it can also hamper individuals' self-development by reducing the type of information and the variety of opinions they are exposed to, leading to ideological confinement. An aggravating factor is the fact that many individuals do not realise that the content they see is selected or ranked by algorithms. For example, more than half of the participants of a study conducted about Facebook users in 2015

[40] Danielle Keats Citron, Frank Pasquale; The scored society: due process for automated predictions; Washington Law Review; (89); 2014; Available at SSRN: https://ssrn.com/abstract=2376209.

[41] Rachel Botsman; Big data meets Big Brother as China moves to rate its citizens; Wired; 2017; http://www.wired.co.uk/article/chinese-government-social-credit-score-privacy-invasion.

[42] Luciano Floridi; On human dignity as a foundation for the right to privacy; Philosophy & Technology; (29,4); 2016.

[43] Jasper Jackson; Eli Pariser: activist whose filter bubble warnings presaged Trump and Brexit; The Guardian; 2017; https://www.theguardian.com/media/2017/jan/08/eli-pariser-activist-whose-filter-bubble-warnings-presaged-trump-and-brexit.

were not aware of the News Feed curation algorithm.[44] Furthermore, being aware that an algorithm is used to filter and rank information does not mean knowledge of the underlying logic of this algorithm or the reasons for whcih a specific piece of content is presented to a given person. This lack of transparency opens the door to all kinds of manipulation. It can undermine individuals' autonomy, to either serve economic interests (for example in the case of micro-targeting ads), or for political purposes (when interest groups or states try to influence voters).

If there is little doubt about the potential impact of content personalisation, the scale of the filter bubble effect in practice is still a matter of debate. For example, Frederik Borgesius[45] and his colleagues concluded a study conducted in 2015 with the observation that 'in spite of the serious concerns voiced, at present, there is no empirical evidence that warrants any strong worries about filter bubbles.' Indeed, homophily[46] is a natural trend and how much it is amplified by ADS is not easy to assess. In addition, filtering could also be used in a different, more transparent way, to provide individuals with more control over their content. As an illustration of this approach, the social media aggregator Gobo[47] allows its users to set the parameters of the algorithm according to what they want to see. For example, a user can define the proportion of political news that matches or challenges their own political perspective, or the proportion of serious or 'fun' news. Gobo is still a preliminary prototype, but it shows that ADS can also be used to broaden the diversity of information to which individuals are exposed. They can therefore also contribute to empowering people and help them reinforce rather than undermine their autonomy.

In some cases, the reason for filtering can be compliance with moral norms. However, the precise nature of what should be considered as morally acceptable or not may vary among cultures and is difficult to implement automatically. A typical illustration of this issue is the difficulty for Facebook to implement its ban of nude photographs. Facebook had to reverse its decision and alter the results of its filtering algorithm in the face of much protest after censoring an image of the Venus of Willendorf, one of the oldest pictures of nude females in the history of art, or the Pulitzer prize-winning photograph of a naked girl fleeing napalm bombs during the Vietnam war. Again, this type of filtering can be perfectly legitimate and even welcome in certain situations, for example to protect children from certain types of content, but when it concerns adults, it can be seen as a form of paternalism restricting individuals' autonomy.

Challenging ADS decisions: Another major issue with opaque ADS is that they make it difficult to challenge a decision based on their results. This is in contradiction, for example, with defence rights and the principle of adversarial proceedings in most legal systems. In the majority of jurisdictions, the judge has a duty to state the reasons on which their decision was based and the parties in the lawsuit have the right to challenge this decision. Beyond breaching the principle of adversarial proceedings, the use of ADS in courts of law raises far-reaching questions about the reliance on predictive scores to make legal decisions, in particular for sentencing. As stated by Angèle Christin and her co-

The use of ADS in courts raises far-reaching questions about the reliance on predictive scores to make legal decisions. Another major issue with opaque ADS is that they make it difficult to challenge a decision based on their results. This is in contradiction with rights of defence and principles of adversarial proceedings.

[44] Motahhare Eslami, Aimee Rickman, Kristen Vaccaro, Amirhossein Aleyasen, Andy Vuong, Karrie Karahalios, Kevin Hamilton, Christian Sandvig, "I always assumed that I wasn't really that close to (her)": reasoning about invisible algorithms in the news feed; Conference on Human Factors in Computing Systems (CHI); ACM; 2015.

[45] Frederik J. Zuiderveen Borgesius, Damian Trilling, Judith Möller, Balazs Bodo, Claes H. de Vreese, Natali Helberger; Should we worry about filter bubbles?; Internet Policy Review; (5,1); 2016.

[46] Homophily is defined as the trend for people to associate with others similar to them, to be more open to information or opinion confirming their preconceptions and to be more influenced by people similar to them.

[47] https://gobo.social

authors, 'perhaps even more problematic is the theory of justice implicitly embedded in the algorithms'.[48] The point is that most ADS used in this context are risk-assessment tools: based on a number of factors about the defendants' criminal history, sociological data or demographic features, they provide an estimation of their risk of recidivism. As a result, they privilege one objective (incapacitation, defined as prevention from reoffending) to the detriment of other traditional justifications of punishment in law, such as retribution (taking into account the severity of the crime), rehabilitation (social reintegration) and deterrence. Two main approaches to sentencing are often distinguished: the deontological (or retributive) approach and the utilitarian (or consequentialist) approach. In the deontological approach, offenders should be punished 'because they deserve it and the severity of their punishment should be proportional to their degree of blameworthiness'. Therefore, 'assessing the risk of future crime plays no role in sentencing decisions'.[49] In contrast, risk assessment is a key instrument to implement the utilitarian approach, in which punishment is justified by the ability to decrease the probability of future crimes.

Adverse side effects of ADS: ADS and data analysis tools in general, can provide individuals with many new services, for example to improve their self-knowledge and possibly adopt new, more healthy habits thanks to quantified-self devices, discover potential ways to save money by analysing their shopping history, or learn new skills through personalised educational tools. However, when ADS are used to perform activities that were previously accomplished by human beings, they may also have adverse effects. The first is a threat to employment. Many studies have been conducted about the types of jobs that are threatened by the development of AI, with contrasting conclusions. It seems unavoidable that certain types of jobs will disappear and many others will be transformed, but it is not clear to what extent they will be replaced by new jobs (such as data scientists or ADS experts). As this issue merits a study of its own, we do not discuss this further in this study.

Another adverse effect of the replacement of human activities by automatic systems is the loss of skills or expertise no longer exercised. A typical example is the increasing number of taxi drivers who are unable to find their way in a city without a navigation system. With the development of voice-to-text applications, it is not unlikely that many people will not be able to write in the future, at least not properly, whether by hand or using a keyboard. If the current trend is sustained, some experts also fear that, in the long term, algorithms could outperform human beings in all areas and possibly even take control and dominate the world. The possibility of this extreme scenario actually happening is very controversial. It relies on the transition from weak artificial intelligence, with algorithms dedicated to specific tasks, to strong artificial intelligence able to perform all essential human tasks.[50] However, whether this is unlikely or unrealistic, it could at least be used as a dystopian scenario and an indication of potential extreme risks to be assessed in future artificial intelligence services and tools.

3.1.3. Opportunities and risks related to healthcare, quality of life, wellbeing and physical integrity

Benefits for health: ADS can have an impact on healthcare, quality of life, wellbeing and even the physical integrity of individuals affected by their decisions. ADS are already in use in the medical sector and can potentially contribute to improve the decisions taken by practitioners and specialists in many ways:

---

[48] Angèle Christin, Alex Rosenblat, Danah Boyd; Courts and predictive algorithms; Data & Civil Rights: A new era of policing and justice; 2015. http://www.law.nyu.edu/sites/default/files/upload_documents/Angele%20Christin.pdf

[49] John Monahan; Risk assessment in criminal sentencing; University of Virginia School of Law; Public Law and Legal Theory Research Paper Series; (2015,03); 2015.

[50] https://en.wikipedia.org/wiki/Technological_singularity

- Medical imaging: image analysis systems can detect pathologies that are difficult to identify even by experts. Examples include quantitative retinal image analysis and early identification of melanomas.[51] Skin cancer can already be detected from images with a level of accuracy that is comparable to a dermatologist.

Image analysis systems can detect pathologies that are difficult to identify even by experts. Examples include quantitative retinal image analysis and early identification of melanomas.

- Diagnosis: IBM Watson is used by some oncology departments to suggest treatments or options to doctors on cancer cases.[52] More generally, treatment recommendations are especially useful for rare diseases that practitioners may have never previously encountered.[53]
- Surgery: robots are increasingly used to help surgeons perform meticulous movements in tight spaces with greater dexterity.
- Personalised medicine: it will be easier in the future to tailor treatments based on the medical history, genetic lineage, diet and other specific conditions of the patient.

Women, the elderly and minorities are less well represented in control trials. The consequences may be that they are less likely to receive the right treatment because their symptoms do not match those of typical 'white adult man'.

Similarly, quantified-self or medical applications are being developed to help people improve their health by monitoring their physical activities or eating habits. As stated in a recent MITRE report:[54]

> 'there are many impressive smartphone attachments and apps currently available for monitoring of personal health. These devices 1) empower individuals to monitor and understand their own health, 2) create large corpora that can, in theory, be used for AI applications, and 3) capture health data that can be shared with clinicians and researchers.'

Opacity issues: The stakes are very high in the health sector and opacity is often unacceptable. For example, Rich Caruana and his co-authors report a case where an ADS based on neural networks was not used in a health-care project because of its lack of intelligibility.[55] The goal of the system was to predict the level of risk (probability of death) of patients with pneumonia, to decide whether they should be admitted to hospital or treated at home. The neural network based ADS predicted that patients suffering from asthma had a lower risk of dying from pneumonia. This prediction went against the knowledge and experience of doctors. It turns out that it reflected a bias in the training dataset: patients with asthma received more intensive care, which effectively lowered their risk of dying from pneumonia. Needless to say, if this ADS had been used to make decisions regarding admissions to hospitals, it would have put the lives of patients suffering from asthma at risk. Other ADS were not better in this respect but, because they were more intelligible, they could be corrected to avoid this type of bias and produce results in line with the experience of professionals.

Different types of biases: Another risk of the use of ADS in the medical sector is an increase in inequality due to the fact that learning data are often biased. Women, the elderly and minorities are less well represented in control trials, and the consequences may be that they are less likely to receive the right treatment because their symptoms do not match those of a typical 'white adult man'. More generally, as pointed out by a number of experts, the use of artificial intelligence also

---

51 Artificial Intelligence for health and health care; JASON; The MITRE Corporation, JSR-17-Task-002, 2017.

52 http://www.wired.co.uk/article/ibm-watson-medical-doctor

53 https://www.wired.com/2014/06/ai-healthcare/

54 https://www.wired.com/2014/06/ai-healthcare/

55 Rich Caruana, Yin Lou, Johannes Gehrke, Paul Koch, Marc Sturm, Noémie Elhadad; Intelligible models for healthcare: predicting pneumonia risk and hospital 30-day readmission; Knowledge Discovery and Data Mining Conference (KDD); ACM; 2015.

raises ethical questions. As stated by David Magnus, director of the Stanford Center for Biomedical Ethics:

> 'bias can play into health data in three ways: human bias; bias that is introduced by design; and bias in the ways health care systems use the data. […] You can easily imagine that the algorithms being built into the health care system might be reflective of different, conflicting interests. What if the algorithm is designed around the goal of saving money? What if different treatment decisions about patients are made depending on insurance status or their ability to pay?'[56]

On the other hand, those promoting the use of artificial intelligence in healthcare argue that it can be a way to overcome the cognitive bias of physicians.[57] These biases, which are inherent in any human decision, typically include availability (experience with past cases) and anchoring heuristics[58] (relying on an initial diagnostic impression, despite subsequent information to the contrary).

Automatic control systems: Other uses of ADS that can have an impact on the physical integrity of individuals include automatic control systems for cars, planes or underground rail systems. Again, ADS can have a positive or a negative impact in these situations. The positive impact is the potential to lower the rate of accidents. In many situations an automatic system can indeed take reasonable decisions based on multiple parameters much more efficiently than humans. In the case of autonomous vehicles, the most difficult issues are related to the fact that they are supposed to evolve in human environments and therefore capable of reacting in many unpredictable situations. Ethical discussions are often based on extreme scenarios, such as the dilemma between running over a group of pedestrians or sacrificing the driver and the lives of the passengers to save the pedestrians.[59] As stated by Johannes Himmelreich, more mundane situations can also raise complex and subtle issues:

> 'For example, the design of self-driving cars needs to balance the safety of others – pedestrians or cyclists – with the interests of cars' passengers. As soon as a car goes faster than walking pace, it is unable to prevent from crashing into a child that might run onto the road in the last second. But walking pace is, of course, way too slow. Everyone needs to get to places. So how should engineers strike the balance between safety and mobility? And what speed is safe enough?'[60]

Another related issue raised by autonomous vehicles is liability. If the level of driver control over the vehicle is very limited or null (or if they can only take back control in an emergency), then they should not be liable in the case of an accident. But who should be liable and how the levels of control should be characterised are open questions.[61] Kenneth S. Abraham and Robert L. Rabin make the following statement:

> 'We are on the verge of another new era, requiring another new legal regime. This time, it is our system of transportation that will be revolutionized. Over time, manually-driven cars are going

---

56 https://medicalxpress.com/news/2018-03-artificial-intelligence-medicine-ethical.html

57 https://www.cio.com/article/3203950/artificial-intelligence/ai-as-a-way-to-overcome-cognitive-bias-in-physicians.html

58 https://www.cio.com/article/3203950/artificial-intelligence/ai-as-a-way-to-overcome-cognitive-bias-in-physicians.html

59 Jean-François Bonnefon, Azim Shariff, Iyad Rahwan; The social dilemma of autonomous vehicles; Science; (352,6293); 2016.

60 https://theconversation.com/the-everyday-ethical-challenges-of-self-driving-cars-92710

61 Even though the Society of Automotive Engineers (SAE) has defined a five-tiered levels of automation (see footnote below).

to be replaced by automated vehicles. The new era of automated vehicles will eventually require a legal regime that properly fits the radically new world of auto accidents.'[62]

A recent report by the Future of Privacy Forum includes a summary of potential harms for individuals related to the use of ADS covers many issues discussed in this section.[63] These harms, listed in figure 2, are grouped into four broad categories:

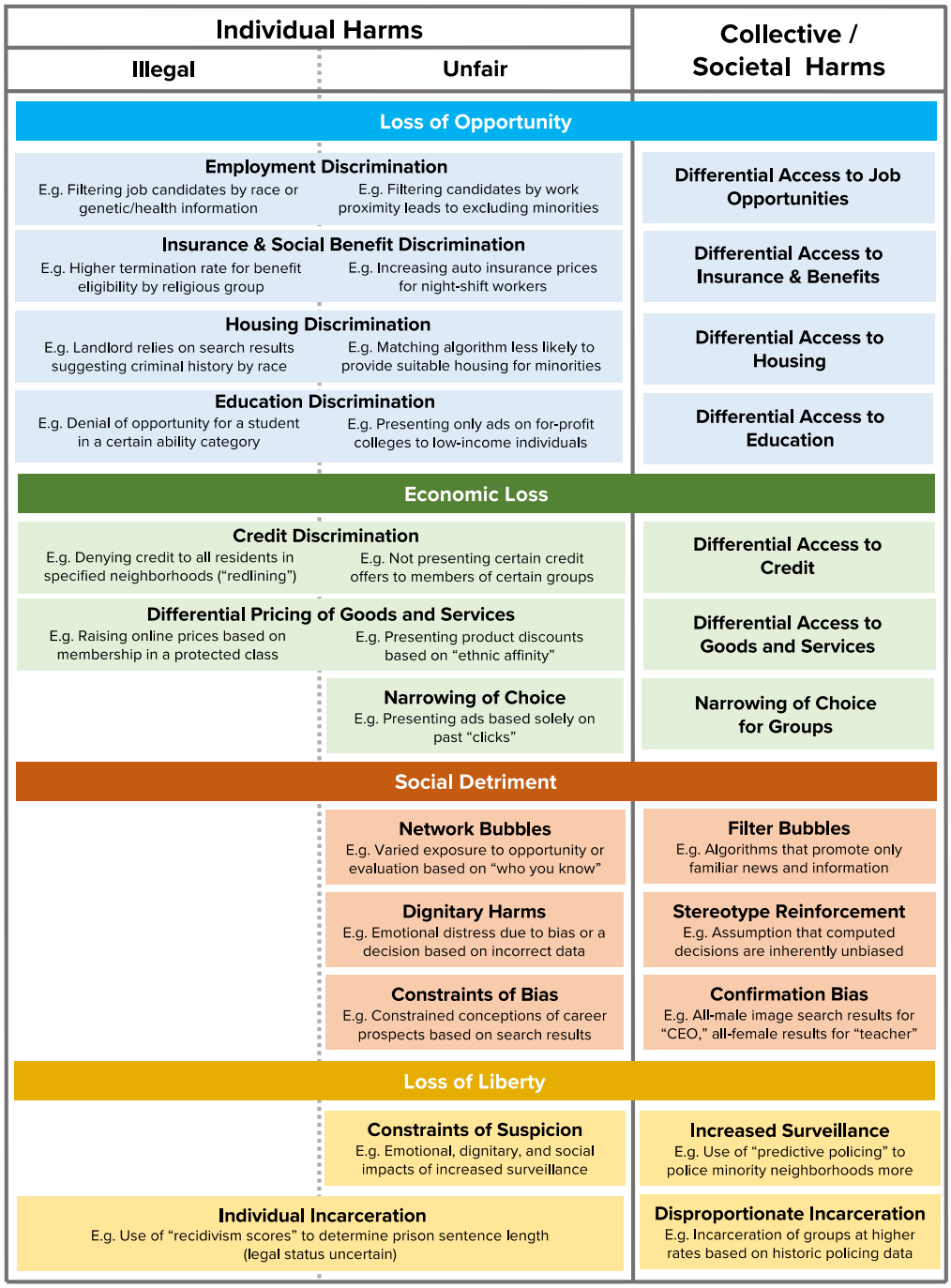
- loss of opportunity,
- economic loss,
- social detriment, and
- loss of liberty.

The Future of Privacy Forum report focuses only on harm and does not discuss opportunities. Nor does it consider certain aspects such as issues of opacity and the specific risks raised by automatic control systems. In addition, the distinction between illegal and unfair is based on US law and needs to be challenged. Nevertheless, this type of classification can be useful in the context of algorithmic impact assessments (AIA) as suggested in Chapter 8.

[62] Kenneth S. Abraham, Robert L. Rabin; Automated vehicles and manufacturer responsibility for accidents: a new legal regime for a new era; Virginia Law Review; forthcoming 2019.

[63] https://fpf.org/wp-content/uploads/2017/12/FPF-Automated-Decision-Making-Harms-and-Mitigation-Charts.pdf

Figure 2 – Potential harm caused by automated decision-making



Source: https://fpf.org/wp-content/uploads/2017/12/FPF-Automated-Decision-Making-Harms-and-Mitigation-Charts.pdf.

3.2. Opportunities and risks for the public sector

The use of ADS can also bring many new benefits and risks to the public sector.[64] In the following section, we distinguish what is at stake when using ADS in four different public sector activities:

[64] Peter Stone, Rodney Brooks, Erik Brynjolfsson, Ryan Calo, Oren Etzioni, Greg Hager, Julia Hirschberg, Shivaram Kalyanakrishnan, Ece Kamar, Sarit Kraus, Kevin Leyton-Brown, David Parkes, William Press, AnnaLee Saxenian, Julie

- public services,
- public safety and security,
- cyber-defence, and
- the safeguarding of democracy and national sovereignty.

3.2.1. Public services

ADS are currently being used by government and public agencies to provide new services or improve existing ones in many areas, such as energy, education, healthcare, transportation, justice systems and security. Examples of applications of ADS in this context are predictive policing, smart metering, video protection and university enrolment. They will also contribute to improving the quality of healthcare, education and job skill training. ADS, or smart technologies in general, such as mobility management tools, or water and energy management systems, can enhance city management. For example, they can help lower energy consumption, reduce traffic congestion or pollution and improve waste management. ADS can make government agencies themselves more efficient and help increase the quality of their services. They can also contribute to administration decisions making them more transparent and accountable, provided however that they are themselves transparent and accountable (see Chapter 4). ADS, and specifically ML techniques, can also contribute to society by producing new knowledge. For example, researchers have used ML algorithms to discover nearly 6 000 previously unknown species of virus.[65]

> ADS can contribute to making administration decisions more efficient, transparent and accountable, provided however that they are themselves transparent and accountable.

Healthcare analytics have the potential to revolutionise the medical sector by analysing the clinical records of millions of patients to enable personalised diagnosis and treatment. Similarly, as mentioned in Section 3.1, the emergence of mobile health, which exploits the data collected from a patient's smartphone, is very promising. However, these developments come with important risks for privacy. The data used to train these systems can be leaked and misused, for example by health insurance companies. Furthermore, another important hindrance for the adoption of healthcare analytics is the opacity of ADS results. This issue, which concerns many applications of ADS, is discussed in Chapter 4.

ADS create many new 'security vulnerabilities' that can be exploited by people with malicious intent, in particular by hackers or foreign organisations. Since ADS play a pivotal role in the workings of society, in nuclear plants, smart grids, hospitals, or cars for example, hackers who are able to compromise these systems have the capacity to cause major damage. Furthermore, these systems will be harder to protect, since these attacks are likely to become more automated and more complex and the risk of cascading failures will be harder to predict. A smart adversary may either attempt to discover and exploit existing weaknesses in the algorithms or create one that they will later exploit. This could be achieved by a poisoning attack, i.e. by interfering with the training data if machine learning is used. In addition, attackers might also use ML algorithms to automatically identify vulnerabilities, and optimise attacks by studying and learning in real time about the systems they target.

Shah, Milind Tambe, Astro Teller; Artificial Intelligence and Life in 2030; One Hundred Year Study on Artificial Intelligence: Report of the 2015-2016 Study Panel; Stanford University; 2016; http://ai100.stanford.edu/2016-report.

[65] Amy Maxmen; Machine learning spots treasure trove of elusive viruses; Nature News; 2018; https://www.nature.com/articles/d41586-018-03358-3.

Security failures may also occur accidentally, without malicious intent. With the deployment of ADS, large or small-scale accidents, such as autonomous cars running over pedestrians,[66] are very concrete threats that must be considered in order to prevent a loss of trust in transport systems and in automated systems in general.[67]

3.2.2. Public safety and security

ADS can help improve the security of infrastructures. They are already used to offer better protection against sophisticated security threats. They can help automate complex processes to detect attacks and undertake countermeasures. For example, data deception technology can be used to trick attackers, analyse their behaviours and take defensive actions against advanced attacks. They can also be used to detect white-collar crimes, such as money laundering or credit card fraud.

Many cities are already using ADS for public safety and security, e.g. by deploying surveillance cameras with face recognition, drones or predictive policing applications. These technologies may help police target their activities, prioritise tasks and solve crime cases. However, these systems come with significant risk to privacy. Surveillance technologies may become overwhelming and oppressive. Furthermore, they can potentially amplify bias and stigmatisation and have disparate impact on citizens.[68] In addition, most ADS operate as 'black boxes' and therefore lack transparency, making their efficiency debatable. For example, the promise of predictive policing is to tell law-enforcement officers the areas of highest risk for future crimes by using complex algorithms and past crime data. However, according to a study, these systems may merely reinforce bad policing habits in historically over-policed communities, thereby creating new sources of tension in these locations.[69]

Surveillance ADS may help police target their activities, prioritise tasks and solve crime cases. However, these systems come with significant risk to privacy. However, they may also amplify bias and stigmatisation and have disparate impact on citizens.

3.2.3. Cyber-defence

ADS are already in use in cyber-defence and they are bound to play an increasing role in this area. Existing machine learning technologies enable a high degree of automation in labour-intensive activities such as satellite imagery analysis. A more ambitious and controversial use of ADS in this context is to build autonomous weapon systems. A number of countries are increasing their studies and development of such systems as they perform increasingly elaborate functions, including identifying and killing targets with little or no human oversight or control.

Autonomous weapon systems can reduce casualties by replacing human soldiers in dangerous missions and protecting them, for example, from potentially harmful chemical substances. Furthermore, they may be more efficient for certain tasks as they are not subject to physical, physiological and mental constraints. They can also limit collateral damage thanks to their fine-grain targeting capabilities. Finally, autonomous weapons can help reduce

In July 2015, an open letter calling for a ban on autonomous weapons was released by artificial intelligence scientists and experts.

[66] Sam Levin and Julia Carrie Wong, Self-driving Uber kills Arizona woman in first fatal crash involving pedestrian, The Guardian, march 2018 (accessing July 10th, 2018).

[67] Amodei, Dario, Olah Chris, Steinhardt Jacob, Christiano Paul, Schulman John, Mane Dan, Concrete Problems in AI Safety, eprint arXiv:1606.06565, 2016, https://arxiv.org/pdf/1606.06565.pdf

[68] Even though one may argue that they could also mitigate some of the human decision-making biases.

[69] Lum, K. and Isaac, W. (2016), To predict and serve?. Significance, 13: 14-19. doi:10.1111/j.1740-9713.2016.00960.x

costs.[70] However, using ADS in this context obviously raises serious moral issues and voices have been raised against this type of application. For example, in July 2015, an open letter calling for a ban on autonomous weapons was released by artificial intelligence scientists and experts. The letter warns:

> 'Artificial Intelligence (AI) technology has reached a point where the deployment of such systems is – practically, if not legally – feasible within years, not decades, and the stakes are high: autonomous weapons have been described as the third revolution in warfare, after gunpowder and nuclear arms.'[71]

The development of autonomous weapons is hard to control and their proliferation is a risk. Furthermore, since they are more powerful and affordable than conventional weapons they may fall into the hands of dangerous organisations such as terrorist groups. Finally, since autonomous weapons embed many algorithms, they are prone to cyber-attacks. If they were actually deployed, the risk of malfunctioning, error or misuse should first be carefully addressed.

3.2.4. Democracy and sovereignty

It is of note that the public sector may also be impacted by ADS deployed by other entities, be they private stakeholders or foreign powers. An illustrative example is the deployment of the Waze system. Waze is a smartphone travel application to find the best path to destination, usually suggesting routes off the main roads, through side streets or residential areas.[72] This upsets residents and goes against the goal of city planners to keep cars on the main axes. The biggest problem here comes from the fact that private organisations like Waze do not have the same goals as city planners. This is a source of tension. A potential risk of ADS within this context, whereby public agencies accept that certain choices be made or influenced by private ADS, means they may lose control and sovereignty over public policy decisions.

> The main problem stems from the fact that a private stakeholder does not have the same goals as city planners, which is a source of tension. A potential risk of ADS in this context is the loss of control of public agencies over their decisions.

ADS can however make a positive contribution to democracy by, for example, allowing people to express their opinions on social networks and make them accessible to a wide audience of potentially interested people. However, the same technologies may be used by states to control people, for example by identifying political opponents and trying to intimidate them. More generally, states and interest groups could be tempted to use these technologies to influence citizen behaviours, which could lead to what has been called *'anticipative conformism*' by Antoinette Rouvroy.[73] These technologies can also be used to distort information, in order to damage the integrity of democratic discourse and the reputation of government or political leaders. It seems that this new form of AI-enhanced propaganda has been used in the alleged Russian operation to influence the 2016 USA presidential election. It leverages social media targeted advertising, psychological profiling and the propagation of fake news using bots of automated fake social network accounts. On the strategic side, a small group of nations and companies are investing massively in AI research and are likely to achieve dominance in related technologies, which could

---

[70] Amitai Etzioni, Oren Etzioni; Pros and Cons of Autonomous Weapons Systems; Military Review; 2017.

[71] Autonomous weapons: an open letter from AI and robotics researchers, future of life institute, July 2015, https://futureoflife.org/open-letter-autonomous-weapons/.

[72] Elizabeth Weise; Waze and other traffic dodging apps prompt cities to game the algorithms; USA Today; 2017; https://www.usatoday.com/story/tech/news/2017/03/06/mapping-software-routing-waze-google-traffic-calming-algorithmsi/98588980/.

[73] Antoinette Rouvroy; The end(s) of critique: data behaviourism vs. due-process privacy; Due Process and the Computational Turn; Routledge; 2012; http://works.bepress.com/antoinette_rouvroy/44.

lead to a strong imbalance of power. These global technological monopolies may threaten the sovereignty of many countries that will not be able to rely on their own technological means. This could lead to frustration from citizens and create more international tension.

3.3. Opportunities and risks for the private sector

The opportunities of ADS, and AI in particular, are endless for the private sector, but risks are also numerous. Any task that is repetitive, pressured by time, or that could benefit from the analysis of high volumes of data, is a prime target for ADS. These tasks concern low-skilled as well as highly-skilled personnel, for example in sectors like banking, insurance or justice.

Efficient and robust systems based on machine learning have replaced standard algorithms for setting inventory levels and optimising supply chains in many companies. In finance, ADS are used to decide which trades to execute (e.g. in high speed trading systems), and increasingly credit decisions are also made with the help of ADS. For example, JPMorgan Chase has deployed a system for reviewing commercial loan contracts, performing what it took loan officers 360 000 hours to do in only a few seconds. E-commerce sites employ ADS to optimise inventory and improve product recommendations to customers. Some companies, such as Mastercard, are using facial recognition tools to allow 'pay by face'. They also use elaborate ML-based analytics systems that predict whether a user is likely to click on a particular advertisement to improve online advertising placement and targeting. Supervised learning systems are now used by the pharmaceutical industry to develop better and more personalised drugs. Finally, an application of ADS that concerns all companies is its potential to improve their IT security by automatically detecting malware.

The above points represent just a small selection of examples of use of ADS in the private sector. In general, ADS are driving changes at three levels in industry: **tasks, business processes** and **business models**:

- An example of task redesign is the use of vision systems to detect the degradation or end of life of a mechanical component, leaving more time for technicians to focus on potential problems.
- An example of process redesign is the modification of the workflow and layout of packing warehouses following the introduction of robots and optimisation algorithms in a company.
- Finally, car sharing services are an example of a new business model that would not be possible without ADS.

Although ADS can be highly beneficial, they come with their own risks for the private sector. In particular, the results of ADS are often difficult to explain. This can reduce consumer trust and creates four main risks:

- There may be 'hidden' biases derived from the data provided to train the system. This can be difficult to detect and correct. In some cases, these biases can be characterised as discriminations and be sanctioned in court.
- It can be difficult, if not impossible, to prove that the system will always provide correct outputs, especially for scenarios that were not represented in the training data. This lack of verifiability can be a concern in mission-critical applications.
- In case of failure, it might be very difficult, given the models' complexity, to diagnose and correct the errors and to establish responsibilities.

- Finally, as previously mentioned, malicious adversaries can potentially attack the systems by poisoning the training data or identifying adversarial examples. These attacks can be difficult to detect and prevent.

While these risks are very real it must be recognised that human beings also have biases, make mistakes, are not always rational, and that their decision process cannot really be transparent. The advantage of ADS in this respect is that they can be audited systematically. Their disadvantage is that they can amplify biases and errors and make it more difficult to allocate liabilities.

Human beings also have biases, make mistakes, are not always rational, and their decision process cannot really be transparent. The advantage of ADS in this respect is that they can be audited systematically. Their disadvantage is that they can amplify biases and errors and make it more difficult to allocate liabilities.

We conclude this section with two systemic risks related to the use of ADS. First, ADS are likely to affect company organisation and management. The expression 'fourth industrial revolution' has been coined to describe this dramatic change. Certain types of jobs will change enormously or no longer exist, whilst new ones will appear. Some economists argue that automation will supplant jobs in manufacturing, but will also offer opportunities to replace them with more rewarding ones. All jobs might be affected, not only jobs that do not need an advanced level of education and expertise. For example, hairdressers are probably less likely to be affected than accountants or lawyers. A doctor using an ADS to scan medical data and to monitor patients will still need to interact with the patients and treat their diseases. The impact of this revolution is obviously not limited to the private sector: it is social and political, and education and training systems will need to be adapted.

Finally, ADS and AI can drive innovation by making it possible to analyse large data sets to develop new services. They create new business opportunities for large and small players. That said, most ADS require a lot of data and many of these datasets lie in the hands of a small group of dominant players. Furthermore, there is a trend for large high-tech companies to buy the most promising AI start-ups.[74] Productivity requires competition, and this is at risk with the current concentration in the market.

[74] Vinod Iyengar; Why AI consolidation will create the worst monopoly in US history; Techcrunch; 2016; https://techcrunch.com/2016/08/24/why-ai-consolidation-will-create-the-worst-monopoly-in-us-history/.

4. Desiderata for algorithms

In this chapter, we review the main approaches proposed in the literature to reduce the risks identified in Chapter 3. These approaches rely on notions such as transparency, explainability, interpretability and accountability, which are used with varied meanings in the literature. For the sake of clarity, we first define the concepts considered in this report precisely (Section 4.1), before comparing them with other definitions and terms used in the literature (Section 4.2).

4.1. Introduction to the main properties used in this document

Several key properties are generally required to enhance trust in algorithmic systems:

- **Safety**, defined as the absence of error in a system, especially errors that can cause damage. For algorithms, safety can be seen as the capacity to deliver correct results i.e. results consistent with their specifications, in the absence of adversarial attack.
- **Security**, defined as the protection of the system against adversarial attacks that could threaten its integrity or disrupt its services. The typical objectives of an adversary are to breach properties such as the **confidentiality, integrity** or **availability** (sometimes represented by the CIA acronym).
- **Privacy**, which relies on the protection of personal data.[75] In contrast to security, privacy can be threatened by the custodian of the personal data itself and the victim is the data subject rather than the organisation that holds the data.

The above requirements apply equally to ADS as to any algorithmic system. However, ADS are specific types of algorithms with strong impact on individuals, as discussed in Chapter 3. They should therefore also meet additional requirements, which can be classified into two main categories:

- **Intrinsic requirements**, such as fairness, absence of bias or non-discrimination, which can be expressed as properties of the algorithm itself (as a mathematical function from its inputs to its outputs) in its application context. As discussed in Section 5.1, different properties can be proposed to capture these requirements. The choice of a specific property is not technical but subjective, contextual and political. When a property has been defined, it can be checked a posteriori (verification) or established a priori (by design). In this report, we equate 'fairness' with 'absence of undesirable bias' and characterise 'discrimination' as a specific form of unfairness related to the use of specific types of data (such as ethnic origin, political opinions, gender, etc.). The use of discriminatory features is prohibited by law in certain types of context such as credit, employment, housing, etc. The specific list of prohibited attributes, contexts and precise means to assess discrimination depend on national laws and jurisdictions.
- **Extrinsic requirements**, such as understandability, are defined as the possibility to provide comprehensible information about the link between the inputs and the outputs of the ADS. This information can take many different forms depending on the recipients (designer of the ADS, user, person affected by the decisions, auditor, etc.), their level of expertise and their objectives. The two main forms of understandability considered in the literature are:

[75] From the legal point of view, privacy can be distinguished from personal data protection, but we do not make the distinction here and refer to privacy in a general sense.

- **Transparency**, defined as the availability of the ADS code with its design documentation, parameters and learning dataset when the ADS rely on machine learning. Transparency does not necessarily mean public availability. It also encompasses cases in which the code is disclosed to specific entities for audits or verifications.
- **Explainability**, defined as the availability of explanations about the ADS. In contrast with transparency, explainability requires the delivery of information beyond the ADS itself. Several explanation modes can be distinguished:
  - Explanations can be of three different types: operational (informing how the system actually works), logical (informing about the logical relationships between inputs and results) or causal (informing about the causes for the results).
  - Explanations can be either global (about the whole algorithm) or local (about specific results).[76]
  - Explanations can take different forms (decision trees, histograms, picture or text highlights, examples, counterexamples, etc.).

The strengths and weaknesses of each explanation mode should be assessed in relation to the recipients of the explanations (e.g. professional or individual), their level of expertise and their objectives (understanding the results to make a decision, challenging a decision, verifying compliance with legal obligations, etc.).

In a nutshell, these two forms of understandability correspond to two strategies: **show** (transparency) or **explain** (explainability).

Accountability is another key desideratum that is often put forward in the context of ADS. Reuben Binns[77] defines accountability as follows:

> 'a party A is accountable to a party B with respect to its conduct C, if A has an obligation to provide B with some justification for C, and may face some form of sanction if B finds A's justification to be inadequate'.

Binns also notes that:

> 'in the context of algorithmic decision-making, an accountable decision-maker must provide its decision-subjects with reasons and explanations for the design and operation of its automated decision-making system'.

Therefore, accountability can be seen as a requirement on a process (obligation to provide justification), which applies to the two categories of requirements for ADS discussed above (intrinsic and extrinsic requirements), each case corresponding to specific types of 'justifications' (proof of non-discrimination, source code, local or global explanation, etc.). Another essential facet of accountability is the possibility of sanctions, which is also orthogonal to the two (intrinsic and extrinsic) categories of requirements.

---

[76] Riccardo Guidotti and his co-authors use the expressions 'model explanation' and 'outcome explanation' to refer respectively to global explanations and local explanations: Riccardo Guidotti, Anna Monreale, Franco Turini, Dino Pedreschi, Fosca Giannotti; A survey of methods for explaining black box models; arXiv:1802.01933 [cs.CY]; https://arxiv.org/abs/1802.01933.

[77] Reuben Binns; Algorithmic accountability and public reason; Philosophy & Technology; 2017.

4.2. Other definitions and terms used in the literature

Many papers use terms such as transparency, explainability, interpretability, accountability or fairness with different meanings or without defining them properly (and often without introducing clear distinctions between them). To place our definitions within a more general context, we focus in the remainder of this section on some alternative definitions or interpretations of these terms used in the literature and compare them with our own. A summary of these variations is presented in figure 3 below.

Fairness is sometimes defined as the fact that the provider of the ADS does not misrepresent its functionalities or divert it against the interest of the users of the ADS or people affected by its results.[78] This version of fairness is more a subjective requirement on the behaviour and claims of the **provider** of the ADS than a requirement on the ADS itself. We therefore stick to the more restrictive definition of fairness introduced in Section 4.1.

Transparency is used with very different meanings in the literature, ranging from the specific obligation to disclose the code of the algorithm (with the learning dataset when the ADS relies on machine learning) to a generic interpretation encompassing any means to reduce the opacity of an ADS. As an illustration of the first trend, Mike Ananny and Kate Crawford[79] do not provide a single definition of transparency but describe several types of transparency (upwards versus downwards, outwards versus inwards, event versus process, etc.). They discuss several limitations of transparency as a tool for accountability, including the fact that 'seeing inside a system does not necessarily mean understanding its behaviour or origins'. They also argue that 'the ideal of transparency places a tremendous burden on individuals to seek out information about a system, to interpret that information, and determine its significance'. For the sake of clarity, we use the same specific (restrictive) meaning here, which makes it possible to highlight the differences between transparency, explainability and accountability (or, in other words, between respectively **showing, explaining** and **justifying**). In contrast, Zachary Lipton[80] seems to refer to transparency as the explanation of the operational aspects of the ADS (how the ADS actually works), as opposed to the explanation of its results, which he calls 'post-hoc interpretability'. The distinction he makes between (operational) transparency and post-hoc interpretability is the same as the difference between 'the processes by which we humans make decisions and those by which we explain them'. In the terms of this report, Lipton's 'post-hoc interpretability' corresponds to logical explanations, as opposed to operational explanations.

> We use a restrictive meaning of transparency here, which makes it possible to highlight the differences between transparency, explainability and accountability (i.e. between respectively, showing, explaining and justifying).

Explainability, interpretability: Lilian Edwards and Michael Veale introduce two categories of explanations,[81] model-centric and subject-centric explanations. According to their definitions, 'model-centric explanations provide broad information about a ML model which is not decision or input-data specific' whereas 'subject-centric explanations are built on and around the basis of an

---

78 Comment permettre à l'homme de garder la main ? Les enjeux éthiques des algorithmes et de l'intelligence artificielle, CNIL Report, 2017.

79 Mike Ananny, Kate Crawford; Seeing without knowing: Limitations of the transparency ideal and its application to algorithmic accountability; New Media and Society; (20,3); 2018.

80 Zachary C. Lipton; The mythos of model interpretability; ICML Workshop on human interpretability in machine learning (WHI); 2016.

81 Lilian Edwards, Michael Veale; Slave to the algorithm? Why a right to an explanation is probably not the remedy you are looking for; Duke Law & Technology Review; (18); 2017; Available at SSRN https://ssrn.com/abstract=2972855 or http://dx.doi.org/10.2139/ssrn.2972855.

input record'. This distinction corresponds to our notions of global explanations (about the whole algorithm) and local explanations (about specific results).

Riccardo Guidotti and his co-authors argue that the notions of interpretability, explainability and comprehensibility are strongly interrelated:

> 'To interpret means to give or provide the meaning or to explain and present in understandable terms some concept. Therefore, in data mining and machine learning, interpretability is defined as the ability to explain or to provide the meaning in understandable terms to a human. These definitions assume implicitly that the concepts expressed in the understandable terms composing an explanation are self-contained and do not need further explanations. Essentially, an explanation is an 'interface' between humans and a decision maker that is at the same time both an accurate proxy of the decision maker and comprehensible to humans.'[82]

Dhurandhar et al. also consider interpretability as a synonym of explainability:

> 'from our human perspective, interpretability typically means that the model can be explained, a quality which is imperative in almost all real applications where a human is responsible for consequences of the model.'[83]

Accountability: The characterisation of accountability in a document recently issued by the World Wide Web Foundation[84] is general and policy-oriented but consistent with our interpretation:

> 'Accountability is usually referred to as the duty governments and other authorities have to present themselves before those whose interest they represent or are otherwise bound to, and to justify how power was exercised, and resources were used'.

As noted by the World Wide Web Foundation and previously by Nicholas Diakopoulos,[85] transparency can be a mechanism that facilitates accountability.

The distinction between transparency and accountability is also stressed by Lilian Edwards and Michael Veale:[86]

> 'Despite the sometimes almost unthinking association of transparency and accountability, the two are not synonymous. Accountability is a contested concept, but in essence involves a party being held to account having to justify their actions, field questions from others, and face appropriate consequences. Transparency is only the beginning of the process.'

The emphasis on justification and sanctions is in line with the constitutive elements of accountability discussed by Mark Bovens.[87]

Accountability has also been introduced as a basic principle in data protection regulation since the publication of the OECD Guidelines on the Protection of Privacy and Transborder Flows of Personal

---

82 Riccardo Guidotti, Anna Monreale, Franco Turini, Dino Pedreschi, Fosca Giannotti; A survey of methods for explaining black box models; arXiv:1802.01933 [cs.CY]; https://arxiv.org/abs/1802.01933.

83 Amit Dhurandhar, Vijay Iyengar, Ronny Luss, Karthikeyan Shanmugam; A formal framework to characterize interpretability of procedures; ICML Workshop on Human Interpretability in Machine Learning; 2017.

84 World Wide Web Foundation, Algorithmic Accountability, July 2017.

85 Nicholas Diakopoulos; Accountability in algorithmic decision making; Communications of the ACM; 2016.

86 Lilian Edwards, Michael Veale; Slave to the algorithm? Why a right to an explanation is probably not the remedy you are looking for; Duke Law & Technology Review; (18); 2017; Available at SSRN https://ssrn.com/abstract=2972855 or http://dx.doi.org/10.2139/ssrn.2972855.

87 Mark Bovens; Analysing and assessing accountability: a conceptual framework; European Law Journal; (13,4); 2007.

Data[88] in 1980. However, the OECD Guidelines do not contain a precise definition of the term.[89] As noted by Charles Raab, the word accountability is often equated with responsibility or liability.

> 'In most European languages, due mainly to differences in the legal systems, the term 'accountability' cannot be easily translated. As a consequence, the risk of varying interpretations of the term, and thereby lack of harmonisation, is substantial. Other words that have been used to capture the meaning of accountability, are 'reinforced responsibility', 'assurance', 'relialability', 'trustworthiness' and in French, *'obligation de rendre des comptes'*, etc.'[90].

In its 2010 Opinion[91] on the principle of accountability, the Article 29 Data Protection Working Party observes that:

> 'The term 'accountability' comes from the Anglo-Saxon world where it is in common use and where there is a broadly shared understanding of its meaning – even though defining what exactly 'accountability' means in practice is complex. In general terms though its emphasis is on showing how responsibility is exercised and making this verifiable. Responsibility and accountability are two sides of the same coin and both essential elements of good governance. Only when responsibility is demonstrated as working effectively in practice can sufficient trust be developed.'

The key aspect of accountability shared by all definitions is therefore the obligation to justify, which applies to the different types of requirements discussed above.

88 Organisation for Economic Co-operation and Development. Guidelines on the Protection of Privacy and Transborder Flows of Personal Data. 1980.

89 The precise wording of the Guidelines is the following: `A data controller should be accountable for complying with measures which give effect to the principles stated above'.

90 Charles Raab; The Meaning of 'Accountability' in the Information Privacy Context; Managing Privacy Through Accountability, Daniel Guagnin et al. eds; Palgrave Macmillan; 2012.

91 Article 29 Data Protection Working Party; Opinion 3/2010 on the principle of accountability; 2010.

Figure 3 – Summary of alternative definitions found in the literature.

Definitions / Terms	Definition used in this report	Alternative definitions found in the literature
Fairness	Absence of undesirable bias	No misrepresentation of the functionality of the system
Transparency	Availability (public or controlled) of the ADS code with its design documentation, parameters and learning dataset.	Generic meaning (all means to reduce opacity, including the code availability and explainability) or specific meanings (focusing on the code availability or explainability of its operational aspects).
Explainability	Availability of explanations about the ADS. Explanations can be global or local, can be of different types (operational, logical, causal) and take different forms (decision trees, rules, counterexamples, etc.)	Global versus local explanations is sometimes called model-centric versus subject-centric. Interpretability or understandability are sometimes used as synonyms of explainability or closely related notions.
Accountability	Obligation to provide some justification for a decision and possibility to face sanctions if justifications are inadequate.	Most existing definitions focus on justifications and sanctions even though some uses of the term are vague and seem to equate accountability with responsibility or liability.

5. Technical issues and approaches

In this chapter, we review the technical issues and solutions available to meet the desiderata presented in Section 4.1. Sections 5.1, 5.2 and 5.3 focus on safety, security and privacy respectively while Section 5.4 is devoted to fairness and 5.5 to explainability. Considering that the disclosure of code and design documents raise more legal than technical issues, transparency is discussed in Chapter 6.

5.1. ADS Safety

Safety is an important issue to consider, especially in the case of ADS embedded in physical systems, whose failure can cause fatal damage. In this section, we use the word 'accident' for 'unintended and harmful behaviour that may emerge from systems when we specify the wrong objective function, are not careful about the learning process, or commit other machine learning-related implementation errors'.[92]

There has been a lot of discussion on extreme scenarios such as the risk of super-intelligence, i.e. the risk of machine intelligence surpassing human intelligence.[93] However, it is probably more useful at this stage to discuss the risks of less speculative scenarios. An illustrative example of failure that deserves attention is the 2016 Tesla autonomous car accident where a driver died in a fatal crash while using the autopilot mode.[94] In this case, the embedded sensors failed to distinguish a white tractor-trailer crossing the highway against a bright sky. This type of accident has to be identified and addressed before an ADS is deployed on a large scale.

In 2016, a Tesla autonomous car that was using autopilot mode was involved in a fatal crash. The autopilot sensors on the car failed to distinguish a white tractor-trailer crossing the highway against a bright sky.

Amodei and his co-authors explore several types of accidents related to machine learning, and present relevant research directions to protect against them.[95] The paper illustrates each type of accident with a fictional robot designed to clean up an office:

- **Negative side effects:** the ADS, while trying to achieve its goal, causes unintended and negative consequences in its environment. For example, the robot may knock over a vase when moving. Another example of an ADS with negative side effects is the Waze service. Very often, the routes suggested by Waze go through residential areas, creating anger and frustration for residents and city planners.[96] One potential solution to this problem is to develop what Gurses et al. call 'protection optimisation technologies (POTS)'.[97] POTS analyse how events affect users and environments, and then manipulate them to influence system outcomes, e.g. by altering the optimisation constraints and poisoning system inputs. Another direction in protection is to

---

92 Amodei, Dario, Olah Chris, Steinhardt Jacob, Christiano Paul, Schulman John, Mane Dan; Concrete Problems in AI Safety; arXiv:1606.06565; 2016; https://arxiv.org/pdf/1606.06565.pdf.

93 Nick Bostrom; Superintelligence: paths, dangers, strategies; OUP Oxford; 2014.

94 Danny Yadron, Dan Tynan; Tesla driver dies in first fatal crash while using autopilot mode; The Guardian; 2016; https://www.theguardian.com/technology/2016/jun/30/tesla-autopilot-death-self-driving-car-elon-musk.

95 Dario Amodei, Chris Olah, Jacob Steinhardt, Paul Christiano, John Schulman, Dan Mane;
Concrete Problems in AI Safety; arXiv:1606.06565; 2016; https://arxiv.org/pdf/1606.06565.pdf.

96 Elizabeth Weise; Waze and other traffic dodging apps prompt cities to game the algorithms; USA Today; 2017; https://eu.usatoday.com/story/tech/news/2017/03/06/mapping-software-routing-waze-google-traffic-calming-algorithmsi/98588980/.

97 Seda Gurses, Tebekah Overdorf, Ero Balsa; POTs: the revolution will not be optimized?; arXiv:1806.02711; 2018; https://arxiv.org/pdf/1806.02711.pdf.

penalise systems according to their detrimental impact on the environment. These approaches could mitigate accidents that can be anticipated, but cannot protect against unanticipated ones.

- **Reward hacking:** the ADS might game its reward function to increase its reward in an unintended way. For example, the cleaning robot might hide dust with materials it cannot see through. Note that from the ADS' point of view, this strategy is valid if it is a way to meet its objectives. Amodei et al. propose several preliminary approaches to prevent reward hacking. They suggest that using multiple awards might improve robustness, since they might be more difficult to game. Another proposed approach is to simply cap the maximum possible reward. The authors admit that it is very difficult to fully solve this problem, and that several protection approaches should probably be combined to make ADS more robust in this respect.

A cleaning robot might hide dust with materials it cannot see through. From the ADS' point of view, this strategy is valid if it is a way to meet its objectives.

- **Scalable oversight:** ADS need to be trained to be able to solve complex tasks. For example, the cleaning robot should learn to handle candy wrappers differently from stray cell-phones. However, the objective function might be too expensive to evaluate frequently or the training data might not be available. The challenge is to ensure that the ADS finds a way to do the right thing despite limited resources and information. Amodei et al. propose some potential counter-measures for semi-supervised reinforcement learning systems.

- **(Un)safe exploration:** Sometimes ADS need to engage in exploration and take actions to learn about their environment and update their model. However, exploration may involve performing dangerous actions, such as, for a cleaning robot, putting a mop in an electrical outlet. Common exploration policies may choose actions at random, or view unexplored actions optimistically towards the ADS goals. For anticipated and known dangerous actions, designers can, of course, make sure that they are avoided. However, in more complex domains, anticipating all possible dangerous actions is very challenging, if not impossible. Amodei et al. propose several approaches to mitigate the risks. One proposal is to confine exploration to safe actions. Another one is to use simulated exploration.

- **Robustness to distributional change:** Problems may occur when the training environment does not match the operational environment, or when there is a shift in the operational environment over time. For example, a speech recognition system trained on clean speech will perform poorly on noisy speech. Similarly, a cleaning robot trained to clean factory floors is likely to perform poorly if used to clean offices. One class of potential approaches is to train specialised ADS to specific environments. Another approach to this problem is to assume a partially specified model, in which only assumptions about some aspects of a distribution are made. Finally, another approach is to train the ADS on multiple distributions in the hope that the ADS will also perform well in a real environment.

While many of the failures described above can be addressed with ad hoc solutions, there is a strong need to define a unified approach to prevent ADS from causing unintended harm. A minimum requirement should be to perform extensive testing and evaluation before any large-scale deployment. It is also important to provide accountability, including the possibility of independent audits as discussed in Chapter 8, and to ensure a form of human oversight.

5.2. ADS Security

As discussed in Chapters 2 and 3, ADS will increasingly be used in critical contexts. It is therefore important to guarantee that they are secure against malicious adversaries.

The objectives of an adversary might be to breach the confidentiality, integrity or availability of the ADS.

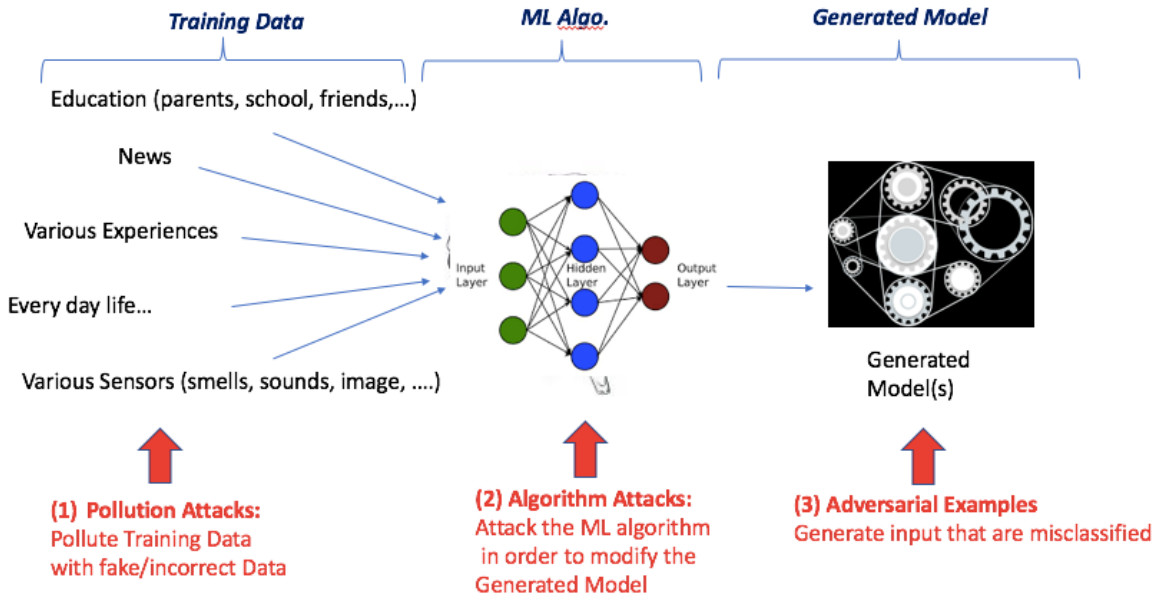
- The goal of a confidentiality attack is to extract part of the ADS internal state, typically the data or the model. For example, the adversary may want to retrieve some of the data that was used to train the model or to get information about the model itself. These attacks may have an impact on privacy or intellectual property.
- Attacks on integrity alter the results provided by the ADS. The adversary may want to modify the outputs of an ADS for example to make a gain or get some advantages (such as a loan or a job).
- Finally, attacks on availability disrupt the services provided by an ADS. The adversary might want, for example, to prevent the ADS from operating normally, by altering some of its parameters or performing a denial of service attack (DoS).

Since most ADS rely heavily on ML algorithms, in the rest of this section we consider security properties in the context of these algorithms. Furthermore, the confidentially property, which is related to privacy, will be addressed in the following section.

To illustrate this, one may consider an ADS built using a ML algorithm trained with a given dataset. As seen in figure 4, an adversary can threaten the integrity or availability of such ADS in different ways:

- by attacking the training dataset, for example, by injecting fake data,
- by attacking the ML algorithm, or
- by exploiting the generated model (the ADS) at run-time.

Figure 4 – Different types of integrity attacks on ML systems.



The attacks on the ML algorithm, sometimes called 'logic attacks', require the adversary to have physical access to the systems where the algorithm is running. These attacks are not specific to ADS and can be mitigated by various security measures, such as access control or hardware security. These measures are not discussed further. We focus on the attacks that target the training datasets (Section 5.2.1) or the generated model (Section 5.2.2) and possible protections against these attacks (Section 5.2.3).

5.2.1. Attacks on the training phase

The goal of an attack on the training phase is to influence the generated model by compromising its integrity or availability. Integrity attacks alter the generated model towards a specific goal, for example to maliciously obtain a loan or to go through an intrusion detection system (IDS).[98] For a ML classifier, the goal of an integrity attack could be to assign an incorrect class to a legitimate input. In contrast, availability attacks tend to affect the quality, performance or access to the system. The final goal may be to create sufficient errors to make the ADS unusable. Although their goals are different, these attacks are similar in nature and are typically performed by altering or poisoning the training dataset by injecting adversarial data (**injection attacks**) or by removing or modifying existing records (**modification attacks**). The modification can be performed, in a supervised setting, by modifying the data labels[99] or the data itself.[100] Note that these attacks require that the adversaries have access to the pre-processed training dataset. If this is not possible, the adversary can poison or inject the training data before pre-processing. For example, Perdisci et al. showed that it is possible to prevent a work signature detection tool from learning valid signatures by polluting the worm traffic data-flows.[101]

5.2.2. Attacks on the execution phase

Attacks on the execution phase do not intend to modify the ADS generated model, but instead seek to exploit some of its weaknesses. The idea is to compute some inputs, called **adversarial examples**, which will trigger the desired, incorrect, outputs.[102] When the ADS is a classifier, the adversary seeks to have the perturbed inputs assigned to incorrect classes. An example of this in an image recognition system is presented in figure 5. This system successfully recognises a panda. However, by adding a little bit of noise, the authors of this study show that the resulting image, which still looks like a panda to a human, is misclassified by the algorithm as a gibbon.[103]

---

[98] Marco Barreno, Blaine Nelson, Russell Sears, Anthony D. Joseph, J. Doug Tygar; Can machine learning be secure?; *ACM Symposium on Information, computer and communications security* (ASIACCS '06); ACM; http://dx.doi.org/10.1145/1128817.1128824.

[99] Battista Biggio, Blaine Nelson, Pavel Laskov; Support vector machines under adversarial label noise; Asian Conference on Machine Learning; 2011.

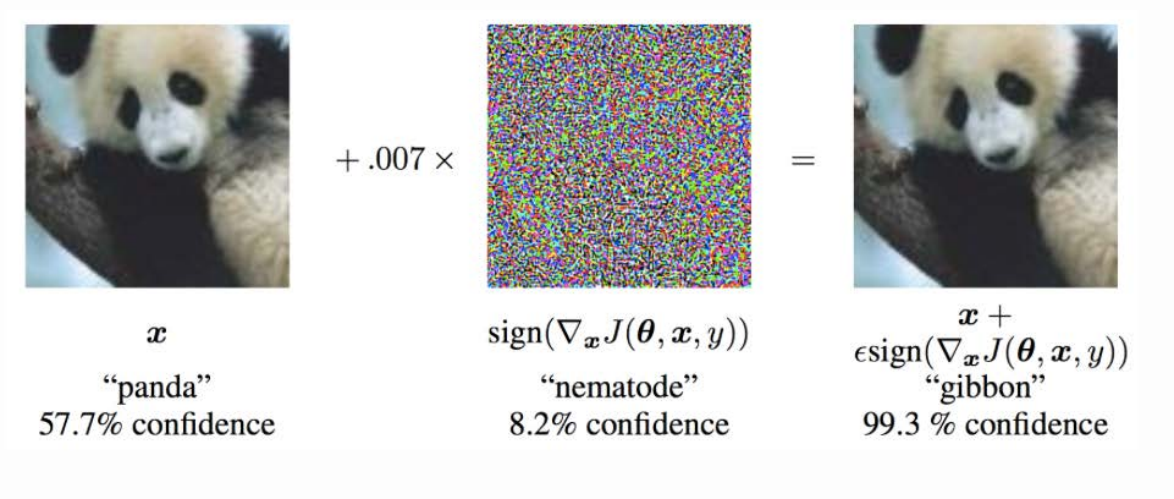
[100] Marius Kloft, Pavel Laskov; Online anomaly detection under adversarial impact; International Conference on Artificial Intelligence and Statistics; 2010.

[101] Roberto Perdisci, David Dagon, Wenke Lee, Prahlad Fogla, Monirul Sharif; Misleading worm signature generators using deliberate noise injection; IEEE Symposium on Security and Privacy; IEEE; 2006.

[102] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, Rob Fergus; Intriguing properties of neural networks; International Conference on Learning Representations; Computational and Biological Learning Society ; 2014.

[103] Patrick McDaniel, Nicolas Papernot, Z. Berkay Celik; Machine learning in adversarial settings; IEEE Security and Privacy; (14,3); 2016.

Figure 5 – Demonstration of fast adversarial example generation applied to GoogleNet on ImageNet



Source: https://arxiv.org/pdf/1412.6572.pdf.

By adding an imperceptibly small vector whose elements are equal to the sign of the elements of the gradient of the cost function with respect to the input, one can change the GoogleNet classification of the image.

This attack can, for example, be exploited by an adversary against an image recognition system. For example, Sharif et al. have developed inconspicuous attacks against biometric systems that allow an attacker to evade recognition or impersonate another individual.[104] Their attacks, illustrated in figure 6, are carried out through printing a pair of spectacle frames that allow the attacker to evade recognition (**dodging attack**), or even to impersonate another individual (**impersonation attack**).

Figure 6 – Impersonation using spectacle frames

Left: Actress Reese Witherspoon, Image classified correctly with probability 1. Middle: Perturbing frames to impersonate (actor) Russell Crowe. Right: The target Russell Crowe.

Source: https://www.cs.cmu.edu/~sbhagava/papers/face-rec-ccs16.pdf.[105]

Without too much explanation of the technical details, these attacks work by perturbing an input *x* with the smallest possible noise *r,* such that the resulting adversarial example, $x^* = x + r$ remains in the correct input domain *D,* but is assigned to the wrong label. These adversarial examples are

[104] Mahmood Sharif, Sruti Bhagavatula, Lujo Bauer, Michael K. Reiter; Adversarial generative nets: neural network attacks on state-of-the-art face recognition; arXiv:1801.00349; 2017.

[105] Mahmood Sharif, Sruti Bhagavatula, Lujo Bauer, Michael K. Reiter; Accessorize to a crime: Real and stealthy attacks on state-of-the-art face recognition; ACM SIGSAC Conference on Computer and Communications Security; ACM; 2016.

possible because the ADS model is not perfect and does not perfectly match the actual decision model. This is illustrated by figure 7(a), which shows a system that classifies images of pandas and gibbons. The black line defines the human decision boundary.[106] The dotted line defines the model decision boundary.[107] Since the ADS model is not perfect, its decision boundary does not exactly match the human decision boundary. As shown in figure 7(b), it is therefore possible to manipulate an image of a panda by adding a little bit of noise so that it moves outside the model decision boundary, and is therefore classified as a gibbon, but remains within the human decision boundary (it is seen as a panda by a human being).

Goodfellow et al. introduced the **fast gradient sign method** to generate adversarial examples.[108] Follow-up work optimised the method by reducing the size[109] of the perturbation, or by minimising the number of perturbed features.[110]

Figure 7(a) – Image classification system (gibbons versus pandas).

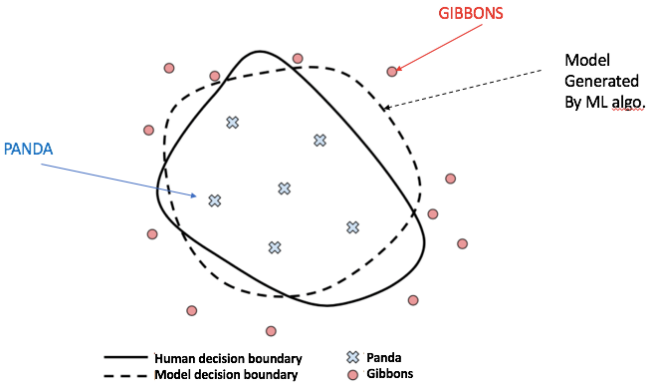
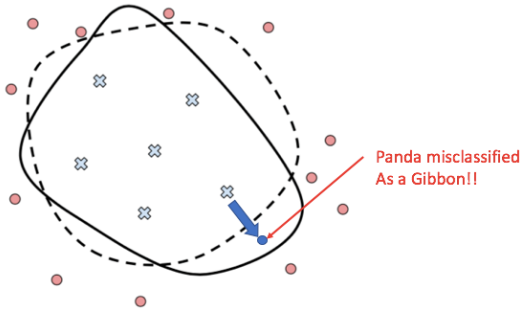


Figure 7(b) – Generating adversarial examples.



[106] All images that are within the area delimited by the black line are pandas, whereas images that are outside of this boundary are gibbons.

[107] All images that are within the area delimited by the dotted line are classified by the model as pandas, whereas images that are outside of this boundary are classified as gibbons.

[108] Ian J. Goodfellow, Jonathon Shlens, Christian Szegedy; Explaining and harnessing adversarial examples; International Conference on Learning Representations; Computational and Biological Learning Society; 2015.

[109] Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, Pascal Frossard; Deepfool: a simple and accurate method to fool deep neural networks; arXiv:1511.04599; 2015.

[110] Nicolas Papernot, Patrick McDaniel, Somesh Jha, Matt Fredrikson, Z Berkay Celik, Ananthram Swami; The limitations of deep learning in adversarial settings; 1st IEEE European Symposium on Security and Privacy; IEEE; 2016.

Adversarial examples are typically constructed by perturbing input data. However, as shown in a recent paper, it is also possible to define a class of adversarial examples that are synthesised entirely using a conditional generative model.[111]

Note that most of the previous attacks assume 'white box' scenarios, in which attackers have access to the internal workings of the model. However, the 'black box' scenario is probably a more realistic threat model. For example, an attacker who wants to attack an image recognition system or a spam filter rarely has access to the internals of the model. Instead, they often have access to the system as an oracle, i.e. it can query the ADS with their own inputs and can observe the generated outputs. Attacks on 'black box' systems, also called 'black box' attacks, are more challenging but not impossible. A key property in this respect is adversarial example **transferability**, i.e. the property that can be exploited whereby adversarial examples crafted for a given classifier are likely to be misclassified by other models trained for the same task.[112] Intriguingly, this property holds even when the models are trained with different datasets. Papernot et al. designed a 'black box' attack based on this property.[113] The proposed attack queries the ADS, and then exploits the results to generate a substitute model. The substitute model is then used to craft adversarial examples that are misclassified by the ADS.

5.2.3. Protections against ADS security attacks

Attacks on the training phase. Most defence mechanisms against attacks on the training phase rely on the fact that poisoning samples are outliers that are typically outside the expected input distribution. Rubinstein et al.[114] propose a solution against poisoning attacks relying on techniques from robust statistics. They show that poisoning has little effect on the robust model, whereas it significantly distorts the model produced by the original principal component analysis (PCA) method.

Another proposal to secure the training phase relies on the regularisation of the optimisation problems solved to train ML models.[115] This technique has the effect of smoothing the solution, which removes the complexity that an adversary may try to exploit. The authors also propose the use of disinformation techniques to alter the data seen by the adversary to prevent them from learning the decision boundaries. Finally, they suggest using randomisation in the placement of the boundary in order to make the attacks more difficult.

Attacks on the execution phase. Two main strategies can be used to protect an ADS against adversarial examples. The first strategy is **reactive**: it attempts to detect adversarial examples. Solutions of this category include ***adversarial detecting,***[116] **input reconstruction**[117] and **network**

[111] Yang Song, Rui Shu, Nate Kushman, Stefano Ermon; Generative Adversarial Examples; arXiv:1805.07894; 2018.

[112] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, R. Fergus; Intriguing properties of neural networks; arXiv:1312.6199; 2017.

[113] N. Papernot, P. McDaniel, I. Goodfellow, S. Jha, Z. B. Celik, A. Swami; Practical black-box attacks against deep learning systems using adversarial examples; arXiv:1602.02697; 2016.

[114] B. I. P. Rubinstein, B. Nelson, L. Huang, A. D. Joseph, S.-h. Lau, S. Rao, N. Taft, J. D. Tygar; Antidote: Understanding and defending against poisoning of anomaly detectors; 9th ACM SIGCOMM Conference on Internet measurement; ACM; 2009.

[115] M. Barreno, B. Nelson, R. Sears, A. D. Joseph, J. D. Tygar; Can machine learning be secure?; ACM Symposium on Information, computer and communications security; ACM; 2006.

[116] J. H. Metzen, T. Genewein, V. Fischer, B. Bischoff; On detecting adversarial perturbations; 5th International Conference on Learning Representations (ICLR); 2017.

[117] S. Gu, L. Rigazio; Towards deep neural network architectures robust to adversarial examples; International Conference on Learning Representations (ICLR); 2015.

V-verification.[118] The second strategy is 'proactive': it aims at making systems more robust to adversarial examples. Solutions of this category include **network distillation**,[119] **adversarial (re)training**[120] and **classifier robustifying**.[121]. Due to the variety of adversarial examples, several defence strategies can be performed together (in parallel or sequentially) to deal with them.

It is out of the scope of this document to present these solutions in detail. For more information about these defence mechanisms, we refer interested readers to the survey papers by Papernot et al.[122] and Yand et al.[123]. It is however worth noting here that existing defence solutions are unsatisfactory. A recent study analysed ten detection proposals and showed that they can all be defeated.[124] More effective solutions still need to be proposed and evaluated.

5.3. ADS Privacy

An adversary may want to compromise the confidentiality of an ADS for example by trying to extract information about the training data or by retrieving the ADS model itself. These attacks raise privacy concerns, since training data often contain personal data. They may also undermine intellectual property, as the ADS model and the training data can be proprietary and confidential to their owner.

The rest of this section describes these two types of attack and presents some solutions to protect ADS against them.

5.3.1. Extraction of training data

Attackers may want to retrieve some of the data used to train the system. Two main types of scenarios can be considered:

- 'White box' attacks rely on the assumption that the attacker has access to the model and tries to learn about the training data by 'inverting' it.
- 'Black box' attacks do not assume access to the model: an adversarial client can only submit queries to the model and make predictions based on the answers.

Most of the research work in this area focuses on 'black box' attacks because they are more realistic and more powerful. Fredrikson et al.[125] have defined a **model inversion attack**, in the context of genomic privacy, which is able to use 'black box' access to prediction models to estimate aspects of the genotype of a person. Their attack works for any setting in which the inferred feature is drawn from a small set. In follow-up work, Fredrikson et al.[126] demonstrated how the confidence

[118] G. Katz, C. Barrett, D. Dill, K. Julian, M. Kochenderfer; Reluplex: An efficient SMT solver for verifying deep neural networks; arXiv:1702.01135, 2017.

[119] N. Papernot, P. McDaniel, X. Wu, S. Jha, A. Swami; Distillation as a defense to adversarial perturbations against deep neural networks; 2016 IEEE Symposium on Security and Privacy (SP); IEEE; 2016.

[120] I. J. Goodfellow, J. Shlens, C. Szegedy; Explaining and harnessing adversarial examples; arXiv:1412.6572; 2014.

[121] J. Bradshaw, A. G. d. G. Matthews, Z. Ghahramani; Adversarial examples, uncertainty, and transfer testing robustness in gaussian process hybrid deep networks; arXiv:1707.02476; 2017.

[122] Nicolas Papernot, Patrick McDaniel, Arunesh Sinha, Michael Wellman ; Towards the Science of Security and Privacy in Machine Learning; 3rd IEEE European Symposium on Security and Privacy; 2016.

[123] Xiaoyong Yuan, Pan He, Qile Zhu, Rajendra Rana Bhat, Xiaolin Li; Adversarial Examples: Attacks and Defenses for Deep Learning; arXiv:1712.07107; 2017.

[124] Nicholas Carlini, David Wagner; Adversarial Examples Are Not Easily Detected: Bypassing Ten Detection Methods; ACM Workshop on Artificial Intelligence and Security; 2017.

[125] M. Fredrikson, E. Lantz, S. Jha, S. Lin, D. Page, T. Ristenpart; Privacy in pharmacogenetics: An end-to-end case study of personalized warfarin dosing; USENIX Security Symposium; 2014.

[126] M. Fredrikson, S. Jha, T. Ristenpart; Model inversion attacks that exploit confidence information and basic countermeasures; 22nd ACM SIGSAC Conference on Computer and Communications Security; ACM; 2015.

information returned by many ML classifiers enables new model inversion attacks. Ateniese et al.[127] showed that it is possible to infer information about ML classifier training sets by building a novel meta-classifier and training it to hack other classifiers.

Membership attack is a specific type of model inversion attack, where the attacker is seeking to test whether a given point was used in the training dataset. Shokri et al. show how to conduct this type of attack against 'black box' models [128]. The proposed attack turns machine learning against itself by training an attack model that can distinguish the target model's outputs on members versus non-members of its training dataset. They basically turn the membership inference problem into a classification problem. They successfully demonstrated their attack against 'black box' models trained in the cloud using Google prediction API and Amazon ML.

5.3.2. Model extraction

Attackers may also seek to recover information about the model of the ADS. It is generally assumed that these attackers can freely query the ADS and observe its outputs. Model extraction attacks may undermine privacy since, as discussed above, the model can be used to retrieve some of the training data. They may also have intellectual property implications when the model is proprietary and should remain confidential.

Tramer and al. show how to extract the parameters of a model from its predictions.[129] Their most successful attacks rely on the information returned by the ML prediction APIs of cloud-based services such as those provided by Google, Amazon and Microsoft. These services return high-precision confidence values in addition to class labels. By querying $d + 1$ random d-dimensional inputs, an attacker can, with high probability, solve the unknown $d + 1$ parameters defining the model. This model extraction attack, although simple and non-adaptive, can be applied to many systems. The most obvious countermeasure is to restrict the information provided by the ML services, but this information may be of interest to the users.

5.3.3. Toward privacy-preserving solutions

Various proposals have been made to address the privacy attacks presented in Sections 5.3.1 and 5.3.2. Most of them anonymising the training datasets and the generated models, i.e. designing privacy-preserving ML algorithms. The anonymisation model most frequently used in this context is differential privacy,[130] a rigorous framework to anonymise and analyse the privacy guarantees provided by algorithms. For example, Abadi et al.[131] introduce an algorithm for non-convex deep learning models with strong differential privacy guarantees.

---

[127] G. Ateniese, L. V. Mancini, A. Spognardi, A. Villani, D. Vitali, G. Felici; Hacking smart machines with smarter ones: How to extract meaningful data from machine learning classifiers; International Journal of Security and Networks; (10,3); 2015.

[128] R. Shokri, M. Stronati, V. Shmatikov; Membership inference attacks against machine learning models; IEEE Security and Privacy; 2017.

[129] F. Tramer, F. Zhang, A. Juels, M. K. Reiter, T. Ristenpart; Stealing machine learning models via prediction APIs; Usenix Security; 2016.

[130] C. Dwork, A. Roth et al.; The algorithmic foundations of differential privacy; Foundations and Trends in Theoretical Computer Science; (9); 2014.

[131] M. Abadi, A. Chu, I. Goodfellow, H. B. McMahan, I. Mironov, K. Talwar, L. Zhang; Deep learning with differential privacy; ACM CCS; 2016.

In order to reduce the risk of data leakage, some proposals rely on the distribution of the learning phase. In other words, the training data does not leave the devices which collect them. The main idea behind this approach, called **federated learning**, is to learn a shared model by aggregating locally-computed updates.[132] As an illustration, Shokri et al.[133] support distributed training of deep learning networks in a privacy-preserving way using differential privacy. Their system relies on the input of independent entities which collaborate to build a ML model without sharing their training data. To this end, they selectively share subsets of noisy model parameters during training. Their approach makes it possible to provide a solution for multiple organisations, for example hospitals, which combine their data to train a deep-learning model, but without having to share it.

In order to reduce the risk of data leakage, some proposals rely on the distribution of the learning phase. In other words, the training data does not leave the devices that collect them.

Another solution is the CryptoNets[134] system which is based on neural networks and can be applied to encrypted data. It allows a user to upload encrypted data to a cloud service that can then apply a neural network to the data without accessing the plaintext. The encrypted prediction can then be returned to the user who decrypts it. As a result, the cloud service does not gain any information about the raw data nor about the predictions, since they are both encrypted, using what is commonly referred to as 'homomorphic encryption'.

5.4. ADS Fairness

As ADS replace or support human decision-makers in a number of sensitive domains such as justice, health or education, it is important to ensure that they do not result in decisions that are considered unfair or discriminatory. In this section, we first discuss the various sources of unfairness (Section 5.4.1), before presenting several definitions of fairness (Section 5.4.2) and technical solutions to build fairness-aware ADS (Section 5.4.3). We conclude with comments on potential tensions and trade-offs between different objectives (Section 5.4.4).

5.4.1. The various sources of unfairness

ADS are often based on machine learning algorithms trained on collected data. There are multiple potential sources of unfairness in this process.[135] Unfair treatment can result, for example, from the content of the training data, the way the data is labelled or the feature selection.[136]

Biased training data. If the training data contains biases or historical discriminations, the ADS will inherit them and incorporate them into its future decisions. For example, as illustrated in figure 8, word embeddings trained on Google News articles exhibit gender stereotypes that are propagated on a daily basis.[137]

---

[132] H. Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, Blaise Agüera y Arcas*;* Communication-Efficient Learning of Deep Networks from Decentralized Data ; arXiv:1602.05629; 2016.

[133] R. Shokri and V. Shmatikov; Privacy-preserving deep learning; ACM Computer Communication Security (CCS); 2015.

[134] Gilad-Bachrach, Ran, et al. ; Cryptonets: Applying neural networks to encrypted data with high throughput and accuracy; International Conference on Machine Learning; 2016.

[135] Barocas, Solon, Selbst, Andrew D.; Big Data's Disparate Impact; 104 California Law Review; (671); https://ssrn.com/abstract=2477899; 2016.

[136] Gal Yona ; A Gentle Introduction to the Discussion on Algorithmic Fairness; 2017; https://towardsdatascience.com/a-gentle-introduction-to-the-discussion-on-algorithmic-fairness-740bbb469b6.

[137] Tolga Bolukbasi, Kai-Wei Chang, James Zou, Venkatesh Saligrama, Adam Kalai ; Man is to computer programmer as woman is to homemaker? debiasing word embeddings; *30th International Conference on Neural Information Processing Systems* (NIPS'16); 2016.

Figure 8 – The most extreme occupations as projected on to the she/he gender direction on g2vNEWS

Extreme *she* occupations

1. homemaker	2. nurse	3. receptionist
4. librarian	5. socialite	6. hairdresser
7. nanny	8. bookkeeper	9. stylist
10. housekeeper	11. interior designer	12. guidance counselor

Extreme *he* occupations

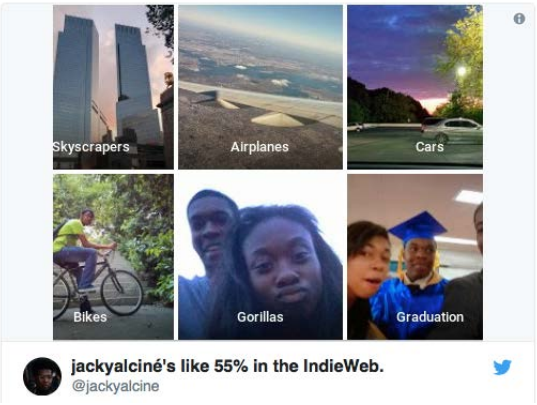
1. maestro	2. skipper	3. protege
4. philosopher	5. captain	6. architect
7. financier	8. warrior	9. broadcaster
10. magician	11. figher pilot	12. boss

Source: https://arxiv.org/pdf/1607.06520.pdf)

A solution to mitigate this issue is to constantly re-train machine learning models with 'fresh' data, under the (optimistic) assumption that society evolves and historical bias will correct itself with time. Another direction is to try to eliminate bias from the training data, by pre-processing it to remove existing biases.[138] This task is challenging, since not all biases are known and many of them can be indirect. For instance, as shown by Tolga Bolukbasi and his co-authors,[139] the fact that the word **receptionist** is much closer semantically to **softball** than **football** may arise from the female associations with both **receptionist** and **softball**.

Accuracy disparity. In 2015, Jacky Alciné, a Brooklyn resident, noticed while browsing his Google Photos app that pictures of him and a friend, both of whom are black, were tagged under the label 'gorillas' (see figure 9). This mistake was clearly not intentional, but resulted from an error by the Google image classification algorithm.

Figure 9 – Example of accuracy disparity: incorrect tagging of pictures



Source: https://mashable.com/2015/07/01/google-photos-black-people-gorillas/#SkdKmWWBtuqQ)

ADS, and more generally machine learning algorithms, are systems trained to recognise and leverage statistical patterns in data. However, they are not perfect and perform classification or prediction errors. The accuracy rate of an ADS is often related to the size of the training dataset: a

[138] Tolga Bolukbasi, Kai-Wei Chang, James Zou, Venkatesh Saligrama, Adam Kalai; Man is to computer programmer as woman is to homemaker? debiasing word embeddings; *30th International Conference on Neural Information Processing Systems* (NIPS'16); 2016.

[139] Ibid.

large training dataset leads to less errors, and less data leads to worse predictions. Minorities tend to be under-represented in datasets, and are therefore subject to much poorer accuracy. This is considered unfair, since different groups of the population get different prediction error rates and they cannot do much about it. A recent study evaluated three commercial gender classification systems and showed that darker skinned females are the most misclassified group (with error rates of up to 34.7 %). All classifiers returned better results for lighter skinned individuals and males and the worst performances of all were observed for darker skinned females (see figure 10).[140] This is not because darker females are more difficult to classify, but simply because they were under-represented in the training datasets.

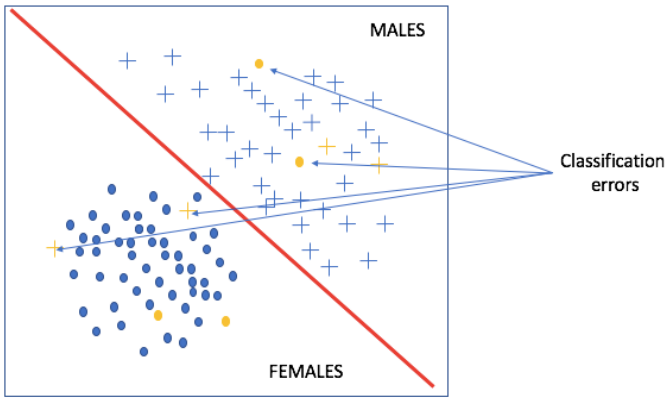
Figure 10 – Accuracy of facial recognition systems

Source: https://www.nytimes.com/2018/02/09/technology/facial-recognition-race-artificial-intelligence.html)

It is therefore important to evaluate the performance of ADS systems for different minorities. In fact, in a system that achieves 95 % accuracy, the 5 % error may be uniformly distributed over the whole population or result from very good accuracy for the majority and very poor accuracy for some minorities. Since minorities are, by definition, smaller groups, their impact on the overall performance of the system can be negligible. This argument is illustrated by the toy example in figure 11. It shows a simple linear binary classification that classifies the inputs of a dataset according to their gender i.e. into male (crosses) or female (points). This dataset is composed of members of a majority group (blue) and members of a minority group (yellow) according to some protected attribute. The accuracy of the classification algorithm is perfect, i.e. equal to 100 %, for the members of the majority group (the blue points and crosses are perfectly separated), but perform very poorly (50 % accuracy), for the members of the minority group (the yellow points and crosses are not separated by the line). However, since the size of the minority group is small compared to the overall population, the overall accuracy remains very high.

140 Joy Buolamwini, Timnit Gebru ; 1st Conference on Fairness, Accountability and Transparency; PMLR; (81); 2018.

Figure 11 – A simple classification example



Automated decisions tend to treat those who belong to the statistically dominant groups more accurately because they are over-represented in the training datasets. Differences in classification accuracy between different groups are a major and underappreciated source of unfairness.

5.4.2. Definitions of fairness

Discussions of fairness in ADS are often too rhetorical and lack rigour and precision. In fact, characterising the notion of fairness is far from trivial and many different and sometimes incompatible definitions have been proposed.[141] There is a growing body of work on this topic and we point the reader to Barocas and Selbst[142] and the survey by Romei and Ruggieri[143] for more comprehensive introductions. At a high level, existing definitions of fairness usually rely on 'protected groups'. These are defined via sensitive attributes such as race or gender. They then define some statistical properties and require them to be approximately equalised across these groups. For example, a definition (called **disparate impact** or **statistical parity**) can require that the rate of positive classification be equal across the groups. Another (called **equalized odds**), requires that the false positive and false negative rates be equal across the groups. In the rest of this section, we present some of the existing definitions in greater detail.

Disparate treatment and impact. Anti-discrimination laws generally distinguish **disparate treatment** and **disparate impact**. Disparate treatment addresses intentional discrimination. This includes: (i) decisions explicitly based on a protected characteristic and (ii) intentional discrimination via proxy variables. A treatment is disparate when it depends on protected attributes. More formally, an ADS does not suffer from disparate treatment if:

$$P(Decision=Accept|\, x, z) = P(Decision=Accept|x)$$

where x is a vector of non-sensitive attributes of a user, z is the sensitive attribute (such as race) and the expression $P(A|x,z)$ means the probability of A, knowing x and z.

A simple idea to achieve this property is to remove the sensitive attributes from the training data. However, only removing the sensitive attributes will often be insufficient, since protected attributes can be redundantly encoded into other attributes, i.e. they can be correlated with them. For example, it is well known that, in certain cities, there is a strong correlation between the religion or

[141] Richard Berk, Hoda Heidari, Shahin Jabbari, Michael Kearns, Aaron Roth; Fairness in Criminal Justice Risk Assessments: The State of the Art; Sociological Methods & Research; 2018.

[142] Barocas, Solon Selbst, Andrew D.; Big Data's Disparate Impact; 104 California Law Review; (671); https://ssrn.com/abstract=2477899; 2016;

[143] Romei, A., Ruggieri, S.; A multidisciplinary survey on discrimination analysis; The Knowledge Engineering Review; (29,5); doi:10.1017/S0269888913000039; 2014.

ethnic origin of an individual and the area code of the part of the city where they live. Removing the religion or ethnic origin attributes from the dataset will therefore not be sufficient, since they can be predicted with good accuracy from the area code information. Identifying all correlated attributes can be very challenging. A promising approach uses machine learning for this task.[144]

As explained above, because attributes are correlated, the fact that an ADS does not exhibit disparate treatment does not necessarily mean that it will not impose disparate impacts on a particular group. A decision has a disparate impact when it has a 'disproportionately adverse' effect on members of a protected group. In other words, the rate of positive classification (acceptance) in each protected group should be very similar. For example, if race is the sensitive attribute, an algorithm has no disparate impact if:

$$P\{Decision = Accept|Race = White)\sim P\{Decision = Accept|Race = Black\}$$

Note that 'disproportionately adverse' is often defined using the 80 % Rule:[145] the ratio between the two probabilities should not be less than 0.80.

Equalized predictive values (EPV). EPV is another measure of fairness, which is widely accepted and adopted by the psychometrics community.[146] It guarantees that a system is fair in the sense of being free of predictive biases.

As an illustration, considering the FICO credit score mentioned in Section 3.1, EPV basically states that, supposing a person was given a positive decision, the probability that they pay back the loan should be equalised across different groups:

$$P\{Will\ pay\ back|Race = White, Decision = Accept\}\sim P\{Will\ pay\ back|Race = Black, Decision = Accept\}$$

Disparate mistreatment and 'equalised odds'. A recent paper proposes the definition of **disparate mistreatment** that measures how the misclassification rates of an algorithm differ for different groups.[147] In other words, instead of studying and comparing the outcomes of a decision algorithm, the authors propose to verify that the error rates of the algorithm are similar for different groups. The intuition is that an ADS potentially causes harm to an individual when it misclassifies them. They propose to use five different types of error rates: overall misclassification rate; false positive rate; false negative rate; false omission rate; and false discovery rates. Similarly, Hardt et al. proposed an alternative called **equalised odds**, which also states that the error (misclassification) should be equalised across different groups.[148] Taking the example of the FICO credit score application, the following properties should be satisfied:

- Equalised true positive ensures that people who pay back their loan have equal opportunity to get a loan:

$$P\{Decision = Accept|Race = White, Will\ pay\ back\}\sim P\{Decision = Accept|Race = Black, Will\ pay\ back\}$$

[144] Rich Zemel, Yu Wu, Kevin Swersky, Toni Pitassi, Cynthia Dwork ; 30th International Conference on Machine Learning; PMLR; (28,3); 2013.

[145] For example, the US Equal Employment Opportunity Commission refers to a 80% rule as a measure of disparate impact.

[146] A Chouldechova ; *Fair prediction with disparate impact: A study of bias in recidivism prediction instruments*, Big data ; (5,2); 2017.

[147] Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez Rodriguez, and Krishna P. Gummadi ; Fairness Beyond Disparate Treatment & Disparate Impact: Learning Classification without Disparate Mistreatment; *26th International Conference on World Wide Web* (WWW '17); 2017.

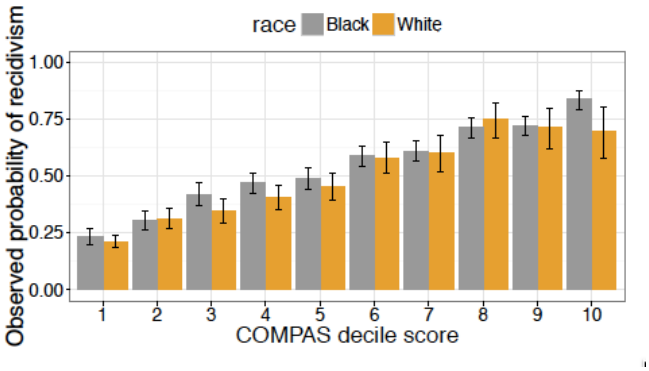
[148] Moritz Hardt, Eric Price, Nathan Srebro; Equality of opportunity in supervised learning ; *30th International Conference on Neural Information Processing Systems* (NIPS'16); 2016.

- Equalised false positive ensures that people who do not pay back their loan have equal opportunity to get a loan:

$$P\{Decision = Accept|Race = White, Won't\ pay\ back\}\sim P\{Decision = Accept|Race = Black, Won't\ pay\ back\}$$

Incompatibility of different definitions of fairness. Considering that several definitions of fairness exist, it is useful to understand how they relate to each other. Unfortunately, many of them are incompatible. For example, Chouldechova showed, in a mathematical sense, that it is impossible to develop a system that simultaneously satisfies 'equalised odds' and 'equalised predictive values' definitions.[149] Chouldechova illustrated the results using the COMPAS application and showed that although COMPAS satisfies EPV (see figure 12), it does not satisfy 'equalised odds' (see figure 13), and therefore disparate has impacts on black people.

Figure 12 – Observed probability of recidivism according to the score provided by the COMPAS system

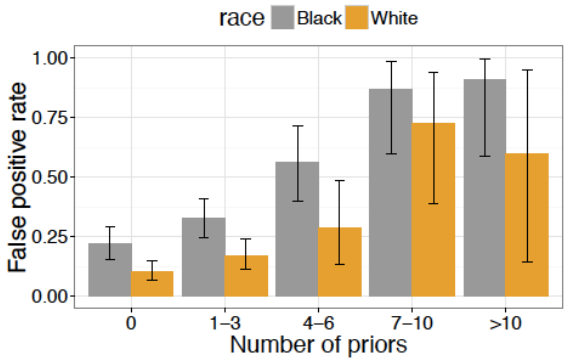


Source: *https://arxiv.org/pdf/1610.07524.pdf*

The results in figure 12 show that these probabilities are very similar for black and white defendants.

[149] Alexandra Chouldocheva; Fair prediction with disparate impact: a study of bias in recidivism prediction instruments; Big Data, Special issue on Social and Technical Trade-Offs; 2017.

Figure 13 – False positive rates across prior record count for defendants charged with a misdemeanour offence using the COMPAS data made available by ProPublica[150]



Source: https://arxiv.org/pdf/1610.07524.pdf.

The results in figure 13 show that the false positive rates are much larger for black than for white defendants.

One of the lessons to be drawn from these incompatibility results is that experts can only provide precise definitions and explain them, whereas the ultimate choices in terms of fairness are not technical, but a matter of public policy.

5.4.3. Towards fairness-aware algorithms

Several research groups have focused on the design of ADS that attempt to address fairness and discrimination issues.[151] A detailed description of these schemes is outwith the scope of this report. In a nutshell, fairness adds an extra constraint to the learning algorithm. It identifies the parameters (i.e. hypothesis) that minimise the classification errors on the training data, while satisfying the fairness constraint. There is therefore generally a trade-off between fairness and accuracy.

Fairness-aware ADS rely on one of the following approaches:[152]

- **Pre-processing approach:** This consists of pre-processing the training data to remove the sources of unfairness or to map the training data into a space where the dependencies between sensitive attributes and class labels disappear.[153] Kamiran and Calders,[154] and Hajian *et* al.[155] adopt this approach by performing a controlled distortion of the training data that leads to an

[150] Jeff Larson, Surya Mattu, Lauren Kirchner and Julia Angwin; How We Analyzed the COMPAS Recidivism Algorithm; Propublica; https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm; 2016.

[151] Michael Feldman, Sorelle A. Friedler, John Moeller, Carlos Scheidegger, Suresh Venkatasubramanian; Certifying and Removing Disparate Impact; *21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (KDD '15); 2015.

[152] Hajian S., Domingo-Ferrer J.; Direct and Indirect Discrimination Prevention Methods, in Discrimination and Privacy in the Information Society; Studies in Applied Philosophy, Epistemology and Rational Ethics; (13); Springer; 2013.

[153] C. Dwork, M. Hardt, T. Pitassi, O. Reingold, and R. Zemel; Fairness through awareness; Innovations in Theoretical Computer Science; 2012. M. Feldman, S. A. Friedler, J. Moeller, C. Scheidegger, S. Venkatasubramanian; Certifying and Removing Disparate Impact; KDD; 2015. F. Kamiran, T. Calders; Classification with No Discrimination by Preferential Sampling; BENELEARN; 2010.

[154] Kamiran F., Calders T.; Classification with no discrimination by preferential sampling; 19th Machine Learning conference of Belgium and The Netherlands; 2010.

[155] Hajian S., Domingo-Ferrer J., Martinez-Balleste A.; Rule protection for indirect discrimination prevention in data mining; Modeling Decisions for Artificial Intelligence (MDAI 2011); Lecture Notes in Computer Science; (6820); Springer; 2011.

unbiased dataset. The main limitation of this approach is that it treats the learning algorithm as a 'black box', which can result in an unpredictable loss in accuracy.

- **In-processing approach:** The second approach modifies the classifier algorithm to limit discrimination.[156] For example, through a novel leaf re-labelling approach, Calders and Verwer[157] propose embedding a non-discriminatory constraint into the algorithm decision tree learner by changing its splitting criterion and pruning strategy.
- **Post-processing approach:** Instead of cleaning the original dataset or changing the data mining algorithms, this approach modifies the resulting data mining models. For example, Pedreschi et al.[158] propose a confidence-altering version of the CPAR algorithm (classification based on predictive association rules).[159]

The above approaches are complementary and they can be combined. For example, Zemel et al.[160] propose a scheme that applies the first two approaches by jointly learning a fair representation of the data and the classification parameters. Many of the existing proposals are restricted to a narrow range of classifiers and can only accommodate a single, binary, sensitive attribute. In other words, they do not generalise to multiple (e.g. gender and race) or polyvalent sensitive attributes (e.g. race, that has more than two values). However, more practical schemes will undoubtedly be proposed in forthcoming years.

5.5. ADS Explainability

Technical solutions for explainability can be classified according to different dimensions:Technically speaking, three main approaches can be followed to implement explainability requirements:

- **The 'black box' approach:** this approach consists in analysing the behaviour of the ADS without 'opening the hood', that is to say without any knowledge of its code. Explanations are constructed from observations of the relationships between the inputs and the outputs of the system. This is the only possible approach when the operator or provider of the ADS is uncollaborative (does not agree to disclose the code).

The first challenge for 'black box' explanations is the construction of explanations based on observations of the ADS. In some cases, the observation itself is also a challenge because the ADS is integrated within a complex system involving multiple parties.

- **The 'white box' approach:** in contrast to the 'black box' approach, the 'white box' approach assumes that it is possible to analyse the ADS code.

[156] T. Kamishima, S.Ahako, H.Asoh, J. Sakuma; Fairness-aware Classifier with Prejudice Remover Regularized; PADM; 2011. G. Goh, A. Cotter, M. Gupta, M. Friedlander; Satisfying Real-world Goals with Dataset Constraints; NIPS; 2016. T. Calders, S. Verwer; Three naive bayes approaches for discrimination-free classification; Data Mining journal; special issue with selected papers from ECML/PKDD; 2010.

[157] Calders T., Verwer S.; Three naive Bayes approaches for discrimination-free classification. Data Mining and Knowledge Discovery; (21,2); 2010.

[158] Pedreschi D., Ruggieri S., Turini F.; Measuring discrimination in socially-sensitive decision records; 9th SIAM Data Mining Conference (SDM 2009); 2009.

[159] Yin X. & Han J.; CPAR: Classification based on Predictive Association Rules; SIAM ICDM; 2003.

[160] R. Zemel, Y. Wu, K. Swersky, T. Pitassi, C. Dwork; Learning fair representations; Intl. Conf. on Machine Learning; 2013.

- **The constructive approach:** in contrast to the first two approaches, which assume that the ADS already exists, the constructive approach is the design of ADS taking into account explainability requirements ('explainability by design').

In the following sections, we discuss the 'black box' techniques (Section 5.5.1), the 'white box' techniques (Section 5.5.2) and the constructive techniques (Section 5.5.3). An important question is the means to evaluate or compare explanations, which is the topic of Section 5.5.4. Considering that the objective of this document is not to cover the field extensively but to provide an overview of the main approaches and challenges, in each section we focus on some representative examples of existing techniques and highlight their main features. The reader can refer to one of the surveys published on this topic for a more comprehensive account of the state of the art.[161]

5.5.1. 'Black box' approaches to explainability

Solutions following the 'black box' approach do not make any assumption about the code or underlying model of the ADS, except for its existence and the possibility to observe its outputs. The first challenge in this context is the construction of explanations based on observations of the ADS.[162] In some cases, the observation itself is also a challenge because the ADS is integrated within a complex system involving multiple parties and which can only be partially observed or controlled. This is typically the case for ADS used in web services such as recommendation systems or personalised advertisement systems.

Local Interpretable Model-agnostic Explanations (LIME)[163] is an example of a state-of-the-art 'black box' explanation system. The underlying idea is that, even if an algorithm can be very complex and difficult to explain globally, it may be possible to provide local explanations that are both faithful and understandable (see figure 14). LIME is a generic framework in which an explanation is defined as a member *g* of a set of interpretable models *G*, endowed with a measure of complexity. A notion of faithfulness is also defined to measure the difference between an explanation *g* and the model *f* of the ADS around *x*. The system draws samples around the point of interest *x* to build a faithful explanation of the model in its vicinity.[164] LIME is defined in a very general way and can be instantiated with different types of interpretable models (set *G*), such as linear functions or decision trees. Explanations can also be presented to the user in different ways, for example as histograms (see figure 15), sets of words or images. As an illustration, a picture classified as a dog may appear with the head and the legs highlighted, meaning that they were the determining factors for the classification. In contrast, if the highlighted part of the picture were a ball, this might be a sign that the ADS has produced the right result for wrong reasons, which can be useful for the designer to improve the system.

---

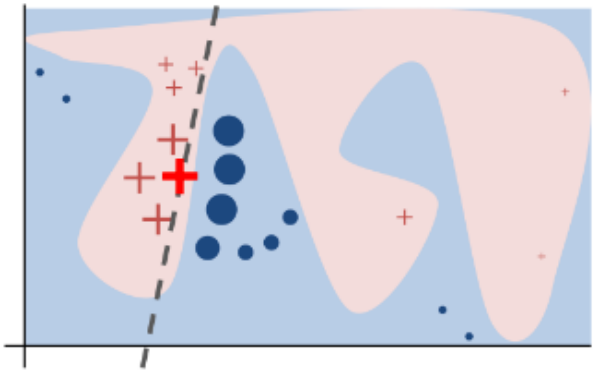
[161] Riccardo Guidotti, Anna Monreale, Franco Turini, Dino Pedreschi, Fosca Giannotti; A survey of methods for explaining black box models; 2018; https://arxiv.org/abs/1802.01933. Carmen Lacave, Francisco J. Diez; A review of explanation methods for Bayesian networks; The Knowledge Engineering Review; (17,2); 2002.

[162] This approach is also called 'reverse engineering' by some authors.

[163] Marco Tulio Ribeiro, Sameer Singh, Carlos Guestrin; 'Why should I trust you?' Explaining the predictions of any classifier; Knowledge Discovery and Data Mining Conference (KDD); ACM; 2016.

[164] Samples are weighted according to their proximity to the point of interest.

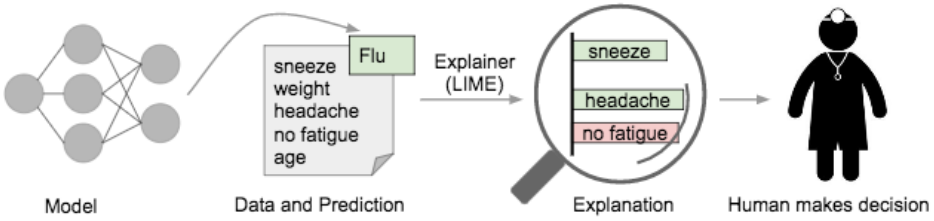
Figure 14 – Local explanation of a complex model by a linear model using LIME



Source: Marco Tulio Ribeiro, Sameer Singh, Carlos Guestrin, 'Why should I trust you?' Explaining the predictions of any classifier, Proceedings of the Knowledge Discovery and Data Mining Conference (KDD), ACM, 2016.

In figure 14, the dashed grey line approximates the frontier between the pink and the blue spaces around the highlighted red cross.

Figure 15 – Explanation in the form of histograms using LIME



Source: Marco Tulio Ribeiro, Sameer Singh, Carlos Guestrin, 'Why should I trust you?' Explaining the predictions of any classifier, Proceedings of the Knowledge Discovery and Data Mining Conference (KDD), ACM, 2016.

In figure 15, the two most influential factors for the 'flu' diagnosis are 'headache' and 'sneeze') while 'no fatigue' goes against this diagnosis.

An earlier 'black box' approach to explainability is the TREPAN algorithm, introduced as a technique to extract decision trees from neural networks.[165] In fact, TREPAN works as a decision tree learning algorithm using the ADS to be explained as an oracle. TREPAN uses the answers returned by the ADS to build the decision tree. These answers are the labels (classes) of the instances queried by TREPAN. TREPAN uses a 'best-first' expansion strategy to expand the tree to increase its fidelity to the ADS. In contrast to traditional decision tree learning algorithms, TREPAN can benefit from the fact that it is not limited to a fixed set of training data.

Other types of explanations can also be extracted following the 'black box' approach. For example, Mark W. Craven and Jude W. Shavlik describe techniques to generate 'if-then-else' rules from neural networks which are used as oracles as in TREPAN.[166] The conditions can be either conjunctive rules, as in '**if a and not b then c**' or 'M-of-N' rules, as in '**if 2 of {a,b,c} then d**'. The main challenge in these approaches is to reduce the complexity of the exploration since the search can be exponential in the number of input features. Marco Tulio Ribeiro and his co-authors have recently proposed an

[165] Mark Craven, Jude W. Shavlik; Extracting tree-structured representations of trained networks; *Conference on Advances in Neural Information Processing Systems; 1996.*

[166] Mark Craven, Jude W. Shavlik; Using sampling and queries to extract rules from trained neural networks; *International Conference on Machine Learning (ICML); 1994.*

optimised algorithm to efficiently compute 'if-then-else' rules. This algorithm is implemented in a model-agnostic explanation system called Anchor.[167]

The above solutions rely on the assumption that the ADS to be explained can be used as an oracle (it is possible to submit queries to the ADS and observe its answers). However, in certain situations access to the ADS is an issue because it is embedded within a larger system whose execution is affected by many parameters that cannot be observed. This is typically the case for ADS used in personalised advertisements or recommendation systems. Several tools have recently been proposed to address this problem, in particular to shed light on the practices of profiling and micro-targeting. For example, AdFischer is a tool to study online tracking through automated controlled experiments.[168] It basically makes it possible to simulate new users visiting web pages corresponding to particular interests and to analyse the advertising served to these users. AdFischer uses machine learning to detect differences of patterns in these advertisements. As a result, AdFischer can provide the features that have the strongest impact on the choice of advertising served to a user. Since many factors potentially influence the choice of the advertisements which are not under the control of the experimentation, it is necessary to conduct a large number of tests and to measure the statistical significance of the results.[169] AdFischer has been applied to the detection of discrimination and to highlight the use of certain topics of interest such as 'substance abuse' in the selection of advertisements.

The 'black box' approach to explainability has received a lot of attention from the research community because it is the only possible option when the code of the ADS is not available. Another advantage of the approach is its generality since it does not depend on the underlying technique or model of the ADS.

Sunlight is another tool that provides explanations about web targeting with statistical confidence.[170] The designers of Sunlight have placed a strong emphasis on three main principles:

- the generality of the approach (large-scale experiments with a wide variety of data),
- the robustness of the results (statistical justification), and
- interpretability (understandability by non-expert users).

Sunlight also generates fictitious profiles and analyses the advertisements served as outputs. In order to enhance interpretability, Sunlight focuses on simple explanations which take the form of disjunctions of features that have the strongest impact on the result. The disjunction is derived from the sparse linear model learned from the observations of the system (inputs and outputs). Sunlight has also been used to detect targeting based on sensitive topics such as health, religious affiliation or sexual orientation. One of the strengths of Sunlight is its scalability, in the sense that it is able to deal with multiple input features.

The 'black box' approach to explainability has received a lot of attention from the research community because it is the only possible option when the code is not disclosed by the operator or

[167] Marco Tulio Ribeiro, Sameer Singh, Carlos Guestrin; Anchors: high precision model-agnostic explanations; Thirty second AAAI Conference on Artificial Intelligence; 2018.

[168] Amit Datta, Michael Carl Tschantz, and Anupam Datta; Automated experiments on ad privacy settings; a tale of opacity, choice, and discrimination; Privacy Enhancing Technologies (PET); 2015.

[169] The measure is implemented in AdFischer by evaluating the p-value of a permutation test, which amounts to comparing the observed test statistics with the result obained from a random permutation test.

[170] Sunlight: fine-grained targeting detection at scale with statistical confidence; 22nd ACM SIGSAC Conference on Computer and Communications Security (CCS); ACM; 2015.

provider of the ADS. Another advantage of this approach is its generality, since the solutions do not depend on the underlying technique or model of the ADS.

5.5.2. 'White box' approaches to explainability

In contrast to the 'black box' approach, 'white box' explanation systems do rely on the analysis of the ADS code. In addition to the type of explanations that they can generate, 'white box' solutions differ in terms of the ADS they can handle (Bayesian networks, neural networks of limited depth, deep neural networks, etc.), their way to handle continuous data (e.g. through discretisation) and their complexity. An example of early work in this direction is the Elvira system for the graphical explanation of Bayesian networks.[171] Basically, the user of Elvira can change certain assumptions, such as 'fever' in the case of a medical ADS, and observe the impact of this assumption on the result, for example the probability of a disease. Elvira also offers qualitative information about pairs of variables for example by colouring a link between two variables if higher values of the first one lead to higher values of the second. Elvira follows a 'white box' approach in the sense that it makes it possible for the user to see and edit the model (Bayesian network).

The 'white box' approach has also been explored for ADS based on neural networks. For example, the tool proposed by Matthew D. Zeiler and Rob Fergus makes it possible to visualise the input stimuli of a network that excite individual feature maps at any layer in the model.[172] The challenge in this context is to be able to map the activities in intermediate layers back to the input pixel space so as to make the explanation understandable. This information is useful for the designers of an ADS to get insight into its internal operations to detect potential problems and improve the ADS. It is also possible to produce more widely accessible explanations for neural networks by showing the features that have the strongest influence in a decision. A 'white box' approach to reach this goal is called 'contribution propagation'[173] or 'relevance propagation'.[174] The idea consists in propagating the prediction score backwards in the network and redistributing scores to neurons at a lower level depending on their contributions to the neuron at the upper level. This backward propagation leads to interpretable patterns in the input domains that are associated with a given classification.

5.5.3. Constructive approaches to explainability

The constructive approach can be applied in the ideal situation where explainability requirements can be taken into account in the design phase of the ADS. Two options are possible to achieve explainability by design:

1. Relying on an algorithmic technique which, by design, meets the intelligibility requirements whilst providing sufficient accuracy.
2. Enhancing an accurate algorithm with explanation facilities so that it can generate, in addition to its nominal results (classification) a faithful and intelligible explanation for these results.

An example of the first approach is 'generalised additive models' (GAMs) and their extensions, the 'generalised additive models plus interactions' (GA^2M) proposed by Yin Lou, Rich Caruana and

---

171 Carmen Lacave, Roberto Atienza, Francisco J. Diez; Graphical explanation in Bayesian Networks; ISMDA Conference; Spinger; LNCS 1933; 2000.

172 Matthew D. Zeiler, Rob Fergus; Visualising and understanding convolutional networks; ECCV; Springer, LNCS 8689; 2014.

173 Will Landecker, Michael D. Thomure, Luis M. A. Bettencourt, Melanie Mitchell, Garrett T. Kenyon, Steven P. Brumby; Interpreting individual classifications of hierarchical networks; 2013 IEEE Symposium on Computational Intelligence and Data Mining (CIDM); 2013.

174 Grégoire Montavon, Wojciech Samek, Klaus-Robert Müller; Methods for interpreting and understanding deep neural networks; Digital Signal Processing; (73); 2018.

Johannes Gehrke.[175] GAMs are defined as linear combinations of shape functions on individual features. Shape functions can be arbitrarily complex, which makes it possible to get a high level of accuracy. In addition, because shape functions apply to individual features and are combined using a linear function, the results of a GAM remain interpretable: each shape function can be visualised through a two-dimensional plot and the contribution of each feature can be understood from the weights of the linear function. As a result, GAMs are both more accurate than linear models and more intelligible than state-of-the-art techniques such as deep neural networks or support vector machines. However, the fact that interactions between features are not possible makes GAMs less powerful than techniques such as random forests that do not have this restriction. To alleviate this limitation, Yin Lou, Rich Caruana, Johannes Gehrke and Giles Hooker[176] have proposed an extension called 'generalised additive models plus interactions' (GA^2M). In GA^2M, it is possible to express two-dimensional interactions between features. This makes the model more powerful while maintaining a high level of intelligibility. A two-dimensional interaction between features *x* and *y* can be represented by a heat map on a two-dimension *x*-*y* space. The main challenge for GA^2M is to limit the number of pairs of features to consider, using statistical relevance tests. The authors have shown the accuracy, intelligibility and scalability of GA^2M on real healthcare problems (pneumonia risk prediction and hospital readmission).[177]

An example of the second approach is the technique proposed by Tao Lei, Regina Barzilay and Tommi Jaakkola to produce explanations in the form of subsets of input texts justifying a prediction.[178] These subsets must meet two essential requirements:

- They must correspond to short and coherent pieces of text to ensure intelligibility.
- To ensure correctness (faithfulness of the explanation), the application of the ADS to the subset must lead to the same prediction as its application to the entire text.

As an illustration, an explanation for a five-star rating for 'colour' in the analysis of a beer review can be an excerpt of the text such as 'a very pleasant ruby red-amber colour'. The technique involves two main components, a rationale generator used to generate short sequences of words and an encoder used to minimise the discrepancy between the true prediction (for the whole text) and the explanation prediction (for the excerpt). Different algorithms can be used to implement the rationale generator and the encoder. The generation of explanations is an unsupervised learning process (without any explicit explanation annotations) based on these two components.

Another strategy, followed in particular by J Bien and R Tibshirani, consists in generating 'prototypes' of each class. Prototypes are representative samples (in the input domain) leading to a particular classification. They must satisfy a number of criteria. For example, their number must be limited and they must be varied to ensure a good coverage of the input data set: each piece of input data must have a prototype of its class and no prototypes of a different class in its neighbourhood. As an illustration, for a handwritten digit recognition system, the system generates a small set of images corresponding to each digit as prototypes. Each set must provide sufficient variety to represent the corresponding class well, which in this case corresponds to the different ways to write

[175] Yin Lou, Rich Caruana, Johannes Gehrke; Intelligible models for classification and regression; Knowledge Discovery and Data Mining Conference (KDD); ACM; 2012.

[176] Yin Lou, Rich Caruana, Johannes Gehrke, Giles Hooker; Accurate intelligible models with pairwise interactions, Proceedings of the Knowledge Discovery and Data Mining Conference (KDD); ACM; 2013.

[177] Rich Caruana, Yin Lou, Johannes Gehrke, Paul Koch, Marc Sturm, Noémie Elhadad; Intelligible models for healthcare: predicting pneumonia risk and hospital 30-day readmission; Knowledge Discovery and Data Mining Conference (KDD); ACM; 2015.

[178] Tao Lei, Regina Barzilay Tommi Jaakkola; Rationalizing neural predictions; Conference on Empirical Methods in Natural Language Processing (EMNLP); 2016.

a digit. This strategy can be implemented as an optimisation problem, which can be solved using standard optimisation techniques.

5.5.4. Qualities of explanations

The explanations generated by the above methods can take very different forms and a number of criteria can be used to evaluate them.

Intelligibility, understandability. Since the primary goal of an explanation is to enhance the understanding of the ADS or its results, the first yardstick is intelligibility. Even though intelligibility is highly dependent on the form of the explanation, it is often measured using size criteria. Examples include the size or depth of a decision tree, the number of rules or the length of a textual explanation. Other, complementary, metrics such as readability or availability have been proposed to assess intelligibility.[179] However, intelligibility is a complex and subjective notion that can only be assessed precisely through experimental means. For example, larger decision trees are sometimes easier to understand than smaller trees. The notion of monotonicity is also sometimes associated with intelligibility: basically, a monotonic decision function (such as a decreasing probability of purchase when the cost increases) is easier to grasp for a human than a non-monotonic function. In addition, it often corresponds to the intuition that humans may have about the expected results of an ADS.

Fidelity, accuracy. The second key requirement for explanation is accuracy in the sense of fidelity to the ADS. This objective is more difficult to meet for global explanations than for local explanations. Indeed, if an explanation were absolutely accurate while covering the whole model, it would reflect the complexity of the ADS and would probably not be intelligible. Fidelity is therefore a relative rather than an absolute requirement for explanations. It is generally integrated within explanation systems through a proximity measure. For example, LIME includes a notion of faithfulness to measure the distance between an explanation and the true model of the ADS in the vicinity of a given input value. Similarly, the encoder proposed by Tao Lei, Regina Barzilay and Tommi Jaakkola minimises the discrepancy between the true prediction and the explanation prediction.[180]

Precision, level of detail. Explanations can also differ in terms of the level of precision that they provide. For example, an explanation may be only a list of features used to get a result with or without their respective weights; it can highlight excerpts of a text in red or use different colours to provide more information about the impact of these excerpts on the results of the ADS, etc.

Completeness. Another relevant property for explanations is completeness. Indeed, if an explanation includes only some of the factors that have influenced a decision, it might be misleading (unless the rule used for choosing the features is made clear). As an illustration, a study conducted by Athanasios Andreou and his colleagues has shown that the explanations provided by Facebook about its personalised advertisement system are both incomplete and misleading, because the unique feature shown is the most prevalent[181] attribute, rather than the most interesting from the perspective of the users.[182]

---

[179] Dayana Spagnuelo, Cesare Bartolini, Gabriele Lenzini, Metrics for transparency; Data Privacy Management and Security Assurance (DPM); LNCS 9963; Springer; 2016.

[180] Tao Lei, Regina Barzilay and Tommi Jaakkola; Rationalizing neural predictions; Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP); 2016.

[181] The most common attribute in the community of users.

[182] Athanasios Andreou, Giridhari Venkatadri, Oana Goga, Krishna P. Gummadi, Patrick Loiseau, Alan Mislove; Investigating ad transparency mechanisms in social media: a case study of Facebook's explanations; Network and Distributed Systems Security Symposium (NDSS); 2018.

Consistency. Consistency is another quality of explanations that can be defined in different ways. Consistency may concern a single explanation, several explanations (e.g. different inconsistent explanations provided for the same type of results), or the content of an explanation in relation to common sense or the knowledge of the users (for example, the fact that patients suffering from asthma had lower risk of dying from pneumonia).[183] Consistency has an impact on intelligibility and trust in the ADS.

5.5.5. Evaluation of explainability

As discussed in Section 4.1, the quality of an explanation should be assessed in relation to its intended recipients, their level of expertise and their objectives. The specific context should therefore be taken into account to determine the significance of each of the above criteria in the assessment of an explanation system. In addition, we should emphasise that:

- Some of these criteria may be in tension. For example, higher levels of accuracy and level of precision may reduce intelligibility.
- The evaluation of most of these criteria is a difficult (and often partly subjective) task.

In order to make this task more systematic and rigorous, Finale Doshi-Velez and Been Kim propose a taxonomy for the evaluation of 'interpretability' (taken in the same sense as 'explainability' here). They distinguish three levels of evaluation:[184]

- Functionally-grounded evaluations that are based on formal definitions of interpretability (for example use of decision trees as explainable models). These are less expensive because they do not require human experiments but they must rely on assumptions (the formal definitions) which have already been validated.
- Human-grounded evaluations involve simple human experiments, for example to assess what kinds of explanations are better understood.
- Application-grounded evaluations involve field experiments with the actual (or future) users of a system (e.g. doctors). They provide the most precise assessments but they are also the most expensive.

5.6. Challenges

As shown in this chapter, designing ADS that are safe, secure, privacy-preserving, fair and explainable is still very challenging and deserves more effort and research. Even well-engineered computer systems can result in unexpected errors and unexplained outcomes for several reasons.

Safety. ADS can cause unintended and negative consequences in their environment. For example, this may happen when the training environment does not match the operational environment. While many of these failures can be addressed with ad hoc solutions, there is a strong need to define a unified approach. A minimum requirement should be to perform extensive testing before deployment and provide accountability.

Security. ADS are complex and subject to many different types of attacks. For example, an adversary can threaten the integrity and availability of an ADS by polluting its training dataset, attacking its

[183] Rich Caruana, Yin Lou, Johannes Gehrke, Paul Koch, Marc Sturm, Noémie Elhadad; Intelligible models for healthcare: predicting pneumonia risk and hospital 30-day readmission; Knowledge Discovery and Data Mining Conference (KDD); ACM; 2015.

[184] Finale Doshi-Velez, Been Kim; Towards a rigorous science of interpretable machine learning; https://arxiv.org/abs/1702.08608.

underlying algorithm or exploiting the generated model at run-time. As shown earlier, existing countermeasures are unsatisfactory and more effective solutions need to be developed.

Privacy. ADS are often trained on personal data. Several attacks have been devised to either extract information about the training data or to retrieve the ADS model. These attacks raise privacy concerns. Some solutions, using cryptography or distributed architectures, have been proposed but are still preliminary and often have a detrimental impact on the performance of the ADS. More research is required to propose privacy-preserving ADS that achieve acceptable performance and privacy trade-offs.

Fairness. As shown earlier, there are different definitions of fairness, and new definitions are regularly proposed. For example, most existing definitions are statistical and are defined with respect to group averages. Definitions of fairness that consider individuals instead of groups are also worth considering.[185] Finally, it has been shown that it is impossible to satisfy all notions of fairness and, at the same time, maximise accuracy and fairness.[186] It is therefore necessary to consider challenging trade-offs, and these trade-offs have to be discussed by stakeholders, not statisticians or computer scientists. For example, it is not up to computer scientists to decide between different definitions of fairness and its trade-off with accuracy. As stated by Richard Berk and his co-authors, 'these are matters of values and law, and ultimately, the political process. They are not matters of science'.[187]

Explainability. ADS are often complex systems that are difficult to understand. 'Hand-coded' ADS code can be audited, but the task is not always easy since they generally consist of complex modules made of a large number of code lines developed by groups of engineers. ADS that are based on machine learning are even more challenging to understand, and therefore to explain, since their models are generated automatically from training data. Data have many properties and features, and each of them can influence the generated models. Furthermore, as noted by J. Burrell,[188] 'while datasets may be extremely large but possible to comprehend and code may be written with clarity, the interplay between the two in the mechanisms of the algorithm is what yields the complexity and thus the opacity.'

[185] Aaron Roth; Between 'statistical' and 'individual' notions of fairness in Machine Learning; https://aaronsadventures.blogspot.fr/2017/11/between-statistical-and-individual.html; 2017.

[186] Richard Berk, Hoda Heidari, Shahin Jabbari, Michael Kearns, Aaron Roth; Fairness in Criminal Justice Risk Assessments: The State of the Art; Sociological Methods & Research; 2018.

[187] Richard Berk, Hoda Heidari, Shahin Jabbari, Michael Kearns, Aaron Roth; Fairness in Criminal Justice Risk Assessments: The State of the Art; Sociological Methods & Research; 2018.

[188] Burell J.; How the machine 'thinks': Understanding opacity in machine learning algorithms; Big Data & Society; 2016.

6. Legal instruments

While the main focus of this document is the technical dimension of ADS, in this chapter we briefly discuss legal instruments that can be used to meet to objectives set forth in Chapter 4. It is a matter of fact that the technical solutions described are necessary, but cannot by themselves solve all the issues raised by ADS. They must be associated with other types of measures, in particular legal requirements in terms of transparency, explainability or accountability. In fact, various existing laws already apply to ADS and can, to a greater or lesser extent, address some of the requirements identified in Chapter 4. To cite but a few:

- Laws and European Directives against discrimination such as Directive 2006/54/EC on Equal Opportunities and Equal Treatment of Women and Men in Employment and Occupation, the Racial Equality Directive 2000/43/EC or the Employment Equality Directive (2000/78/EC).

- General consumer protection laws: for example, Directive 2006/123/EC on Services in the Internal Market states that 'Member States shall ensure that the general conditions of access to a service, which are made available to the public at large by the provider, do not contain discriminatory provisions relating to the nationality or place of residence of the recipient, but without precluding the possibility of providing for differences in the conditions of access where those differences are directly justified by objective criteria.'

- Sectoral laws such as regulations of the healthcare and banking sectors. For example, in the United States, the Equal Credit Opportunity Act (ECOA), which is enforced by the Federal Trade Commission (FTC), provides a right to be informed about the reasons of rejection of an application. Another example is the transparency requirements for high-speed trading algorithms.

- Regulations related to open access to administrative documents such as the Freedom of Information Act in the USA or the United Kingdom.

Needless to say, this list is far from exhaustive, but the point is that existing laws do not address all the issues identified in Chapter 4. In addition, certain laws may also constitute an obstacle to transparency or accountability, in particular laws protecting intellectual property and trade secrets. We discuss this issue in Section 7.2. Adapting or strengthening existing laws is often seen as a necessity to take the risks posed by the development of new technologies into account. To address this need and to try to meet the desiderata set forth in Chapter 4, a number of legal safeguards have been proposed or adopted during the last decade. An exhaustive review of these regulations and proposals is beyond the scope of this document. In this chapter, we illustrate the most significant approaches, considering two complementary dimensions:

- Substance: what are the actual rights or obligations introduced by new regulations or proposed by academics? Which desiderata (among those listed in Chapter 4) are targeted by new regulations or proposals? What is their intended scope (e.g. sectoral or general)?

- Means of enforcement: is the regulation (or would the proposal be) legally binding or not? Is it or would it be implemented through dedicated bodies (or supervisory authorities) or does it (or would it) fall within regular jurisdictions?

We first discuss the situation in Europe with the new General Data Protection Regulation (Section 6.1). We then focus on an example of recent developments in an EU Member State, France (Section 6.2), before sketching proposals that originate in the United States (Section 6.3).

6.1. European level: General Data Protection Regulation

One of the most widely discussed and commented regulations passed during recent years is the European General Data Protection Regulation (GDPR). Although it is too early to assess its practical impact, which is highly dependent on its interpretation and implementation by data protection authorities and courts, the GDPR has the potential to enhance personal data protection in Europe. In particular, it introduces:

- new rights for individuals (such as the right to portability, stricter rules for information and consent, enhanced erasure rights, etc.),
- new obligations for data controllers (data protection impact assessments, data protection by design and default, data breach notifications, etc.),
- new action levers such as collective actions and higher sanctions,
- better coordination mechanisms between supervisory authorities and a new body, the European Data Protection Board (EDPB), which replaces former Article 29 Working Party and which has extensive powers and binding decisions in particular for dispute resolution between national supervisory authorities.

The interested reader can find more details in Paul de Hert and Vagelis Papakonstantinou's analysis of the GDPR.[189] The effectiveness of the GDPR in terms of transparency or explainability is a topic of intense debate. Some authors claim that the GDPR introduces a 'right to explanation' for ADS,[190] while others argue that this right does not exist in the GDPR.[191] The issue of whether the GDPR provides a true right to explanation can be separated into two questions:

The effectiveness of the GDPR in terms of transparency or explainability is a topic of intense debate. Some authors claim that the GDPR introduces a 'right to explanation' for ADS, while others argue that this right does not exist in the GDPR.

1. Is such a right really set forth in the GDPR and, if so, what does 'explanation' mean exactly in this context?
2. If this right is set forth in the GDPR, what are the conditions for its application and is it likely to be effective?

At the core of the debate is Article 22 (figure 16), which defines the rules for 'automated individual decision-making, including profiling'. In addition, Articles 13 and 14 provide that:

> 'the controller shall, at the time when personal data are obtained, provide the data subject with the following further information necessary to ensure fair and transparent processing: […] (f) the existence of automated decision-making, including profiling, referred to in Article 22(1) and (4) and, at least in those cases, meaningful information about the logic involved, as

[189] Paul de Hert, Vagelis Papakonstantinou; The new General Data Protection Regulation: still a sound system for the protection of individuals?; The Computer Law & Security Review; (32,2); 2016.

[190] Bryce Goodman, Seth Flaxman; European Union regulations on algorithmic decision-making and a "right to explanation"; AI Magazine, (38,3); 2017. Andrew D. Selbst, Julia Powles; Meaningful Information and the Right to Explanation; International Data Privacy Law; (7,4); 2017. Gianclaudio Malgieri, Giovanni Comandé; Why a right to legibility of automated decision-making exists in the General Data Protection Regulation; International Data Privacy Law; (7,4); 2017.

[191] Sandra Wachter, Brent Mittelstadt, Luciano Floridi; Why a Right to Explanation of Automated Decision-Making Does Not Exist in the General Data Protection Regulation; International Data Privacy Law; (7,2); 2017.

> well as the significance and the envisaged consequences of such processing for the data subject.'

Sandra Wachter and her co-authors 'doubt both the legal existence and feasibility' of a right to explanation in the GDPR. As far as the legal existence is concerned, they argue that the word 'explanation' occurs only once, in Recital 71, which concerns decisions about a subject 'based solely on automated processing and which produces legal effects concerning him or her or similarly significantly affects him or her'. Recital 71 provides that:

> 'In any case, such processing should be subject to suitable safeguards, which should include specific information to the data subject and the right to obtain human intervention, to express his or her point of view, to obtain an explanation of the decision reached after such assessment and to challenge the decision.'

However, recitals are not legally binding: they only provide guidance for the interpretation of the articles. Therefore, the fact that the word 'explanation' occurs neither in Article 22 nor in Articles 13 or 14 is significant and means, according to Sandra Wachter and her co-authors that 'a right to explanation is thus not currently legally mandated by the requirements set in Article 22(3)'. They do not exclude that future jurisprudence can still interpret it as introducing a right to explanation, but this is 'only one possible future'. In addition, they argue that Articles 13 and 14 concern ex-ante explanations (notifications before a decision is made) but not ex-post explanations about specific decisions.

Figure 16 – GDPR Article 22: Automated individual decision-making, including profiling.

Article 22

1. The data subject shall have the right not to be subject to a decision based solely on automated processing, including profiling, which produces legal effects concerning him or her or similarly significantly affects him or her.

2. Paragraph 1 shall not apply if the decision:

(a) is necessary for entering into, or performance of, a contract between the data subject and a data controller; or

(b) is authorised by Union or Member State law to which the controller is subject and which also lays down suitable measures to safeguard the data subject's rights and freedoms and legitimate interests; or

(c) is based on the data subject's explicit consent.

3. In the cases referred to in points (a) and (c) of paragraph 2, the data controller 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.

4. Decisions referred to in paragraph 2 shall not be based on special categories of personal data referred to in Article 9(1), unless point (a) or (g) of Article 9(2) applies and suitable measures to safeguard the data subject's rights and freedoms and legitimate interests are in place.

Other authors take a less restrictive interpretation and argue that a right to explanation does exist in the GDPR. For example, Andrew D. Selbst and Julia Powles state that:

> 'Whether one uses the phrase 'right to explanation' or not, more attention must be paid to the GDPR's express requirements and how they relate to its background goals, and more thought must be given to determining what the legislative text actually means.'

Their legal analysis of the GDPR leads to the conclusion that:

> 'The GDPR clearly mandates 'meaningful information about the logic' of decisions to which Article 22 applies. If 'meaningful' is to have any substance that appears on its face to be a move in the direction of explanation of some type – and all parties in this debate, including Wachter and others, seem to agree on that point.'

It may also be noted that the right to 'contest the decision' in Article 22(3) would not be meaningful without any right to explanation.

Even if we can agree with the conclusion that a right to explanation exists in the GDPR, the actual effectiveness of this right remains to be seen since it applies only in case of 'decision based solely on automated processing'. Wachter and her co-authors state that:

> 'this creates a loophole whereby even nominal involvement of a human in the decision-making process allows for an otherwise automated mechanism to avoid invoking elements of the right of access (both in the Directive and GDPR) addressing automated decisions.'

This view is supported by the jurisprudence in countries, such as Germany, where this provision, which already existed in European Directive 95/46/EC, has been tested in court.[192] Another potentially strong restriction is that Article 22 applies only to decisions producing legal effects concerning the subject or 'similarly significantly affecting him or her'. Other authors, such as Gianclaudio Malgieri and Giovanni Comandé, encourage a more optimistic interpretation of Article 22, considering that

Even if we can agree with the conclusion that a right to explanation exists in the GDPR, the actual effectiveness of this right remains to be seen since it applies only in case of 'decision based solely on automated processing'.

> 'the threshold for minimum human intervention required so that the decision-making is 'solely' automated can also include nominal human intervention; the envisaged 'significant effects' on individuals can encompass as well marketing manipulation, price discrimination, etc.'

This opinion is in line with the 'Guidelines on automated individual decision-making and profiling for the purposes of Regulation 2016/679' published by the Article 29 Working Party in October 2017.[193] Regarding the restriction to 'decisions based solely on automated processing', the Article 29 Working Party states that:

> 'To qualify as human intervention, the controller must ensure that any oversight of the decision is meaningful, rather than just a token gesture. It should be carried out by someone who has the authority and competence to change the decision. As part of the analysis, they should consider all the available input and output data.'

The Article 29 Working Party also advocates a wide interpretation of 'similarly significantly affecting him or her':

> 'Recital 71 provides the following typical examples: 'automatic refusal of an online credit application' or 'e-recruiting practices without any human intervention'. These suggest that it is difficult to be precise about what would be considered sufficiently significant to meet the threshold. For example, based on the recital each of the following credit decisions fall under Article 22, but with very different degrees of impact on the individuals concerned: (1) renting

[192] Bettina Berendt, Sören Preibusch; Toward accountable discrimination-aware data mining: The importance of keeping the human in the loop – and under the looking-glass; Big Data; (5,2); 2017.

[193] Guidelines on automated individual decision-making and profiling for the purposes of Regulation 2016/679; Article 29 Data Protection Working Party; WP 251; 2017.

> a city bike during a vacation abroad for two hours; (2) purchasing a kitchen appliance or a television set on credit; (3) obtaining a mortgage to buy a first home.'

As a temporary conclusion on the GDPR, we can agree with Andrew D. Selbst and Julia Powles on the fact that 'these issues go to the applicability of the right rather than the shape of the right and will be a matter for future interpretation by legislators, data protection authorities, and courts.'

6.2. France: Law for a Digital Republic

Some countries have also adopted new laws to enhance transparency or explainability of ADS during recent years. For example, the Law for a Digital Republic[194] passed in October 2016, in France, introduces new obligations for two types of users of ADS: administrations and 'digital platform operators'. A digital platform is defined as an online service based on ranking, referencing contents, goods or services or connecting several parties with a view to selling or exchanging contents, goods or services.

In contrast to the GDPR, this law does not restrict the obligations for administrations to 'decisions based solely on automated processing'. It refers instead to decisions taken on the basis of algorithmic processing. Administrations must inform the persons affected by such decisions. In addition, they must, upon request, communicate the rules of the algorithm and the main features of its implementation to individuals. The application decree provides further details. It requires in particular that the information be intelligible and include the criteria used and their weight in the decision for the affected person. These requirements pertain to both local and global explanations. However, these obligations should not adversely affect secrets protected by law.

The requirements of the Law for a Digital Republic are different and less constraining for platform operators. Platform operators must provide clear, fair and transparent information about:

- the general conditions of use of the services, including the modalities for referencing, dereferencing and ranking, and
- the existence of a contractual or corporate relation or compensation that can have an impact on the referencing or ranking.

The application decree adds that the information must include the criteria and main parameters used by the algorithm and, close to each result, the fact that it has been influenced by a contractual or corporate relation or compensation.

6.3. United States

Most proposals from the legal doctrine in the United States emphasise due process and accountability as the most effective way to introduce a form of control over ADS. For example, Danielle Keats Citron and Frank Pasquale[195] state that:

> 'One of the great accomplishments of the legal order was holding the sovereign accountable for decision-making and giving subjects basic rights, in breakthroughs stretching from Runnymede to the Glorious Revolution of 1688 to the American Revolution. New algorithmic decision-makers are sovereign over important aspects of individual lives. If law and due process are absent from this field, we are essentially paving the way to a new feudal order of unaccountable reputational intermediaries.'

---

[194] Loi n° 2016-1321 du 7 octobre 2016 pour une République numérique; JORF n°0235; 2016.

[195] Danielle Keats Citron, Frank Pasquale; The scored society: due process for automated predictions; Washington University Law Review; (89); 2014.

Citron and Pasquale propose two complementary strategies to achieve this goal:

- Full access given to federal regulators (such as the FTC for credit-scoring systems) to assess ADS, in particular to check that they do not lead to unfair or discriminatory decisions. The regulator should be able to audit the ADS on a regular basis and these audits should include sufficient tests of the systems to detect biases or other sources of unfairness. The audit should lead to a 'Privacy and Civil Liberties Impact Assessment' and 'identify appropriate risk mitigation measures'.
- At the individual level, audit trails should be available to make it possible for people affected by ADS to understand the decisions. Such audit trails should record 'the correlations and inferences made algorithmically in the prediction process'. If the protection of the intellectual property of the ADS precludes public access, audit trails can be accessed through trusted neutral experts. Another suggestion made by Citron and Pasquale is to make it possible for people to test different assumptions to understand their impact on a decision.

As far back as 2008, Danielle Keats Citron advocated 'technological due process' as a 'framework of mechanisms capable of enhancing the transparency, accountability, and accuracy of rules embedded in automated decision-making systems'.[196] Focusing on administrative and constitutional law, Citron argues that 'automation jeopardises the procedural protections that have long been deemed foundational to the administrative state'. The proposed framework relies on a systematic approach (based on the distinction between two forms of laws: rules and standards) to find an acceptable balance between automation and human discretion. It also includes a set of procedural measures (in line with the two above strategies) to 'prevent procedurally defective rule-making and arbitrary government decision making'. Danielle Keats Citron also stressed the need to release the source code of ADS to the public and suggested that federal funding for technology purchase should be conditioned on the use of open code. Kate Crawford and Jason Schultz build on this work to develop further requirements for due process to redress privacy harms.[197]

Other scholars suggest additional legislative changes to improve accountability. For example, Deven Desai and Joshua Kroll[198] start from the observation that discriminatory or unfair treatments are sometimes difficult to detect, either for technical reasons, or because companies believe that they can easily evade their obligations and deny any intent to breach the law. Based on this observation, they make several proposals, including changes in trade secrecy law to protect whistleblowers from employers.

Scholars often use other legal fields as sources of inspiration for the regulation of ADS. For example, Deven Desai and Joshua Kroll[199] refer to the Sarbanes-Oxley Act:

> 'As stated during the passage of the Sarbanes-Oxley Act of 2002 ('SOX'), '[w]ith an unprecedented portion of the American public investing in [publicly-traded] companies and depending upon their honesty, ... [the lack of whistleblower protection for private-sector whistleblowers did] not serve the public good.' Similarly, with an unprecedented portion of decision-making with due process and vital verification interests at stake being processed through software, protection for employees who blow the whistle on software companies who knowingly violate the law is vital. In other words, the government needs private actors to aid in law enforcement, and there is a long history of private citizens aiding in law

196 Danielle Keats Citron; Technological due process; Washington University Law Review; (85, 6); 2014.

197 Kate Crawford, Jason Schultz; Big Data and Due Process: Toward a Framework to Redress Predictive Privacy Harms; Boston College Law Review; (93); 2014.

198 Deven R. Desai, Joshua A. Kroll; Trust But Verify: A Guide to Algorithms and the Law; Harvard Journal of Law and Technology; (31,1); 2018.

199 Ibid.

> enforcement by providing support to public prosecution and through private enforcement and private evidence gathering.'

Andrew Tutt uses another source of inspiration, the Food and Drug Administration (FDA), to advocate the creation of a dedicated agency to supervise the development, deployment and use of algorithms:[200]

> 'The products the FDA regulates, and particularly the complex pharmaceutical drugs it vets for safety and efficacy, are similar to black-box algorithms. And the crises the FDA has confronted throughout its more than one hundred years in existence are comparable to the kinds of crises one can easily imagine occurring because of dangerous algorithms. The FDA has faced steep resistance at every stage, but its capacity to respond to, and prevent, major health crises has resulted in the agency becoming a fixture of the American institutional landscape. We could draw on the FDA's history for lessons, and use those lessons as an opportunity to avoid repeating that history.'

Academics often use other legal fields as sources of inspiration for the regulation of ADS. Examples include the Sarbanes-Oxley Act (SOX) for whistleblower protection and the Food and Drug Administration (FDA) as a reference with regard to regulatory agencies.

Tutt argues that, even if the development of ADS is still at an early stage and some stakeholders may worry that regulation could stifle innovation, a regulatory agency is necessary precisely because the use of ADS is growing very quickly and raises significant risks. We come back to the issue of supervisory authorities in the next chapter.

[200] Andrew Tutt; An FDA for algorithms; Administrative Law Review; (83); 2017.

7. Open questions and remaining challenges

The instruments presented in Chapter 5 and 6 are undoubtedly useful but far from sufficient to address all the challenges raised by ADS. Complementary measures are necessary to make technical and legal instruments effective in terms of fairness, explainability and accountability. Furthermore, ADS raise substantive issues that are not yet fully understood and which must be analysed and thoroughly debated. In this chapter, we sketch out the main existing challenges from three different (and complementary) perspectives:

- **Ethical and political:** what should be accepted or not? Under what principles?
- **Legal and social:** what rights and obligations should be enshrined in law? What should be the role of the different stakeholders in the implementation of the rules?
- **Technical:** what guarantees can technical instruments provide? How can one reconcile potentially conflicting objectives such as accuracy and explainability of ADS?

7.1. Ethical and political debate

As illustrated in Chapter 3, ADS raise far reaching issues in many areas such as justice, policing, healthcare, democratic life, etc. ADS exacerbate or force us to rethink existing problems such as discrimination, but they also introduce new ethical questions that are very difficult to address. Examples of critical and complex questions raised by ADS include:

- Is it actually legitimate to use an ADS in the first place? In certain contexts, such as evidence-based sentencing or lethal weapons, this use has been criticised. It is however far from straightforward to establish firm boundaries between acceptable uses of ADS and situations in which they should be banned. The question can also be raised about personalisation. For example, is it acceptable to deny a loan based on the fact that friends of the requester in a social network are deemed credit-unworthy? Is it acceptable to personalise prices based on the location of a consumer in a given country or his assumed capacity to pay a high price? Is it acceptable to make access to certain services conditional upon a 'trust score' or 'social score' derived from the behaviour of the requester? Is it acceptable to grade teachers and decide to renew their contract or not based on their ranking?
- Beyond existing criteria already identified in anti-discrimination laws, what types of treatment should be considered undesirable? Existing laws focus on the protection of well-identified social groups (e.g. based on gender, ethnic origin, religion, etc.) but ADS make it possible to discriminate according to many other criteria that could also be considered as unfair in certain situations. Such unfairness could target virtual groups that are not necessarily identified in society (such as, to take a few random examples, left-handed people, people who have learned the same foreign language, or like the same movies). Where should the line be drawn and under what principles?
- How should online manipulation be characterised and distinguished from (acceptable) influence or 'nudging'? When manipulation is based on the exploitation of human biases how can it be identified and fought when it reproduces existing commercial practice in the digital world?
- In which cases should transparency, explainability or other forms of accountability be required and under what principles? How can this requirement be defined and assessed (e.g. how can explainability be measured)? Should certain types of ADS be forbidden in certain situations when an acceptable level of transparency, explainability or accountability cannot be achieved (for example in court, or to support medical diagnosis)?

- What are the choices to be made in the design of autonomous cars when critical decisions have to be taken (question of life and death)? Should ethical behaviour be encoded in the system and, if so, what should they be and who should be able to decide upon this choice of 'ethical behaviour'? Should otherwise autonomous cars behave like human drivers, which would probably mean in a selfish way?

Some proposals have been made to try to address these issues in a principled way. For example, Brent Daniel Mittelstadt and his co-authors[201] have proposed a conceptual map based on six types of concerns:

1. **Inconclusive evidence**, which addresses the uncertainty surrounding ADS results and their use (e.g. confusing correlation and causation). Inconclusive evidence can typically lead to unjustified actions.
2. **Inscrutable evidence**, which corresponds to the lack of knowledge about the data used or lack of explanation about the link between the conclusions and the inputs.
3. **Misguided evidence**, which corresponds to the fact that input data can be erroneous or biased.
4. **Unfair outcomes**, which includes discriminatory decisions.
5. **Transformative effects**, which can be seen as the side effects of the use of ADS, including their impact on our vision of the world and the social and political organisation of society.
6. **Traceability**, which includes the difficulty 'to identify who should be held responsible for the harm caused'.

The first three concerns are epistemic in the sense that they 'address the quality of evidence produced by an algorithm' while the fourth and fifth are normative in that they focus on the actions (decision based on an ADS). The last one, traceability, has to do with responsibilities. Daniel Mittelstadt and his co-authors[202] use this map to structure the academic discussion on ethics of algorithms.

Another useful conceptual framework to understand the different variants of transparency requirements is the taxonomy proposed by Tal Zarsky.[203] This two-dimensional grid (pictured in figure 17) can be used to analyse the benefits and potential drawbacks of different forms of transparency.

[201] Brent Daniel Mittelstadt, Patrick Allo, Mariarosaria Taddeo, Sandra Wachter, Luciano Floridi; The ethics of algorithms : mapping the debate; Big Data & Society; 2016.

[202] Ibid.

[203] Tal Z. Zarsky; Transparent predictions; University of Illinois Law Review; (1503); 2003.

Figure 17 – Transparency framework proposed by Tal Zarsky

Segments of information flow

1. The collection of data and aggregation of data sets
2. The data analysis phase
3. The usage stage (how the results of the analysis are utilised)

Recipients of transparency information:

1. The general public
2. Internal and external institutions (e.g. committees, independent watchdog groups or experts)
3. Affected individuals

Source: Tal Z. Zarsky, Transparent predictions. University of Illinois Law Review, 1503, 2003).

Another example of the systematic analysis of ethical issues that can be useful in this context is the EDPS Ethics Advisory Group Report,[204] which proposes a list of 'foundational values to digital ethics': dignity, freedom, autonomy, solidarity, equality, democracy, justice and trust. The EDPS report identifies five types of conditions for the development of digital technologies in line with these values: material conditions; cultural conditions; personal conditions; political and socio-structural conditions; and legal conditions.

NGOs also have a strong role to play in this ethical debate. For example, the Electronic Frontier Foundation (EFF) identifies five questions to address in analysing the risks posed by ADS:[205]

- Will this algorithm influence – or serve as a basis for – decisions with the potential to negatively impact people's lives?
- Can the available data actually lead to a good outcome?
- Is the algorithm fair?
- How will the algorithm (really) be used by humans?
- Will people affected by these decisions have any influence over the system?

As stressed by the EFF, 'these questions are just starting points, and they won't guarantee equitable results, but they are questions that all organisations should be asking themselves before implementing a decision-making system that relies on an algorithm.'[206]

7.2. Legal and social perspective

The ethical and political debates suggested in the previous section are prerequisites for further action. Assuming that wide agreement has been reached on some of the issues discussed above,

[204] EDPS Ethics Advisory Group Report; Towards a digital ethics; https://edps.europa.eu/sites/edp/files/publication/18-01-25_eag_report_en.pdf; 2018.

[205] Jamie Williams, Lena Gunn; Math can't solve everything: questions we need to be asking before deciding an algorithm is the answer; Electronic Frontier Foundation; https://www.eff.org/fr/deeplinks/2018/05/math-cant-solve-everything-questions-we-need-be-asking-deciding-algorithm-answer; 2018.

[206] Ibid.

the next step is to decide what the most appropriate instruments are to implement it. As far as law is concerned, the first question to be asked is whether it is necessary (or even desirable) to create new legal instruments. If so, several options can be considered, each of them having their pros and cons:

- Should these instruments pertain to state regulation, self-regulation or co-regulation?[207]
- Should they take the form of hard law or soft law (codes of conduct, guidelines, recommendations)?
- Should they be general or sectoral?

Different options are also possible in terms of enforcement. For example, ADS regulations may be the competence of

- Ordinary jurisdictions;
- Existing regulatory agencies (e.g. the FTC in the USA or Data Protection Authorities in Europe);
- Regulatory agencies dedicated to ADS (such as the 'FDA for algorithms' suggested by Andrew Tutt).[208]

The answers to these questions may be very different for each of the issues discussed in the previous section. For example, there is a pressing need to legislate, or at least to clarify the interpretation of existing laws, about liability rules applicable to autonomous cars and robots in social environments. A lot of progress has been made in this domain on the technical side but providing clear liability rules seems to be a prerequisite for larger scale deployment. More generally, liability is a key issue in all uses of ADS, especially in physical systems.

ADS that are not entirely automated may raise even more complex issues because the frontier between automated decisions and decisions taken on the basis of ADS can be blurred. For example, would a practitioner be entirely responsible for a fatal decision taken on the basis of an ADS that is widely used and generally recognised as trustworthy in the medical sector?

A complementary question is the development of certification schemes for ADS. Certifications and labels, if properly implemented, can be a way to enhance trust in ADS and to verify that they comply with certain rules (such as the absence of bias or discrimination). Certifications can either be made on a voluntary basis (as encouraged by the GDPR) or be a requirement (as for the deployment of medical devices). We return to this issue in Chapter 8.

Another critical issue on the legal side is the potential use of intellectual property rights to set limits on transparency or accountability. Legally speaking, an ADS can be protected in three main ways: through copyright, trade secret or patents. Patents are published and concern the right to manufacture or use an invention: they are thus not an obstacle to transparency or accountability. Copyright applies only to the code of the software itself and does not prevent explanations. The main hindrance may therefore be the protection of trade secrets.

Actually, the protection of trade secrets has already been used by industry as an argument against transparency.[209] The first interesting question in this respect is whether the reverse engineering

---

[207] Linda Senden; Soft law, self-regulation and co-regulation in European law: where do they meet?; Electronic Journal of Comparative Law; (9.1); 2005.

[208] Andrew Tutt; An FDA for algorithms; Administrative Law Review; (83); 2017.

[209] See for example the written evidence submitted by Google to the members of parliament in the UK: http://data.parliament.uk/writtenevidence/committeeevidence.svc/evidencedocument/science-and-technology-committee/algorithms-in-decisionmaking/written/71681.html. Another case is the protection of formulas used for

techniques presented in Chapter 5 could be considered as breaches of trade secrets. It seems that most jurisdictions would not disallow reverse engineering in this context but any uncertainty should be removed to this respect. For example, Article 3 of the European Directive 2016/943 on the protection of undisclosed know-how and business information (trade secrets) against their unlawful acquisition,[210] use and disclosure states that:

> 'The acquisition of a trade secret shall be considered lawful when the trade secret is obtained by any of the following means:
>
> (a) independent discovery or creation;
>
> (b) observation, study, disassembly or testing of a product or object that has been made available to the public …'

However, Article 4(3)c also states that the acquisition of a trade secret shall be considered unlawful if 'in breach of contractual or any other duty to limit the use of a trade secret'. Therefore, the developer or operator of an ADS can prohibit reverse engineering by contractual means (i.e. in its terms of use). In the context of US law, the AINow institute also calls for clarification on this matter:[211]

> 'In order to conduct the research necessary for examining, measuring, and evaluating the impact of AI systems on public and private institutional decision-making, especially in terms of key social concerns such as fairness and bias, researchers must be clearly allowed to test systems across numerous domains and via numerous methodologies. However, certain US laws, such as the Computer Fraud and Abuse Act (CFAA) and the Digital Millennium Copyright Act (DMCA), threaten to limit or prohibit this research by outlawing 'unauthorized' interactions with computing systems, even publicly accessible ones on the internet. These laws should be clarified or amended to explicitly allow for interactions that promote such critical research.'

Another question is whether an ADS provider can, in certain circumstances, be required to disclose the code of the system or information about its logic, even if the provider argues that this information is a trade secret and its disclosure would undermine the competitiveness of the company. As discussed in the recitals of the Directive 2016/943, the main argument for trade secrets is to encourage innovation, which is also a worthy cause. Therefore, the answer to this question depends on the situation and the values at stake. For example, it would not be sensible to require a spam filter or chess game provider to publish the code of his ADS. On the other hand, it does not seem acceptable that ADS used for sentencing in courts or to execute medical diagnoses could not be audited because they are protected by trade secrets. This question has been raised in the famous *Loomis* v. *Wisconsin* case. The Wisconsin Supreme Court considered that Loomis, the defendant's, right to due process was not violated despite the fact that the provider of the COMPAS system used to calculate his risk score refused to disclose information about it (even about the

> Trade secret protection has already been used by industry as an argument against transparency. However, several lawyers have argued strongly against the trade secret privilege in the case of ADS used for sentencing in court.

credit scoring in Germany: according to Sandra Wachter and her co-authors (Sandra Wachter, Brent Mittelstadt, Luciano Floridi; Why a Right to Explanation of Automated Decision-Making Does Not Exist in the General Data Protection Regulation; International Data Privacy Law; (7,2); 2017), the judgements of the German Federal Court show that data subjects are not allowed to get information about the logic of the ADS beyond the features taken into account by the system (excluding their weights).

[210] https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=CELEX:32016L0943&from=EN.

[211] The AI Now Report; The Social and Economic Implications of Artificial Intelligence Technologies in the Near-Term; 2016; https://ainowinstitute.org/AI_Now_2016_Report.pdf.

weights of the input variables). The decision has been confirmed by the United States Supreme Court. However, several lawyers have argued strongly against the trade secret privilege in this context. For example, according to Rebecca Wexler:[212]

> 'trade secrets should not be privileged in criminal proceedings. A criminal trade secret privilege is ahistorical; harmful to defendants; and unnecessary to protect the interests of the secret holder. Meanwhile, compared to substantive trade secret law, the privilege overprotects intellectual property. Further, privileging trade secrets in criminal proceedings fails to serve the theoretical purpose of either trade secret law or privilege law.'

As far as innovation is concerned, Rebecca Wexler also points out that 'despite its name, one stated objective of trade secret law is to facilitate controlled information sharing'. Therefore:

> 'The fact that trade secret law aims, at least in part, to facilitate information sharing for purposes of negotiation, employment, and regulation suggests that the law should also perform this function in criminal proceedings. Revealing trade secrets under duties of confidentiality in business or regulatory contexts is arguably analogous to revealing them under a protective order in a criminal proceeding.'

In addition, one can argue that the controlled disclosure of the code or the logic of an ADS should not be conflated with its public disclosure. In particular, the risks in terms of loss of competitiveness seems a much weaker argument in the case of controlled disclosure, especially if it has to be balanced with fundamental rights such as the right to due process.

7.3. Technical perspective

Technical instruments can play an essential role to meet the desiderata identified in Chapter 4, but they are still in their infancy with many challenges that need to be addressed. These challenges can be classified into two main categories:

- **Conceptual**
  - How to define complex and subjective notions such as discrimination, unfairness, privacy or manipulation. As seen in Chapter 5, these concepts are complex, difficult to formalise and require more work to agree on common definitions.
  - When several definitions of a notion exist, what are their respective strengths and weaknesses? As seen in Section 5.4, several incompatible definitions of fairness have been proposed. For example, a definition can require that the rate of positive classification be equal across the groups (**disparate impact** or **statistical parity**), or that the false positive and false negative rates be equal across the groups (**equalised odds**).
  - What are the best types of explanations depending on the different recipients, their level of expertise and objectives?
- **Operational**:
  - How can the tensions between accuracy, cost and explainability/fairness/privacy be reconciled? Is it feasible or are they inherent limitations? If the latter case is confirmed, what would be the best trade-off for a particular ADS?
  - What are the best technical approaches and mechanisms to provide explainability, fairness and privacy in ADS? As seen previously, different approaches can be taken to

[212] Rebecca Wexler; Life, liberty, and trade secrets : intellectual property in the criminal justice system; (70); Standford Law Review; (1343); 2018; https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2920883.

implement each of these properties. For example, fairness can be achieved using a pre-processing, in-processing or post-processing approach (see Section 5.4.3). What is the best approach for each specific ADS? As far as explainability is concerned, how should the interactions with an explanation system be designed to improve user understanding?[213]

– How should explainability by design be implemented? The same question holds for fairness by design and privacy by design. These properties should be taken into consideration from the beginning of the conception of an ADS, as required by the GDPR for data protection. However, this phase requires strong technical expertise that cannot be expected from all ADS developers. How can guidance and help be provided to designers and developers to implement these principles?

Addressing technical challenges is further complicated by the fact that researchers do not have access to the huge data sets held by private companies, nor do they have access to the algorithms themselves. This imbalance is a significant impediment to the development of knowledge in the field.

– How should explainability, fairness and privacy of ADS be assessed? What metrics and tools should be used? Should a specific risk management methodology be developed?

– How can ADS best be secured against accidents and adversarial attacks? As seen in the report, it is very difficult to protect ADS against unexpected failures and attacks. How should their resistance to these failures and attacks be tested? What is an acceptable level of security and robustness?

Addressing these challenges is further complicated by the fact that researchers do not have access to the huge data sets held by private companies, nor do they have access to the algorithms themselves. This imbalance is a significant impediment to the development of knowledge in the field.

[213] Prashan Madumal, Tim Miller, Frank Vetere, Liz Sonenberg; Towards a grounded dialog model for explainable artificial intelligence; 1st International workshop on socio-cognitive systems; IJCAI; 2018.

8. Policy options

Even though ADS are still in an early stage of development (see Chapters 2 and 3), they are already being used in many different situations and will soon become pervasive across all professional and personal activities. However, as discussed in Chapter 7, many critical questions remain to be solved in this regard and many others are bound to arise in the future. A whole host of reports and studies have been published to inform policy-makers and the public about the precautions and measures to need to be taken to address these issues.[214] Based on these studies and the analysis presented in this report, as a conclusion we put forward a number of options, listed below. These options are mostly organisational or procedural (in the general sense of the term), rather than substantive, since positions on this matter should rather result from public debate than be issued by expert groups. We do however provide guidance on the criteria and issues that should be carefully considered before the adoption of ADS.

We distinguish five complementary types of actions:

1. Development and dissemination of knowledge about ADS,
2. Public debate about the benefits and risks of ADS,
3. Adapting legislation to enhance the accountability of ADS,
4. Development of tools to enhance the accountability of ADS,
5. Effective validation and monitoring measures for ADS.

8.1. Development and dissemination of knowledge about ADS

As discussed in this report, ADS raise complex questions that are not entirely understood by experts, not to mention users or affected people. In addition, it is a very fast-changing area, both from a technical perspective and in terms of usage. The first step to enhance their accountability is therefore to improve and disseminate knowledge about ADS. In particular this means:

- **Developing multidisciplinary and interdisciplinary research in ADS**. Philosophers, experts in ethics, AI, computer science, social science and law should work together to further develop conceptual tools to analyse the ethical issues raised by ADS. More research is also needed to design methods and tools to enhance the security, safety, privacy, fairness and explainability of ADS. Computer scientists, AI experts, psychologists and knowledge engineers should join forces to understand the types of explanations that are the most useful depending on the targeted audiences and their needs. Further progress has also to be made on the characterisation of notions like explainability and fairness and their implementation 'by design'. As shown in Chapter 5 and Chapter 7, implementing privacy-aware, transparent, secure and fair ADS is a very challenging task that deserves further research work. Research should also be performed in order to better understand the risks, to monitor, evaluate and mitigate them. Rigorous algorithm impact assessment (AIA) methodologies should be defined following an interdisciplinary and multi-stakeholder approach.

[214] EDPS Ethics Advisory Group Report; Towards a digital ethics; https://edps.europa.eu/sites/edp/files/publication/18-01-25_eag_report_en.pdf; 2018. Commission Nationale de l'Informatique et des Libertés (CNIL); How can humans keep the upper hand? The ethical matters raised by algorithms and artificial intelligence?; 2018; https://www.cnil.fr/sites/default/files/atoms/files/cnil_rapport_ai_gb_web.pdf. Center for Internet and Human Rights; The ethics of algorithms: from radical content to self-driving cars; GCCS; 2015. Mike Annany; Towards an ethics of algorithms: convening, observation, probability, and timeliness; Science, Technology, and Human Values; 2015.

- A key condition to facilitating research is the possibility for the research community to obtain access, under specific conditions and strict confidentiality, to the datasets held not only by public bodies but also by private companies. This right to access is justified by the fact that such large amounts of data may be considered as 'data of public interest'. As stated by the European Statistical System (ESS) in a recent positioning paper:

'the issue of access to data of public interest cannot be left unanswered now as the risk of fragmented approaches across the EU is increasing, making it even more difficult to address it in the future. A more practical avenue would consist at this stage in affirming in EU law a general principle of access to privately-held data which are of public interest and addressing in broad terms the main elements for such access to be effective at operational level'[215].

The ESS positioning paper focuses on the use of data by statistical offices but the arguments are valid for research in general. This position is in line with a report published by the French Conseil Générale de l'économie (CGE) and Internet Governance Forum (IGF) about 'data of general interest',[216] and with a recent paper by Hetan Shah who goes further, stating that:

> 'intellectual-property rights expire after a fixed time period: what if, similarly, technology companies were allowed to use the data that they gather only for a limited period, say, five years? The data could then revert to a national charitable corporation that could provide access to certified researchers, who would both be held to account and be subject to scrutiny that ensure the data are used for the common good.'[217]

For the same reason, it should be made clear that reverse engineering for the purpose of analysing, explaining or detecting biases in ADS should be considered lawful and should not be limited by trade secret or more generally by intellectual property right laws.

- Ensuring that the issues raised by ADS are properly understood by their designers and developers. Engineers should be trained and supervised, in order to consider essential requirements such as fairness or explainability from the beginning of the design phase and throughout the ADS development cycle. Guidelines describing good development practices should be devised and published. Tools should be provided to help developers implement and test the desired ADS properties. The development of a body of experts in ADS, with the ability to cover both technical and ethical aspects, should also be encouraged. These experts could be integrated into development teams or serve in ADS evaluation bodies.[218]

- Enhancing the level of awareness of users of ADS, be they professionals or individuals, and citizens in general (who can all be affected by the use of ADS). Because ADS are used to make decisions about people, it is of prime importance that all everyone involved have a minimum of knowledge about the underlying processes, their potential and the limitations of the technologies. As stated by Tal Zarksy:

'computerized automations generate an (erroneous) aura of flawless decision-making abilities. This is indeed a serious concern. I doubt, however, whether additional transparency

---

215 European Statistical System; Position paper on access to privately held data which are of public interest; 2017.

216 C. Duchesne, L. Cytermann, L. Vachey, M. Morel, T. Aureau; Rapport relatif aux données d'intérêt général; Conseil Général de l'Economie, Inspection Générale des Finances; 2015.

217 Hetan Shah; Use our personal data for the common good; Nature; (556,7); 2018.

218 This option is in line with the report published recently by a French parliamentary mission led by Cédric Villani: Donner un sens à l'intelligence artificielle. Pour une stratégie nationale et européenne; 2018. https://www.aiforhumanity.fr/pdfs/9782111457089_Rapport_Villani_accessible.pdf.

> provides a sufficient answer. This concern could be resolved through other measures such as educating the public and relevant decision makers of the true nature of automation.'[219]

More generally, digital literacy is essential for citizens to be able to exercise their rights in the digital society.

8.2. Public debate about the benefits and risks of ADS

Enhancing the level of understanding of the technologies involved in ADS is necessary, but not sufficient since many issues raised by ADS are subjective and may be approached in different ways depending on individual perceptions and political views. Considering that ADS can have a major impact on society, they must be subject to public debate. Several conditions have to be met to ensure the quality of this debate:

- It must involve all stakeholders, opinions and interests, including at least experts of all disciplines, policy-makers, professionals, NGOs and the general public.
- It must be conducted in a rigorous way without overshadowing any of the key issues, including the preliminary question of the legitimacy of the use of ADS.[220] Indeed, voices have been raised against the use of ADS in certain contexts. For example, Kelly Hannah-Moffat states that:

  > 'The use of risk tools in sentencing is especially problematic because when used in courts they may offend moral and legal norms as well as country specific constitutional values. […] The trend towards using risk instruments in all sectors of the criminal justice system, therefore, merits further theoretical deliberation and empirical study.'[221]

- In the same vein, Chelsea Barabas and her colleagues argue:

  > 'for a shift away from predictive technologies, towards diagnostic methods that will help us to understand the criminogenic effects of the criminal justice system itself, as well as evaluate the effectiveness of interventions designed to interrupt cycles of crime. In contrast to the current emphasis on machine learning techniques that offer no grounded way of understanding the underlying drivers of crime, these methods should be based in a more rigorous approach that incorporates both qualitative and quantitative data analysis.'[222]

8.3. Adapting legislation to enhance the accountability of ADS

As discussed in Section 7.2, different types of legal instruments can be used to enhance the accountability of ADS. Considering that the technology and its uses are evolving very quickly in this area, it is wise to avoid hasty legislation that could end up creating more problems than those that it attempts to solve. New regulations should be enacted only when the matter has been properly understood, the public debate suggested above has taken place, and it is established that existing laws are insufficient to address the issues. It may be the case that certain sectors require further

---

219 Tal Z. Zarsky; Transparent predictions; University of Illinois Law Review; 2013.

220 Jamie Williams, Lena Gunn; Math can't solve everything: questions we need to be asking before deciding an algorithm is the answer; Electronic Frontier Foundation; 2018. https://www.eff.org/fr/deeplinks/2018/05/math-cant-solve-everything-questions-we-need-be-asking-deciding-algorithm-answer. See also: https://blog.google/topics/ai/ai-principles .

221 Kelly Hannah-Moffat; Actuarial sentencing: an 'unsettled' proposition; Justice Quarterly; (30,2); 2013.

222 Chelsea Barabas, Karthik Dinakar, Joichi Ito, Madars Virza, Jonathan Zittrain; Interventions over predictions: reframing the ethical debate for actuarial risk assessment; Processing of Machine Learning Research; (81); 2018.

regulation or clarifications on the application of existing laws. It can be argued, for example, that liabilities in the area of autonomous cars or robots in social environments should be better defined, or transparency requirements should be imposed for the use of ADS by judges or the medical sector. Some issues, such as the potential ban of lethal weapons, should ideally be regulated at the international level.

As far as enforcement is concerned, we believe that a clear distinction should be made between:

- Ethical committees, with the mission to stimulate discussion, to conduct debate and publish recommendations, and
- Operational bodies, such as accreditation bodies, certification agencies and oversight agencies, which together provide a framework for the monitoring, certification and oversight of specific ADS. Oversight agencies should also have the power to sanction operators of non-compliant ADS (like Data Protection Authorities for non-compliance with the GDPR).

Ethical committees can operate at a general (cross-sector) level while operational bodies should be sectoral because different application areas raise different issues and have different histories, cultures, sets of practices and regulations. To take just one example, the medical sector has a well-established tradition of certification of medical devices and the certification of ADS to support diagnosis should fall within this framework. Operational bodies could still rely on the expertise of the body of experts suggested in Section 8.1 because ADS used in different sectors may involve similar techniques and require similar expertise.

8.4. Development of methodologies and tools to enhance ADS accountability

Tools and methodologies must be developed in order to:

- Help designers and developers build ADS that match the desired properties described in Chapter 4;
- Help third parties to test, validate and possibly certify ADS;
- Help users to interact in a meaningful way with ADS.

Most ADS designers and developers are not experts in privacy, security, fairness or explainability. It is therefore important to provide tools and methodologies to help them reconcile the tensions that exist between accuracy, cost and explainability/fairness/privacy. Recommendation guides are unfortunately not sufficient to do so. Tools and methodologies that consider the whole development cycles of ADS should be developed and disseminated.

Similarly, frameworks, composed of metrics, methodologies and tools that assess the impact of an ADS and test the desired properties of ADS should be developed. These frameworks could be used by designers to test their ADS, and by third-party entities, such as certification authorities, to validate them.

As far as users are concerned, better explanation facilities are required, in particular, more interactive interfaces and dialog models. As stated by Prashan Madumal and his co-authors,[223] 'lack of a general

[223] Prashan Madumal, Tim Miller, Frank Vetere, Liz Sonenberg; Towards a grounded dialog model for explainable artificial intelligence; 1st International workshop on socio-cognitive systems; IJCAI; 2018.

dialog model of explanation that takes into account the end user can be attributed as one of the shortcomings of existing explainable AI systems'.

Developing such tools and frameworks is far from trivial. It requires a large amount of research and study, as discussed in Chapter 7.

8.5. Effective validation and monitoring measures

The GDPR introduces an obligation for data controllers to conduct Data Protection Impact Assessments[224] and encourages 'the establishment of certification mechanisms and data protection seals and marks'. Considering that the stakes are very high regarding ADS, there is no reason why they should not be subject to the same types of precaution. We recommend in particular that:

- ADS should not be deployed without a prior Algorithmic Impact Assessment (AIA) unless it is clear that they have no significant impact on the life of individuals.[225]
- The certification of ADS should be encouraged and even mandatory in certain sectors.

Dillon Reisman and his colleagues have already advocated AIA as a 'practical framework for public agency accountability' in a recent AINow Institute report.[226] Beyond a 'self-assessment of existing and proposed automated decision systems, evaluating potential impacts on fairness, justice, bias, or other concerns across affected communities', they emphasise the need for 'researcher review processes before the system has been acquired'. They also recommend that agencies provide notice to the public about ADS, 'solicit public comments to clarify concerns and answer outstanding questions' and 'provide due process mechanisms for affected individuals or communities to challenge inadequate assessments or unfair, biased, or otherwise harmful system uses'. Even if the AINow report focuses on public agencies, most of its recommendations should also apply to sensitive ADS deployed in the private sector. As outlined by Reisman and his colleagues, there are two major differences between their AIA framework and the DPIA requirements of the GDPR: DPIA 'are not shared with the public and, and have no built-in external researcher review or other individualised due process mechanisms.' We agree that AIA should be more ambitious on these matters because the lack of external review and publicity is a major weakness of the GDPR regarding DPIA.

Conducting an AIA is not simple. Models and tools should be proposed to make it easier, as done for DPIA.[227] Even though many assessment criteria are bound to remain subjective, an AIA framework should provide an evaluation methodology which is as systematic and rigorous as possible. The definition of such an AIA framework is outside of the scope of this document; it should be proposed by the ethical committees or oversight agencies suggested above. We highlight below only some key issues which should be considered in such an AIA:

1. **Legitimacy:** the first question to be addressed is the legitimacy of the use of the ADS. Legitimacy can be addressed at three levels:
    a. The legitimacy of the purpose of the ADS: for example, what role should risk prediction instruments play in criminal sentencing? Is it legitimate to take sanctions based on crimes that have yet not been committed? Is it legitimate or desirable to

[224] Only when a type of processing 'is likely to result in a high risk to the rights and freedoms of natural persons'.

[225] For example, it does not seem necessary to impose an AIA for an ADS included in a consumer electronic chess game.

[226] Dillon Reisman, Jason Schultz, Kate Crawford, Meredith Whittaker; Algorithmic impact assessments: a practical framework for public agency accountability; AINow Institute; https://ainowinstitute.org/aiareport2018.pdf; 2018.

[227] For example, the CNIL, which is the French DPA, has made available a tool to help data controllers conduct DPIA: https://www.cnil.fr/en/may-2018-updates-pia-tool.

grade teachers and use their ranking to decide whether their contract should be renewed or not? Is it legitimate to use the social network connections of a person to take a decision about whether they should be given a loan? More generally, whose interests does the ADS serve?

b. The legitimacy of the underlying technique: for example the use of machine learning techniques is a topic of heated debate because they establish correlation rather than causation relations. As an illustration, Chelsea Barabas and her colleagues argue that 'machine learning should not be used for prediction, but rather to surface covariates that are fed into a causal model for understanding the social, structural and psychological drivers of crime.'[228] D. James Greiner also argues that 'as it has been used in civil rights litigation, regression suffers from several shortcomings: it facilitates biased, result oriented thinking by expert witnesses; it encourages judges and litigators to believe that all questions are equally answerable; and it gives the wrong answer in situations in which such might be avoided. These difficulties, and several others, all stem from the fact that regression does not begin with a paradigm for defining causal effects and for drawing causal inferences.'[229]

c. The legitimacy of the criteria used by the model: for example, if risk prediction instruments are used in criminal sentencing, is it legitimate to use demographic or socioeconomic variables?[230] Is it legitimate to use geographical parameters in personalised pricing? Is it legitimate for insurance companies to use genetic data?

2. **Qualities:** If the system has passed the legitimacy test, a second type of question exists that relates to the intrinsic qualities of the model itself. As discussed in Chapter 5, the expected properties include fairness, privacy, reliability, security, accuracy, etc. Each property may be more or less critical depending on the purpose of the ADS. In any case, the relevant properties should be systematically taken into consideration and rigorously assessed. Great care must be paid in particular to the justification of the choices made when several properties are in tension (such as accuracy and fairness) and when different definitions are available for a given objective (such as fairness or privacy).

3. **Integration within the human environment:** The third main issue is the integration of the ADS within its human environment, including its users, the affected persons, external experts and oversight agencies. This is where issues like transparency or explainability come into play. We believe that transparency and explainability, as they are defined in Chapter 4, should be the rule by default, and should be as broad as possible. When restrictions are applied, they should be justified and the burden of proof should lie with the operator of the ADS. These justifications can be based, for example, on the need to protect intellectual property rights or industrial secrets. In some cases, it can also be argued that the publication of the details of the ADS or its logic could defeat its purpose because it would make it easier to manipulate. This is the case for ADS used for fraud detection or selecting tax payers who will be subject to manual audits for example. In any case, such arguments can be used to justify minimising the information disclosed to the general public but not to evade independent reviews or certifications by accredited bodies. As stated in the AINow report, ADS operators should also provide ways for the people affected to be able to challenge

---

228 Chelsea Barabas, Karthik Dinakar, Joichi Ito, Madars Virza, Jonathan Zittrain; Interventions over predictions: reframing the ethical debate for actuarial risk assessment; Processing of Machine Learning Research; (81); 2018.

229 D. James Greiner; Causal inference in civil rights litigation; Harvard Law Review; (122,2); 2008.

230 Sonja B. Starr; Evidence-based sentencing and the scientific rationalization of discrimination; Stanford Law Review; (66); 2014.

decisions taken on the basis of an ADS. Considering that the behaviour of certain systems continuously evolves, it should also be possible for oversight agencies to audit them on a regular basis.

All types of risks should be considered in an AIA, including individual and collective risks, allocative risks (such as denying a loan) and representational risks (such as labelling images of black people as 'gorillas').[231] It should be clear however that AIA should not only focus on the risks of **using** an ADS: they should also assess the risks of **not using** an ADS. In other words, AIA should consider both the benefits and risks. For example, several studies have been conducted about the use of ADS in the area of justice, some of them focusing on the risks of discrimination,[232] others on the benefits in improving judges' decisions.[233] In addition, the benefit risk balance applies to both the primary functionalities of the ADS and to its transparency and explainability features. For example, transparency and explainability will generally make the ADS more effective because users who do not understand the results of a system may not use them properly. However, if the ADS is not robust or uses rough proxies (such as the number of ongoing lease contracts for an ADS used to make a loan decision), it can be manipulated by the people affected (for example by closing or merging lease contracts). The risks and benefits identified in Chapter 3 can be used to check that all relevant issues have been considered. The conceptual framework proposed by Tal Zarsky can also serve as inspiration for this benefit/risk analysis.[234]

A complementary question is the development of certification schemes for ADS. Certifications and labels, if properly implemented, can be a way to enhance trust in ADS and to verify that they comply with certain rules (such as the absence of bias or discrimination). The implementation of a certification scheme must be carefully thought out to ensure that it can really be trustworthy (which requires serious audits by independent third parties, ideally accredited by a national body), while remaining acceptable from an economic standpoint. We believe that certification requirements and obligations should be sectoral. Indeed, the needs and the risks vary greatly from one type of application to another and sectoral supervisory authorities or agencies are in a better position to define reference evaluation criteria and to control their application. ADS certification can be either on a voluntary basis (as encouraged by the GDPR) or mandatory in certain areas such as justice and healthcare.

Even if the target is not an official certificate or label, it is important to test, validate or have ADS audited by external reviewers, before deployment. These tests should check that the desired requirements are met or lead to recommendations to improve the system. An initiative worth mentioning in this respect is a new company called ORCAA (O'Neil Risk Consulting and Algorithmic Auditing). ORCAA reviews algorithms using an 'ethical matrix' including criteria such as 'accuracy, consistency, bias, transparency, fairness and timeliness'.[235]

8.6. Conclusion

This study presents a number of desired properties or objectives for ADS and different technical and legal instruments to achieve or facilitate them. It also discusses the limitations of the proposed

---

231 Dillon Reisman, Jason Schultz, Kate Crawford, Meredith Whittaker; Algorithmic impact assessments: a practical framework for public agency accountability; AINow Institute; https://ainowinstitute.org/aiareport2018.pdf; 2018.

232 Sonja B. Starr; Evidence-based sentencing and the scientific rationalization of discrimination; Stanford Law Review; (66); 2014.

233 Jon Kleinberg, Himabindu Lakkaraju, Jure Leskovec, Jens Ludwig, Sendhil Mullainathan; Human decisions and machine predictions; National Bureau of Economic Research; NBEC Working Paper 23180; http://www.nber.org/papers/w23180.

234 Tal Z. Zarsky; Transparent predictions; University of Illinois Law Review; (1503); 2003.

235 https://www.wired.com/story/want-to-prove-your-business-is-fair-audit-your-algorithm/.

instruments and puts forward some options. At this stage, we stress that the options are only general suggestions resulting from our analysis, which builds on previous reports and studies. They could serve as a starting point for a multi-stakeholder discussion, as suggested in Section 8.2. We conclude with the desiderata for algorithms and put them into perspective.

We define transparency as the availability of the ADS code with its design documentation, parameters and learning dataset when the ADS relies on machine learning. However, as argued by Kroll et. al. (and many other authors), 'the source code of computer systems is illegible to non-experts'.[236] Transparency should therefore not be seen as the ultimate solution for users or people affected by the decisions of an ADS. Its main benefit is rather:

- For independent experts, evaluation bodies or DPAs for example, to audit ADS and certify them. To achieve this goal, the ADS code does not necessarily need to be made public. To avoid intellectual property issues, it can be made available to evaluators bound by a confidentiality agreement.
- For public scrutiny by the community (in the spirit of open source communities) to detect potential bugs or unacceptable features and possibly suggest improvements. This option is especially relevant for ADS used by administrations.[237]

Kroll et al. also argue that 'even experts often struggle to understand what software code will do: inspecting source code is a very limited way of predicting how a computer program will behave'. Hence the need for explanations, for users, for affected people, and also for designers and developers themselves. As shown in Section 4.1, 'explainability' has different meanings and the needs for explainability vary considerably according to the audience:

- Designers and developers may be interested in all types of explanations, including operational explanations (how the system actually works), logical explanations (the logical relationships between inputs and results) or causal explanations (the causes for the results).
- Users, especially professional users such as medical doctors or judges, should be able to understand the general logic of the ADS (global and logical explanations), i.e. the decisional criteria and their respective weights as well as the reasons for specific results (local and causal explanations).
- Affected people will probably be more interested in the reasons for the decisions that affect them (local and causal explanations) and how they can influence them (counterfactual explanations).[238]

Generally speaking, explainability, just like transparency, should not be seen as an end in itself but as a means to an end. This end can be, for example, to be able to improve the fairness of an ADS or to challenge a decision. It is also important to note that the requirements for explainability vary from one ADS to another, according to the potential impact of the decisions made. For example, Finale Doshi-Velez and Been Kim, argue that sometimes explanations[239] are not even necessary 'either because (1) there are no significant consequences for unacceptable results or (2) the problem is sufficiently well-studied and validated in real applications that we trust the system's decision, even if the system is not perfect'. The second condition applies, for example, to automated metro systems

[236] J.A Kroll et al.; Accountable Algorithms; Univ. Penn. Law Review; 2017.

[237] Except for ADS used for fraud detection whose code may have to remain confidential.

[238] Sandra Wachter, Brent Mittelstadt, Chris Russel; Counterfactual explanations without opening the black box, automated decisions and the GDPR**;** Harvard Journal of Law & Technology; 2018.

[239] What is intended here is probably explanations to the general public. One could argue that explanations can still be useful to experts (to validate the system) and could also possibly be useful to the collectivity (in the spirit of open source communities) in order to detect potential bugs and suggest improvements.

which have been validated and certified by independent experts. However in other situations such as ADS used by medical doctors or judges, explainability should be an absolute requirement. In fact, one could argue that ADS that do not make automated decisions produce results that have to be used, and therefore interpreted, by human beings to make decisions. Human decision-makers must therefore have sufficient understanding of these results and their limitations to be able to use them in an appropriate way. Finale Doshi-Velez and Been Kim also argue that 'the need for interpretability stems from an incompleteness in the problem formalisation, creating a fundamental barrier to optimisation and evaluation'.

However, as discussed in Chapter 7, significant progress is yet to be made to provide and assess explainability tools that can really be useful to non-experts. As argued by Tim Miller and his co-authors:[240]

> 'While the re-emergence of explainable AI is positive, this paper argues most of us as AI researchers are building explanatory agents for ourselves, rather than for the intended users. But explainable AI is more likely to succeed if researchers and practitioners understand, adopt, implement, and improve models from the vast and valuable bodies of research in philosophy, psychology, and cognitive science; and if evaluation of these models is focused more on people than on technology.'

In addition, even if a legal obligation of explainability is desirable for most ADS, this obligation should not be a way for ADS providers or operators to evade their responsibilities. As described by Edwards and Veale,[241] there is a risk that a right to explanation puts the onus on users or affected people to challenge decisions that are wrong. First, even if individuals do have a right to ask for an explanation, this right may not be easy to exercise. If the ADS is endowed with explanation facilities, it may still require that the user have a minimal amount of familiarity with the technology. If they have to follow an administrative procedure, the process may be lengthy and demanding, requiring a certain level of motivation and persistence. As noted by Edwards and Veale: 'a legal right to an explanation may be a good place to start, but it is by no means the end of the story. Rights become dangerous things if they are unreasonably hard to exercise or ineffective in results, because they give the illusion that something has been done while in fact things are no better'.

Second, even if an explanation is obtained, it may not be sufficient to be able to understand the decision up to a point that it can be challenged. In addition, bias or discrimination may only be detected by studying the whole corpus of users. Something that is difficult to do through individual challenges.

Although transparency and explainability of ADS should be required in most cases, we argue that, as far as the protection of individuals is concerned, accountability is the most important requirement. In fact, transparency and explainability may allow for the discovery of deficiencies, but they do not guarantee the reliability, security and fairness of an ADS. Accountability can be achieved via different means such as algorithm impact assessments (AIA), auditing and certification. The main virtue of accountability is to put the onus on the providers or operators of the ADS to demonstrate that they meet the expected requirements. Of course, accountability cannot provide

'A legal right to an explanation may be a good place to start, but it is by no means the end of the story. Rights become dangerous things if they are unreasonably hard to exercise or ineffective in results, because they give the illusion that something has been done while in fact things are no better.'

[240] T. Miller, P. Howe, L. Sonenberg; Explainable AI: beware of inmates running the asylum; IJCAI Workshop on Explainable Artificial Intelligence (XAI); 2017.

[241] L. Edwards, M. Veale; Eslaving the Algorithm: From a "Right to an Explanation" to a "Right to Better Decisions"; IEEE Security and Privacy; 2018.

absolute guarantees either, but if certification is rigorous and audits are conducted on a regular basis, potential issues can be identified and corrective measures taken. In addition, if sanctions are significant enough, an accountability approach provides strong incentives for ADS providers to comply with the requirements. From this perspective, oversight agencies and supervisory authorities should play a central role and it is critically important that they have all the means necessary to carry out their duties. These means are not only in terms of funding and expertise. They should also have the power to access and analyse the details of the ADS, including their source code and training data.

Last, but not least, we believe that, if appropriate accountability measures are taken, ADS also have the potential to improve transparency and reduce unfairness and discrimination. Another benefit of using them, and one that can already be observed, is the fact that they put decisions at the front and centre of public debate. Decisions that, up to now, have been taken far from the sight of citizens.[242]

[242] As an illustration, the controversy about the COMPAS algorithm has triggered a debate about the types of evidence that can be used for sentencing in the USA. In another area, the ADS used to match students and universities in France (APB, for '*Affectation Post Bac*', replaced by '*Parcoursup*' in 2018), have given rise to many discussions about the rules that should be used to allocate students to universities and the level of automation that should be considered as acceptable in this context.

9. Bibliography

Abadi M., Chu A., Goodfellow I., McMahan H.B., Mironov I., Talwar K., and Zhang L.; Deep learning with differential privacy; ACM Computer and Communication Security (CCS); 2016.

Abraham K., Rabin R.; Automated vehicles and manufacturer responsibility for accidents: a new legal regime for a new era; Virginia Law Review; Forthcoming 2019.

Amodei D., Olah C., Steinhardt J., Christiano P., Schulman J., Mane D.; Concrete Problems in AI Safety; ArXiv:1606.06565 [cs]; 2016.

Annany M.; Towards an ethics of algorithms: convening, observation, probability, and timeliness; Science, Technology, and Human Values (41); 2015.

Ananny M., Crawford K.; Seeing without knowing: Limitations of the transparency ideal and its application to algorithmic accountability; New Media and Society (20); 2018.

Andreou A., Venkatadri G., Goga O., Gummadi K., Loiseau P., Mislove A.; Investigating ad transparency mechanisms in social media: a case study of Facebook's explanations; Proceedings of the Network and Distributed Systems Security (NDSS) Symposium; 2018.

Ateniese G., Mancini L. V. , Spognardi A., Villani A., Vitali D., Felici G.; Hacking smart machines with smarter ones: How to extract meaningful data from machine learning classifiers; International Journal of Security and Networks (10); 2015.

Barabas C., Dinakar K., Ito J., Virza M., Zittrain J; Interventions over predictions: reframing the ethical debate for actuarial risk assessment; Processing of Machine Learning Research (81); 2018.

Barreno M., Nelson B., Sears R., Joseph A.D., Tygar J.D.; Can machine learning be secure?; ACM Symposium on Information, computer and communications security (ASIACCS '06); ACM, New York, NY, USA; 2016.

Berendt B., Preibusch S.; Toward accountable discrimination-aware data mining: The importance of keeping the human in the loop – and under the looking-glass; Big Data (5); 2017.

Berk R., Heidari H., Jabbari S., Kearns M., Roth A.; Fairness in Criminal Justice Risk Assessments: The State of the Art; Sociological Methods & Research; 2018.

Biggio B., Nelson B., Laskov P., Support vector machines under adversarial label noise; Journal of Machine Learning Research; 2011.

Binns R.; Algorithmic accountability and public reason; Philosophy & Technology; 2017.

Birnbaum B.; Credit scoring and insurance: costing consumers billions and perpetuating the racial divide; National Consumer Law Center; 2007.

Bradshaw J., de G. Matthews A., Ghahramani Z.; Adversarial examples, uncertainty, and transfer testing robustness in gaussian process hybrid deep networks; arXiv preprint arXiv:1707.02476; 201.

Bonnefon J.F., Shariff A., Rahwan I.; The social dilemma of autonomous vehicles; Science (352); 2016.

Bolukbasi T., Chang K.W., Zou J., Saligrama V., Kalai A.; Man is to computer programmer as woman is to homemaker? debiasing word embeddings; Proceedings of the 30th International Conference on Neural Information Processing Systems (NIPS); 2016.

Bostrom N.; Superintelligence: Paths, dangers, strategies; Oxford University Press; 2014.

Bovens M.; Analysing and assessing accountability: A conceptual framework; European Law Journal (13); 2007.

Burell J.; How the machine 'thinks': Understanding opacity in machine learning algorithms; Big Data & Society, 2016.

Calders, T., Verwer, S.; Three naive Bayes approaches for discrimination-free classification; Data Mining and Knowledge Discovery (21); 2010.

Carlini N., Wagner D.; Adversarial Examples Are Not Easily Detected: Bypassing Ten Detection Methods; ACM Workshop on Artificial Intelligence and Security; 2017.

Caruana R., Lou Y., Gehrke J., Koch P., Sturm M., Elhadad N.; Intelligible models for healthcare: predicting pneumonia risk and hospital 30-day readmission; Proceedings of the ACM Knowledge Discovery and Data Mining Conference (KDD); 2015.

Chen L., Mislove A., Wilson C.; Peeking beneath the hood of Uber; ACM Internet Measurement Conference (IMC);2015.

Christin A., Rosenblat A., Boyd D.; Courts and predictive algorithms; Workshop on Data & Civil Rights: A new era of policing and justice; 2015.

Chouldocheva A.; Fair prediction with disparate impact: a study of bias in recidivism prediction instruments; Big Data, Special issue on Social and Technical Trade-Offs; 2017.

Citron D. K., Pasquale F. A.; The Scored Society: Due Process for Automated Predictions; Washington Law Review (89); 2014.

Citron D. K.; Technological due process; Washington University Law Review (85); 2014.

Craven M., Shavlik J. W.; Extracting tree-structured representations of trained networks; Conference on Advances in Neural Information Processing Systems; 1996.

Craven M., Shavlik J. W.; Using sampling and queries to extract rules from trained neural networks; International Conference on Machine Learning (ICML); 1994.

Crawford K., Schultz J.; Big Data and Due Process: Toward a Framework to Redress Predictive Privacy Harms; Boston College Law Review; 2014.

Danzinger S., Levav J., Avnaim-Pesso L.; Extraneous factors in judicial decisions ; Proceedings of National Academic of Sciences (PNAS); 2011.

Datta A., Tschantz M., Datta A.; Automated experiments on ad privacy settings; Privacy Enhancing Technologies (PET); 2015.

Deng, M., Wuyts, K., Scandariato, R., Preneel, B., Joosen, W.; A Privacy Threat Analysis Framework: Supporting the Elicitation and Fulfilment of Privacy Requirements; Requirements Engineering ; 2011.

Desai D. and Kroll J.; Trust But Verify: A Guide to Algorithms and the Law; Harvard Journal of Law and Technology (31); 2018.

Dhurandhar A., Iyengar V., Luss R., Shanmugam K.; A formal framework to characterize interpretability of procedures; ICML Workshop on Human Interpretability in Machine Learning; 2017.

Diakopoulos N.; Accountability in algorithmic decision making; Communications of the ACM; 2016.

Doshi-Velez F., Kim B.; Towards a rigorous science of interpretable machine learning; arXiv:1702.08608; 2017.

Duchesne C., L. Cytermann, Vachey L., Morel M., Aureau T. ; Rapport relatif aux données d'intérêt général ; Conseil Général de l'Economie, Inspection Générale des Finances ; 2015.

Dwork C., Roth A.; The algorithmic foundations of differential privacy; Foundations and Trends in Theoretical Computer Science (9); 2014.

Dwork C., Hardt M., Pitassi T., Reingold O., Zemel R.; Fairness through awareness; Proc. of Innovations in Theoretical Computer Science; 2012.

Edwards L., Veale M.; Slave to the algorithm ? Why a right to an explanation is probably not the remedy you are looking for; Duke Law & Technology Review (18); 2017.

Edwards L., Veale M.; Eslaving the Algorithm: From a 'Right to an Explanation' to a 'Right to Better Decisions'; IEEE Security and Privacy; 2018.

Edwards M.; Price and prejudice: the case against consumer equality in the information age; Lewis & Clark Law Review (10); 2010.

Eslami M., Rickman A., Vaccaro K., Aleyasen A., Vuong A., Karahalios K., Hamilton K., Sandvig C.; I always assumed that I wasn't really that close to (her)': reasoning about invisible algorithms in the news feed; Conference on Human Factors in Computing Systems (CHI); 2015.

Esteva, Andre, Kuppel B., Novoa R., Ko J., Swetter S., Blau H., Thrun, S.; Dermatologist-level classification of skin cancer with deep neural networks; Nature (542); 2017.

Feldman M., Friedler S., Moeller J., Scheidegger C., Venkatasubramanian S.; Certifying and Removing Disparate Impact; In KDD, 2015.

Floridi L.; On human dignity as a foundation for the right to privacy; Philosophy & Technology (29); 2016.

Fredrikson M., Lantz E., Jha S., Lin S., Page D., Ristenpart T.; Privacy in pharmacogenetics: An end-to-end case study of personalized warfarin dosing; In USENIX Security Symposium; 2014.

Fredrikson M., Jha S., Ristenpart T.; Model inversion attacks that exploit confidence information and basic countermeasures; Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security; 2015.

Feldman M., Friedler S., Moeller J., Scheidegger C., Venkatasubramanian S.; Certifying and Removing Disparate Impact; Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '15); 2015.

Friedman B., Nissenbaum H.; Bias in computer systems; ACM Transactions on Information Systems (14); 1996.

Gilad-Bachrach, Ran, Laine K., Lauter K., Naehrig M., Wernsing J.; Cryptonets: Applying neural networks to encrypted data with high throughput and accuracy; International Conference on Machine Learning; 2016

Goel S., Perelman M., Shroff R., Sklansky D.; Combatting police discrimination in the age of big data; New Criminal Law Review (20); 2017.

Gu S., Rigazio L.; Towards deep neural network architectures robust to adversarial examples; Proceedings of the International Conference on Learning Representations (ICLR); 2015.

Guidotti R., Monreale A., Turini F., Pedreschi D., Giannotti F.; A survey of methods for explaining black box models; arXiv:1802.01933; 2018.

Goodfellow I.J., Shlens J., Szegedy C.; Explaining and harnessing adversarial examples; International Conference on Learning Representations; Computational and Biological Learning Society; 2015.

Goodman B., Flaxman S.; European Union regulations on algorithmic decision-making and a 'right to explanation'; AI Magazine; 2017.

Greiner D.; Causal inference in civil rights litigation; Harvard Law Review (122); 2008.

Overdorf R., Kulynych B., Balsa, E., Troncoso C., Gurses S.; POTs: The revolution will not be optimized?, eprint arXiv:1806.02711; 2018.

Hajian S., Domingo-Ferrer J.; Direct and Indirect Discrimination Prevention Methods; Discrimination and Privacy in the Information Society, Studies in Applied Philosophy, Epistemology and Rational Ethics (3); Springer; 2013.

Hajian, S., Domingo-Ferrer J., Martinez-Balleste , A.; Rule protection for indirect discrimination prevention in data mining; Modeling Decisions for Artificial Intelligence-MDAI 2011; Lecture Notes in Computer Science; 2011.

Hannak A., Soeller G., Lazer D., Mislove A., Wilson C.; Measuring price discrimination and steering on e-commerce web sites; ACM Internet Measurement Conference (IMC); 2014.

Hannah-Moffat K.; Actuarial sentencing: an 'unsettled' proposition; Justice Quarterly (30); 2013.

Hardt M., Price E., Srebro N.; Equality of opportunity in supervised learning; In Proceedings of the 30th International Conference on Neural Information Processing Systems (NIPS); 2016.

de Hert P., Papakonstantinou V.; The new General Data Protection Regulation: still a sound system for the protection of individuals?; The Computer Law & Security Review; (32,2); 2016.

Jones N.; How machine learning could help to improve climate forecasts, Nature (548); 2017.

Joyee De S., Le Métayer D.; PRIAM: A Privacy Risk Analysis Methodology; 11th IEEE International Workshop on Data Privacy Management (DPM); 2016.

Katz G., Barrett C., Dill D., Julian K., Kochenderfer M.; Reluplex: An efficient SMT solver for verifying deep neural networks; arXiv preprint arXiv:1702.01135; 2017.

Kamishima T., Ahako S., Asoh H., Sakuma J.; Fairness-aware Classifier with Prejudice Remover Regularized; PADM; 2011;

Kamiran, F., Calders, T.; Classification with no discrimination by preferential sampling; Proc. of the 19th Machine Learning conference of Belgium and The Netherlands; 2010.

Kleinberg J., Lakkaraju H., Leskovec J., Ludwig J., Mullainathan S.; Human decisions and machine predictions; National Bureau of Economic Research, NBEC Working Paper 23180; 2017.

Kloft M., Laskov P.; Online anomaly detection under adversarial impact; International Conference on Artificial Intelligence and Statistics; 2010.

Kroll J.A., Huey J., Barocas S., Felten E., Reidenberg J., Robinson D., Yu H.; Accountable Algorithms; Univ. Penn. Law Review; 2017.

Lacave C., F. Diez; A review of explanation methods for Bayesian networks; The Knowledge Engineering Review (17); 2002.

Lacave C., Atienza R., Diez F.; Graphical explanation in Bayesian Networks; Proceedings of the ISMDA Conference; Spinger Verlag, LNCS 1933; 2000.

Landecker W., Thomure M.D., Bettencourt L., Mitchell M., Kenyon G., Brumby S.; Interpreting individual classifications of hierarchical networks; IEEE Symposium on Computational Intelligence and Data Mining (CIDM); 2013.

Lecuyer M., Spahn R., Spiliopolous Y., Chaintreau A., Geambasu R., Hsu D.; Sunlight: fine-grained targeting detection at scale with statistical confidence; ACM Conference on Computer and Communications Security (CCS); 2015.

Li Z., Wang C., Han M., Xue Y., Wei W., Li L., Li F.; Thoracic disease identification and localization with limited supervision; arXiv:1711.06373; 2017.

Lei T., Barzilay R., Jaakkola T.; Rationalizing neural predictions; Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP); 2016.

Lipton Z.; The mythos of model interpretability; Proceedings of the ICML Workshop on human interpretability in machine learning (WHI 2016); 2016.

Lou Y., Caruana R., Gehrke J.; Intelligible models for classification and regression; Proceedings of the ACM Knowledge Discovery and Data Mining Conference (KDD); 2012.

Lou Y., Caruana R., Gehrke J., Hooker G.; Accurate intelligible models with pairwise interactions; Proceedings of the ACM Knowledge Discovery and Data Mining Conference (KDD); 2013.

Madumal P., Miller T., Vetere F., Sonenberg L.; Towards a grounded dialog model for explainable artificial intelligence; First international workshop on socio-cognitive systems; 2018.

Malgieri G., Comandé G.; Why a right to legibility of automated decision-making exists in the General Data Protection Regulation; International Data Privacy Law; (7,4); 2017.

Maxmen A.; Machine learning spots treasure trove of elusive viruses; Nature News; 2018; https://www.nature.com/articles/d41586-018-03358-3

McDaniel P., Papernot N., Celik Z.; Machine Learning in Adversarial Settings; IEEE Security and Privacy (14); 2016.

McMahan B., Moore E., Ramage D., Hampson S., Agüera y Arcas B.; Communication-Efficient Learning of Deep Networks from Decentralized Data; arXiv preprint arXiv:1602.05629; 2016.

Metzen J.H., Genewein T., Fischer V., Bischoff B.; On detecting adversarial perturbations; Proceedings of 5th International Conference on Learning Representations (ICLR); 2017.

Miller T., Howe P., Sonenberg L.; Explainable AI : beware of inmates running the asylum; Proc. IJCAI 2017 Workshop on Explainable Artificial Intelligence (XAI); 2017.

Mittelstadt B., Allo P., Taddeo M., Wachter S., Floridi L.; The ethics of algorithms : mapping the debate; Big Data & Society; 2016.

Monahan J.; Risk assessment in criminal sentencing; University of Virginia School of Law, Public Law and Legal Theory Research Paper Series 2015-03; 2015.

Montavon G., Samek W., Müller K.-R.; Methods for interpreting and understanding deep neural networks; Digital Signal Processing; 2018.

Moosavi-Dezfooli S.-M., Fawzi A., Frossard P.; Deepfool: a simple and accurate method to fool deep neural networks; arXiv preprint arXiv:1511.04599; 2015.

O'Neil C.; Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy; Crown Publishing Group, New York, NY, USA; 2017.

Papernot N., McDaniel P., Goodfellow I., Jha A., Celik A., Swami A.; Practical black-box attacks against deep learning systems using adversarial examples; arXiv preprint arXiv:1602.02697; 2016.

Papernot N., McDaniel P., Wu X., Jha S., Swami A.; Distillation as a defense to adversarial perturbations against deep neural networks; IEEE Security and Privacy (SP); 2016.

Papernot N., McDaniel P., Sinha A., Wellman M.; Towards the Science of Security and Privacy in Machine Learning; 3rd IEEE European Symposium on Security and Privacy, London, UK; 2018.

Papernot N., McDaniel P., Jha S., Fredrikson M., Celik Z., Swami A.; The limitations of deep learning in adversarial settings; Proceedings of the 1st IEEE European Symposium on Security and Privacy; 2016.

Pedreschi, D., Ruggieri, S., Turini, F.; Measuring discrimination in socially-sensitive decision records; Proc. of the 9th SIAM Data Mining Conference (SDM 2009); 2009.

Penney J.; Chilling effects: online surveillance and Wikipedia use; Berkeley Technology Law Journal (31); 2016.

Perdisci R., Dagon D., Lee W., Fogla P., Sharif M.; Misleading worm signature generators using deliberate noise injection; IEEE Security and Privacy; 2006.

Raab, C.; The Meaning of 'Accountability' in the Information Privacy Context; *Managing Privacy Through Accountability*; Basingstoke: Palgrave Macmillan; 2012.

Reisman D., Schultz J., Crawford K., Whittaker M.; Algorithmic impact assessments: a practical framework for public agency accountability; AINow Institute white paper; 2018; https://ainowinstitute.org/aiareport2018.pdf .

Ribeiro M., Singh S., Guestrin C.; Why should I trust you ? Explaining the predictions of any classifier ; Proceedings of the ACM Knowledge Discovery and Data Mining Conference (KDD); 2016.

Ribeiro M., Singh S., Guestrin C.; Anchors: high precision model-agnostic explanations; Thirty second AAAI Conference on Artificial Intelligence; 2018.

Rice L., Swesnik D.; Discriminatory effects of credit scoring on communities of color; Suffolk University Law Review (XLVI); 2013.

Romei A., Ruggieri S.; A multidisciplinary survey on discrimination analysis; The Knowledge and Engineering Review (29,5); 2014.

Rouvroy A.; The end(s) of critique: data behaviourism vs. due-process; in Privacy, *Due Process and the Computational Turn*; Routledge; 2012.

Rouvroy A., Poullet Y.; The right to informational self-determination and the value of self-development: reassessing the importance of privacy for democracy; Reinventing data protection; Serge Gutwirth et. al. (eds), Springer; 2009.

Rubinstein B.I.P., Nelson B., Huang L., Joseph A.D. , Lau S., Rao S., Taft N., Tygar J.D.; Antidote: Understanding and defending against poisoning of anomaly detectors; 9th ACM SIGCOMM Conference on Internet measurement; 2009.

Senden L.; Soft law, self-regulation and co-regulation in European law: where do they meet?; Electronic Journal of Comparative Law (9.1); 2005.

Selbst A. D., Powles J.; Meaningful Information and the Right to Explanation; International Data Privacy Law; (7,4); 2017.

Shah H.; Use our personal data for the common good; Nature (556); 2018.

Sharif M., Bhagavatula S., Bauer L., Reiter M.; Adversarial generative nets: Neural network attacks on state-of-the-art face recognition; arXiv preprint arXiv:1801.00349; 2017.

Sharif M., Bhagavatula S., Bauer L., Reiter M.; Accessorize to a crime: Real and stealthy attacks on state-of-the-art face recognition; Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security; 2016.

Shokri R., Stronati M., Shmatikov V.; Membership inference attacks against machine learning models; IEEE Security and Privacy; 2017.

Shokri R., Shmatikov V.; Privacy-preserving deep learning; ACM Computer Communication Security (CCS); 2015.

Song Y., Shu R., Kushman N., Ermon S.; Generative Adversarial Examples; ArXiv preprint arXiv:1805.07894; 2018.

Solon B., Selbst, A. D.; Big Data's Disparate Impact; California Law Review 671; 2016.

Spagnuelo D., Bartolini C., Lenzini G.; Metrics for transparency; Proceedings of Data Privacy Management and Security Assurance; 2016.

Starr S.; Evidence-based sentencing and the scientific rationalization of discrimination; Stanford Law Review (66); 2014.

Stone P., Brooks R., Brynjolfsson E., Calo R., Etzioni O., Hager G., Hirschberg J., Kalyanakrishnan S., Kamar E., Kraus S., Leyton-Brown K., Parkes D., Press W., Saxenian A., Shah J., Tambe M., Teller A.; Artificial Intelligence and Life in 2030. One Hundred Year Study on Artificial Intelligence; Report of the 2015-2016 Study Panel, Stanford University, Stanford, CA; 2016.

Szegedy C., Zaremba W., Sutskever I., Bruna J., Erhan D., Goodfellow I., Fergus R.; Intriguing properties of neural networks; arXiv preprint arXiv:1312.6199; 2017.

Tramer F., Zhang F., Juels A., Reiter M., Ristenpart T.; Stealing machine learning models via prediction APIs; Usenix Security; 2016.

Tutt A.; An FDA for algorithms; Administrative Law Review (83); 2017.

Wachter S., Mittelstadt B., Russel C.; Counterfactual explanations without opening the black box, automated decisions and the GDPR**;** Harvard Journal of Law & Technology; 2018.

Wachter S., Mittelstadt B., Floridi L.; Why a Right to Explanation of Automated Decision-Making Does Not Exist in the General Data Protection Regulation; International Data Privacy Law (7); 2017.

Wright D, De Hert, P.; Privacy Impact Assessment; Springer; 2012.

Yin, X., Han, J.; CPAR: Classification based on Predictive Association Rules; Proc. of SIAM ICDM ; 2003.

Yuan X., He P., Zhu Q., Li X.; Adversarial Examples: Attacks and Defenses for Deep Learning; arXiv preprint arXiv:1712.07107; 2017.

Zafar M., Valera I., Gomez Rodriguez M. and Gummadi K.; Fairness Beyond Disparate Treatment & Disparate Impact: Learning Classification without Disparate Mistreatment; Proceedings of the 26th International Conference on World Wide Web (WWW '17); 2017.

Zarsky T.; Transparent predictions; University of Illinois Law Review; 2003.

Zeiler M., Fergus R.; Visualising and understanding convolutional networks; Proceedings ECCV, Springer, LNCS 8689; 2014.

Zemel R., Wu Y., Swersky K., Pitassi T., Dwork C.; Learning fair representations; Proc. of Intl. Conf. on Machine Learning; 2013.

Zuiderveen Borgesius F.; Online price discrimination and data protection law; Amsterdam Law School Legal Studies Research Paper, No. 2015-32; 2015.

Zuiderveen Borgesius F., Trilling D., Möller J., Bodo B., H. de Vreese C., Helberger N.; Should we worry about filter bubbles?; Internet Policy Review (5); 2016.

The expected benefits of algorithmic decision systems (ADS) may be negated by the variety of risks for individuals (discrimination, unfair practices, loss of autonomy, etc.), the economy (unfair practices, limited access to markets, etc.) and society as a whole (manipulation, threat to democracy, etc.). This study presents existing options to reduce the risks related to ADS and explain their limitations. We sketch some policy options to overcome these limitations to be able to benefit from the tremendous possibilities of ADS while limiting the risks related to their use. Beyond providing an up-to-date and systematic review of the situation, the study gives a precise definition of a number of key terms and an analysis of their differences. The main focus of the study is the technical aspects of ADS. However, to broaden the discussion, other legal, ethical and social dimensions are considered.

QA-06-18-337-EN-N

This is a publication of the Scientific Foresight Unit (STOA)
EPRS | European Parliamentary Research Service

ISBN 978-92-846-3506-1 | doi: 10.2861/536131 | QA-06-18-337-EN-N