

The forgotten human autonomy in Machine Learning

Paula Subías-Beltrán^{1,2}, Oriol Pujol³ and Itziar de Lecuona^{2,4}

¹Eurecat, Centre Tecnològic de Catalunya, Unit of Digital Health, Barcelona, Spain

²Bioethics and Law Observatory - UNESCO Chair in Bioethics, Universitat de Barcelona, Barcelona, Spain

³Dept. de Matemàtiques i Informàtica, Universitat de Barcelona, Barcelona, Spain

⁴Dept. of Medicine, Universitat de Barcelona, Barcelona, Spain

Abstract

There are many rights, freedoms, and principles that build our society and nourish it day after day. No right, freedom, or principle is absolute; they must always be balanced. Our starting point is the respect for internationally recognized human rights and we focus on the principle of autonomy, which is not being adequately treated and protected in the ever-changing algorithmic world. In this article we review some of the most influential bodies of knowledge and ethical recommendations in artificial intelligence, and analyze the extent to which they address the principle of autonomy. We ground the concept of autonomy in operational terms such as being well-informed and being able to make free decisions. Under these two different aspects, we analyze the technical and social risks and propose different ways in which artificial intelligence requires further exploration with the aim of preserving human autonomy.

Keywords

human autonomy, machine learning, AI ethics, bioethics and human rights

1. Introduction


The basis of the European society is the respect for the rights and freedoms constitutionally recognized [1]. But we must keep in mind that no right or freedom is absolute, and that there is always some tension between the rights and freedoms at stake. And this is the real challenge. Weighing what should prevail in specific cases gives rise to ethical debates for which there is not a unique valid answer. But technologists cannot work with uncertain terms.


The need of well-defined concepts leaves no room for interpretation. The articulation of the right to non-discrimination illustrates this point clearly. This right has been mathematically addressed through different perspectives corresponding to the notions of independence, separation, or sufficiency. However, these concepts are mutually exclusive among them [2]. This means that a general regulating wording requiring non-discrimination can be multiply realized. And there is no common agreement on which is the most appropriate realization for a given situation (this can be seen in the discussion between Northpointe [3] and ProPublica [4] about the performance of COMPAS).

IAIL 2022 - Imagining the AI Landscape after the AI Act, June 13, 2022, Amsterdam, Netherlands

✉ paula.subias@eurecat.org (P. Subías-Beltrán); oriol_pujol@ub.edu (O. Pujol); itziardelecuona@ub.edu (I. de Lecuona)

ORCID 0000-0003-1167-1259 (P. Subías-Beltrán); 0000-0001-7573-009X (O. Pujol); 0000-0002-5081-5756 (I. de Lecuona)

 © 2022 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

 CEUR Workshop Proceedings (CEUR-WS.org)

Automation has a long history. Automation has been motivated by the search for efficiency, cost reduction, and even equal access to opportunities, among others. In a society with ambitions to prosper while preserving all rights, automation is necessary to ensure progress. Until recently, only repetitive processes that required solely trivial reasoning were automated. But for some time now, more complex reasoning, such as decision making hitherto relegated to individuals, has started to be automated. It is in this context that machine learning (ML) has appeared as a predictive tool that makes it possible to evaluate the impact of certain actions in the future. Systems using ML are based on finding complex patterns in data that allow to find associations with the desired outcomes to be predicted. The products and systems based on ML are sometimes referred as artificial intelligence (AI) in current folk terminology. ML explicitly refers to the branch of this discipline concerned with systems and methods that improve their performance based on experience. Experience that many times is translated in data. ML contributes to improved decision making by removing noise from reasoning, making automated systems more robust decision-making agents than humans, but the quality of the data they are trained on should be accurate enough to avoid the leakage of undesired judgement that may perpetuate and magnify injustices [5]. The problem, in the case of undesired judgement, arises in systems that process personal data, as they directly shape the being of the individuals that the data represent.

The inexorable advance of technology means that, over time, more and more actions may be automated [6]. But, should they? Parasuraman and Wickens [7] thoroughly analyse the particularities of automation by reflecting on a number of use cases for the different stages of automation. They suggest that reliance and compliance ought to be balanced and recognize the importance of adaptively integrate human and automation performance [7]. The current almost ubiquitous presence of ML makes it necessary to review this issue in terms of its permeability and reach. Laitinen and Sahlgren's work makes progress among these lines by providing a study on the effects that ML may have on people's autonomy by examining the sociotechnical bases of the principle of autonomy [8]. This approach reinforces the importance of the respect for autonomy from a philosophical perspective but, in our view, lacks a technical analysis that reflects how ML can better address this principle. From our point of view, a systematic understanding of how ML may contribute to the respect of people guaranteeing our autonomy is still lacking.

In the last decade, the ML community has been active and concerned with rights and principles such as privacy, transparency, accountability, or fairness [9]. But there are still other rights, freedoms, and principles not properly tackled by ML. And the principle of autonomy is one such example.

Addressing the problems laying at the intersection of many disciplines, such as in this case, is difficult. These here require the scientific point of view, that of the technology itself, together with ethical, normative, and societal perspectives. But even in that case, a transdisciplinary team should be able to go beyond the traditional boundaries of each discipline and face challenges not necessarily well-defined or present in their own field of study. A first step into this direction in order to successfully implement this practice is to work on developing a common language, which would allow us to understand the challenges of each discipline and address them with a holistic approach. One of the ambitions of this work is to bring us a step closer to this end by analysing the challenges that arise from the need to assure the respect for autonomy in the

design, development, and implementation of ML-powered systems.

Stakeholder profiles of ML solutions are very diverse and, today, they have different needs. This article contributes to the creation of an ethical framework through an ethical evaluation of the current limitations and challenges posed by ML, as well as to narrow the gap between the abstract definition of autonomy and its operational translation.

To put it concisely, there are four primary aims of this study:

- emphasise the importance of the respect for autonomy in ML solutions;
- distill the principle of autonomy into two operational axes that enable a practical articulation, namely, being well-informed and being able to make free decisions;
- analyse the European strategy in relation to the principle of respect for autonomy in the design, development, and implementation of ML-based systems; and
- identify and analyse challenges in the design, development, and implementation of ML to ensure respect for autonomy from a technical approach and from the perspective of the rest of stakeholders.

This paper begins introducing the concept of autonomy in section 2, with a reflection of its importance on the development of democratic societies and the different trends and dynamics that are endangering it. Then, the current normative European framework is presented in section 3. This includes a bioethical perspective to analyse ethical, legal, and social issues from an interdisciplinary perspective, as well as a review of the different voices demanding that autonomy be given the weight it deserves. This section closes with an analysis of the AI Act, the first European law on AI. The fourth section is concerned with the thorough presentation of the risks and challenges posed by ML that govern the respect for autonomy. There, we dissect the concept of autonomy to raise the challenges we face in technical development and for other stakeholders. The article ends with a discussion of the reflections raised so far and the conclusions.

2. The principle of autonomy in danger

The word autonomy originates from ancient Greek, and it can be decomposed in two words: *autos*, meaning her/his own, and *nomos*, standing for law. *Autonomy* describes the ability of a person to make her or his own rules in life and to make decisions independently. There is a number of definitions of autonomy, but the idea that freedom is a requisite for people to form their own lives is fundamental to most of its interpretations [10].

Being autonomous can be operationally defined as being well-informed and being able to make free decisions. Being well-informed means having sufficient information that describes the situation of interest in a relevant, pertinent, and neutral manner such that the individual understands impacts and consequences. While being able to make free decisions implies to have a chance to speak and the opportunity to act accordingly, i.e., having decision-making capacity.

The principle of autonomy is substantially related to privacy and self/determination, among other rights, freedoms, and principles; and it is part of the very foundations of human rights. The articulation of these moral norms is not straightforward, since none of them conforms a domineering criterion that overrules the rest by default. The application of such moral code

consists of the balance between competing human rights, freedoms, and principles for each use case.

Democracy (*demo*, meaning people, and *kratos*, standing for rule) is the form of government in European countries. By definition, the principle of autonomy is one of the most important pillars of democracy. If our autonomy is reduced, third parties can nudge us towards different judgements responding to their desires and interests. We tend to focus on violations of our rights and freedoms when they affect us individually. But it is also important to talk about their impact on the collective.

The ever-changing modern world overwhelms us and our way of responding to the excessive stimuli is to become docile [11]. In our day-to-day lives we value the transitory over the permanent, the immediate over the long-term, and utility over any other value. We live a *liquid life*, where everything is constantly changing and individuals are forced to adapt continuously [12]. To cope with this rush, immediacy, and time limitations we are delegating decision-making to automated systems [13]. Instead of treating these systems as the tools they are, we use them as our *alter ego*, forgetting that it is in our hands to take back our agency over ML decisions. As a result, the balance between decision automation and our responsibility to make decisions is out of whack. Thus, autonomy is being relegated to the sidelines. We are giving our autonomy to systems designed to assist us in our decision-making. We are allowing these tools to stop being passive elements and become active actors in decision-making. We are opening the door to them to enter the human decision space.

We have reached a point where these tools may know us better than we know ourselves. The underlying algorithms and data have captured enough relevant information about us that they cannot only understand us, but, by processing that information, can also engineer us. Thus, the owners of these tools may discover what kind of push is needed to convince us of a specific idea. In Yuval Noah Harari's words, we have become *hackable animals* [14, 15]. The increasing available amount of computing power and data is getting us closer to the idea that, soon, some corporations and governments will be able to systematically hack people.

There are many powers that benefit from the fact that we do not prioritize our autonomy. Today, this is a concern that falls on society. Those who move the world are not particularly interested in us imposing our autonomy, and we, as part of a state governed by the rule of law, must be aware and claim the value it deserves. ML practitioners may lack the societal background needed to realize about the implications of limiting the information that people receives to form their judgements. At the same time, the rest of stakeholders may not have enough technological background to properly understand the ins and outs of technological solutions and assess its power and ramifications. In other words, ML practitioners are capable of developing solutions but they cannot really assess whether they respond to the actual societal challenges, while the rest of stakeholders cannot thoroughly comprehend the effect of technological solutions.

It has been reported that many institutions world-wide are currently profiting of our more and more diminished autonomy by making us dependent, exploiting our weaknesses [16]. There is a tendency towards commodification, pricing things before thinking about their value [17], that impacts all areas and practices. This trend gives a competitive advantage to those organisations that are able to adapt more quickly to change. And, this, undoubtedly puts enterprises in a more favourable position in the race of liquid life. Private initiatives are targeted and capable of adapting to the constant uncertainty and speed of change that is now part of our way of

life. However, public initiatives should not be neglected. Both initiatives have taken steps in this direction in pursuit of their profits by de-emphasising human values. We have ceased to be subjects to become objects. We still have the agency, but we do not have enough of time to apply it.

So far, we have focused on respecting people's autonomy in machine-human interactions for those people who are *de facto* autonomous. However, what about people who suffer a reduction of their autonomy? Whether due to natural (i.e., age, medical conditions) or social causes (i.e., immigration, poverty), there are certain discrimination that, if perpetuated, imply the promotion of systemic inequalities. For this reason, we should not only focus on people who are already autonomous to understand how ML may limit their autonomy, but we should also analyse what happens to people who are not and investigate whether ML can be an opportunity to promote their autonomy. Not only those who are *de facto* autonomous are entitled to have their autonomy respected, and ML-based solutions run the risk of inheriting existing problems and reproducing them in the artificial world. This is why we must be aware of the societal inequalities that make it difficult for people to build a self-determined life. Unsympathetic attitudes towards these systemic imbalances conceal unjust realities, where discrimination has conditioned the present limiting people's capacity for action and decision-making.

The COVID-19 pandemic demonstrated that in critical situations a pragmatic approach may be chosen to provide a rapid response to the ongoing conflict, tending to push ethical considerations into the background for the sake of an imminent good. These situations raise multiple ethical debates that are difficult to resolve by finding the right balance between rights, freedoms, and principles. Be that as it may, the multiple use cases that occurred during the COVID-19 pandemic are a good example that in no situation should ethical considerations be put aside [18].

We need to look at the balance between benefits, harms, and obligations. Not only in order to provide guidelines for proper development and ensure prosperous progress, but also to influence policy and legislative changes in the years to come. In other words, we must analyze this complex issue from a transdisciplinary approach, where all stakeholders could be represented and have a voice in the analysis of the implications and consequences of ML systems by building a common language.

3. Current normative European framework

Bioethics is the multi-disciplinary study of the ethical issues emerging from advances in biology, medicine, and technologies, and it is grounded on four pillars: respect for autonomy, beneficence, non-maleficence, and justice. The link between bioethics and human rights [19] was recognized by the *Universal Declaration on Bioethics and Human Rights* (UDBHR) of UNESCO in 2005 [20].

The UDBHR demands the respect for the autonomy of individuals with regard to the power to make decisions, taking responsibility for these decisions and respecting the autonomy of others. Autonomy stands as one of the key principles to protect human sovereignty. To further promote democratic societies as we know them, we must prevent actors who do not serve this interest from deciding the rules of the ML game. The EU must act as a global standard-setter in AI [21].

Today, there are many voices calling for a unified strategy to protect the rule of law from being undermined by flawed ML developments. But this is not a new issue. The founding of the Council of Europe (CoE) in 1949 meant the creation of an international entity to uphold human rights, democracy, and the rule of law in Europe. One of the best known bodies that hangs on the CoE is the European Court of Human Rights, an international court that interprets the European Convention on Human Rights. Another body of CoE is the Directorate General Human Rights and Rule of Law, which has overall responsibility for the development and implementation of the human rights and rule of law standards of the CoE. And there is also the Commissioner for Human Rights, an independent and impartial non-judicial institution dedicated to engaging in dialogue with national authorities and civil society, and analysing, advising, and raising awareness of systematic human rights work. In January 2019 the Directorate General Human Rights and Rule of Law published a proposal of guidelines [22] while in May 2019 the Commissioner for Human Rights issued a Recommendation [23] that addressed the risk of ML-based systems undermining human rights rather than strengthening them. Note that a Recommendation is a non-binding legal instrument proposed by the European Union (EU) that calls upon the party to whom it is addressed to behave in a particular manner without any legal obligation. The approach of both proposals is to empower individuals by ensuring that they are able to understand the why of AI-powered decisions as well as their verification process. Besides, in order to increase trust in AI, they encourage AI developers and vendors to design products in such a way that they safeguard users' freedom of choice over the use of AI.

In 1998 UNESCO set up the World Commission on the Ethics of Scientific Knowledge and Technology (also known as COMEST), with the ambition of creating an advisory body for decision-makers, a body capable of whistleblowing risky situations, and a forum of reflection. In 2019 COMEST published a preliminary study on the Ethics of AI [24], which included multiple ethical considerations with regard to AI development and suggested several elements that could be included in an eventual Recommendation on the Ethics of AI. Regarding the principle of autonomy, COMEST suggested several generic principles among which we find the respect for human autonomy, which they propose to articulate as the demand for human control at all times. This must be understood as a continuous and perpetual monitoring of AI by an educated person or group of people.

In 2007 the European Union Agency for Fundamental Rights was established, which is better known in English as the Fundamental Rights Agency (FRA). The FRA is a EU independent body that helps to safeguard the rights, values, and freedoms that set the EU's Charter of Fundamental Rights at the EU, national, and local level. Due to the role that AI plays in many decisions that affect our daily lives, in 2020 FRA published a report called "Getting the future right – Artificial intelligence and fundamental rights" [25] where they provided an overview of the use of AI in the EU, an analysis of the awareness of fundamental rights and further implications, and a discussion of measures to assess and mitigate the impact of AI on people's fundamental rights. Among others, this report increases the understanding of how AI-based solutions may cut across different rights. However, it does not address the principle of autonomy specifically. One of FRA's main proposal is that people should be able to contest decisions based on the use of AI, which, although unstated, would strengthen people's autonomy. The only place where this principle is explicitly mentioned is in the section that addresses the impact of AI on the rights to respect for privacy life and the protection of personal data. The report defends both

rights on the premise that they both strive to protect similar values, such as autonomy and human dignity. On this account, the two aforesaid rights form an essential prerequisite for the exercise of other fundamental rights.

In June 2018, the European Commission appointed a group of experts to provide advice on its AI strategy. This group was called High-level Expert Group on Artificial Intelligence (AI HLEG), and, since its appointment, acts as the steering group of the European AI Alliance. The AI HLEG, also in 2019, published the Ethics Guidelines for Trustworthy AI [26], where they list seven key requirements that AI systems should meet in order to be trustworthy based on human rights and ethical principles. AI HLEG adopts respect for autonomy as one of the ethical principles on which AI should be built. AI HLEG advocates that AI systems should ensure the full and effective self-determination of their users. They also argue that the allocation of roles between humans and AI systems should follow human-centred design principles and leave a meaningful opportunity for human choice, and they present human supervision as the mechanism to achieve this. In short, AI HLEG defends that AI systems must serve a “democratic, flourishing, and equitable society” by supporting people’s agency and enabling human oversight. According to them, this mechanism will ensure that AI-powered systems do not undermine human autonomy or cause other adverse effects. This has not been without debate about the voices represented in its creation, with a minority of four ethicists and significant industry representation [27]. But should it be the voice of industry the one pulling the strings?

Last but not least, the CoE appointed the Ad hoc Committee on Artificial Intelligence (CAHAI) in December 2019 and it was tasked with reviewing the feasibility and possible elements of a legal framework for AI. In 2020 they published their work named “Towards regulation of AI systems. Global perspectives on the development of a legal framework on Artificial Intelligence systems based on the Council of Europe’s standards on human rights, democracy and the rule of law” that explored the effects of AI over fundamental rights, studied the initial background of AI, and analysed, overall, the already published ethical AI guidelines [28]. CAHAI defends people’s self-determination, which they express as their freedom of choice and autonomy. They fear that AI threatens to reduce the human agency and autonomy needed to make meaningful decisions. However, they do not provide their own definition of autonomy, although they comment that soft law documents often speak of autonomy as a positive freedom (i.e., to flourish, to decide for oneself, and to self-determine one’s own course of action), while a minority of documents, however, express it as a negative freedom¹ (i.e., to be free from technological experimentation, manipulation or surveillance).

All the documents published by the listed organisations work along the same line of argument. They agree that being autonomous agents is the basis for respect for people. They also argue that the principle of autonomy should be translated into the ML systems by ensuring full and effective space for self-determination, as well as the need to establish mechanisms to exercise our autonomy. CAHAI even suggests that autonomy should become a new human right. However, all of them fail to propose an effective way of operationalizing the principle of autonomy so that it can be distilled and transformed into actionable language, moving from the abstract to the practical.

¹The term “negative” is used to define autonomy through complementary concepts. In a mathematical analogy, given the term $A \subseteq B$, A may be defined as $B \setminus \neg A$.

Furthermore, we need to take into account other norms that regulate this issue. Therefore, we must consider the General Data Protection Regulation (GDPR). This regulation presented new rights that react and allow people to exercise control over the algorithmic world. However, it was not enough. There is no symmetry in the relations among the different ML stakeholders: ML professionals have the technical knowledge about the solution and have a tendency to evaluate algorithms solely from the perspective of their accuracy, without assessing whether they are adequate under other social rules. On this account, the opinion of other stakeholders is mistreated during the design, development, and implementation of the systems in favour of the search for the most accurate algorithm. And this imbalance expresses the evident need for better strategies. The challenge is then to build a framework capable of addressing the ML challenges from the prism of human rights [29].

Demands for reliable AI have culminated in the AI Act [30], which puts on the table the EU's intention to become a global AI standard setter. The AI Act proposes a risk-based approach constructed on a pyramid of criticality accompanied with a layered enforcement mechanism. It has received criticisms that may be summarised in two key points: the AI Act does not provide coverage for all people as it does not sufficiently focus on vulnerable groups, and it does not provide meaningful rights and redress mechanisms for people impacted by AI systems, as stated by the Ada Lovelace Institute [31], whose mission is to ensure that AI serves people and society, as well as many other organisations representing civil society [32]. Additionally, returning to the issue at hand, the AI Act does not explicitly and completely address the principle of autonomy, although it does so implicitly and partially. Multiple applications are forbidden, like manipulative, exploitative, and social control practices; AI systems intended to distort human behavior; or those that provide social scoring of natural people. But there are less clear prohibitions, such as the one that forbids "the placing on the market, putting into service or use of an AI system that deploys subliminal techniques beyond a person's consciousness in order to materially distort a person's behaviour in a manner that causes or is likely to cause that person or another person physical or psychological harm" (Title II, Article 5, 1a). But what about practices that are explicitly manipulating our conception of reality, such as image filters that perpetuate impossible standards of beauty? Besides, this framework wrongly implies that a person's behaviour can be altered in an innocuous way, while such practices are used to undermine the essence of our autonomy.

One mechanism proposed by the AI Act is mandatory human oversight in high-risk AI-based systems. The implementation of such a measure implies the involvement of a person or group of people trained to perform the tasks defined in Title III, Chapter 2, Article 14, among which are: to fully understand the capabilities and limitations of the system such that anomalies and malfunctions are identified and mitigated as early as possible, to be able to detect automation biases, to interpret the system output correctly by knowing the particularities of the system, to be able to decide to refuse the use of a high-risk AI system, and to be able to intervene during its execution and even to stop the use of the system if considered appropriate. But, are we ready to respond to this point effectively? On the one hand, there are few profiles trained in all the disciplines necessary to carry out this task correctly. On the other hand, the immediacy required in the responses of AI-powered systems also makes this type of role uncertain due to the different pressures to which it may be subjected. It should be noted that the AI Act makes no mention of the characteristics of the person or group of individuals who can perform this

task. Consequently, we find ourselves in a scenario where there is freedom for individuals both internal and external to an association to carry out this task. An in-house worker may be under pressure from the employer to withhold certain information from the public or may be overworked performing this role in several initiatives and not be able to perform the work adequately due to the pressure of having to cope with everything. On the other hand, an outsourced worker has the benefit of being impartial and protected from employer pressure. But accessing the innards of the system in question is going to be more difficult. Although this mechanism has the potential to bring a lot of value not only in detecting and mitigating risks, but also in bringing ethical debates closer to those who develop the system, are we prepared to put it into practice?

The AI Act proposes a regulation based on several rigid components. This is exemplified by the list of prohibited systems, which covers specific casuistries that may become obsolete due to the rapid development and updating of use cases for data-driven systems. In our opinion, this rigidity calls into question the future relevance of the AI Act and its ability to respond to future developments and emerging risks, so it would be worth exploring the rationale for this rigidity in more detail. As is common practice, we rely on the past to legislate, but how certain are we about its suitability to cover future issues? Compliance with the legal framework is necessary, as it marks out what can and cannot be done. But the legal framework does not provide enough information to know what is right and what is better. Although laws and regulations are gearing towards the idea of ML as a tool to support decisions, this has not yet been translated into practice. Changes in the perception and use of such solutions do not happen at a rapid pace. There is still not a sufficiently respectful data management culture (note the fines that the Dutch Tax and Customs Administration [33], Google [34], and Enel Energia [35] got because of their inadequate data management in the last year). On the other hand, there will come a time when, for our own efficiency, we may want to delegate our decisions to ML systems and, consequently, these will cease to be decision support systems and will take the decisions themselves. And, as of today, the regulation does not contemplate this case. It is necessary to contemplate that this may happen and, therefore, systems must be designed to assume this delegation. This implies that they must be able to understand the context, to understand that a human decision is being replaced in a human context, and that it must preserve the purpose of the person who delegated autonomy to it. Whatever the implementation, the balance must remain positive for the person involved.

The legal framework is insufficient to steer society in the right direction. This is the task of both the ethics of ML and a good ML governance. ML governance is the practice of defining and executing procedures for the proper development and use of ML, while ML ethics evaluates what is socially acceptable or preferable. We must go beyond compliance [36] and set out an adequate ethical framework that complements the legal bodywork and allow us to progress towards the kind of society we want to be. In this way, we will be prepared to react to situations to which the body of law does not yet respond.

4. How machine learning modulates the principle of autonomy

We focus on a social framework where people are autonomous agents and technological solutions should not limit their autonomy by default. Furthermore, we will address the actions that are triggered after human-ML interaction. We evaluate the equilibrium between how much autonomy we should give up, in which cases, and according to which rules. As we stated before, autonomy can be simplified to being well-informed and able to make free decisions. In the following paragraphs we elaborate on the risks and challenges posed by ML in both aspects of autonomy from two different perspectives: that of technological profiles and that of the remaining stakeholders.

4.1. Technological challenges

In a broad sense, ML systems can be grouped based on their usages: descriptive, predictive, and prescriptive. Descriptive analytics summarize historical data and allow us to explore it, predictive systems aim at answering the questions “what will happen?” and “why will it happen?”, and prescriptive systems reply to the questions “what should I do?” and “why should I do it?”. Descriptive systems do not pose any risk from human-machine interaction from an autonomy perspective beyond the human biases that may affect the interpretation of the results of the analyses. These biases can lead to misconceptions of reality that can result in impaired judgement and, consequently, reduced autonomy. But it is not the descriptive systems themselves that alter autonomy; rather, it is the presentation of the results of these systems that can alter our conceptions (e.g., framing effect or distinction bias). Accordingly, in the following paragraphs we will focus on the challenges posed by predictive and prescriptive systems on autonomy.

4.1.1. Being well-informed

Some ML systems such as recommenders or those based on information retrieval techniques are usually based on finding the optimal balance between their explorative and exploitative capabilities. Exploration consists on the examination of the space of solutions (e.g., scouring the entire supermarket for the best products), while exploitation implies finding the optimal solution from a set of known and already explored solutions (e.g., choosing the best products from the aisles visited in a supermarket). Observe that exploitation only considers already known information and disregard future and more profitable solutions that have not already been explored. Recommendation systems can be regarded as soft-prescriptive systems that suggest what to do in the form of a ranked list of options based on an understanding of the user profile and their context. Currently, most of these systems balance exploration and exploitation policies with the goal of maximising specific goals: such as maximising profit or maximising prediction of user’s future choices. This may result in well-known effects such as filter bubbles, epistemic bubbles, or echo chambers [37]. However, to break these effects it is necessary to establish adequate mechanisms that ensure avoiding hindering user’s autonomy. For example, the deliberate exposition of users to a diverse set of information options or choices beyond those that strictly maximize the adequacy to their preferences and points of view. This may be

done either by balancing a good level of exploration while exploiting the space of solutions or by explicitly exploiting alternative options that satisfy a diversity criterion. These options that balance content appropriateness with diversity, would increase the level of well-informedness and, thus, the autonomy of end-users. Regulation could enforce specific mechanisms that ensure the exposition to diverse alternative options and points of view of the user.

Prescribers attempt to quantify the effect of future decisions, but the best future action devised by the system may not be good enough to be accepted. Being well-informed through ML systems not only involves maximising diversity while still being relevant to the user and the organization's goals, but also knowing the degree of confidence and the underlying logic that led to a prescription. Knowing why a prescription has been suggested will make the user better informed. For example, if we are talking about a news prescriber, it will be interesting to know which news you should read because they have a different point of view from yours and which ones because their content matches your interests. In the same way, knowing the confidence level of prescriptions will provide the user with information to assess the degree of reliability that is associated with them. Thus, the user will have sufficient information to weigh the different contextual factors that complement the information received in order to make a well-informed decision.

4.1.2. Being able to make free decisions

In the design of the former systems, developers of ML solutions should be aware of the cognitive biases that come into play when evaluating prescriptions. The presentation of the results can make or break a decision in the light of the same exact information. The anchoring bias, for example, is the tendency to rely on the first piece of information that is received, which may alter the user's decision-making [38].

But, information retrieval systems are not the only way of decision-making by means of ML techniques. On the contrary, the most usual way of decision-making by means of ML techniques directly refers to the so called decision-support systems. These are usually predictive systems. It is important to highlight that a predictive system may be regarded as a system that describes the future. It does not leverage different options, their impact or consequences. However, many times the predictions given by a ML system are taken as optimal decisions at face value. As a ML developer, one cannot ignore that a predictive system will be used as a prescriptive or a decision-making system. That we naturally delegate our decision-making into these systems is a difficulty that is not easily solved. We often do not have the time to add a human-in-the-loop to the loop, and this leads to predictive systems being used as prescriptive systems. As a result we are in front of a problem in no-one's land. From a technological perspective this is simplified and delegated to the person acting as the human-in-the-loop. This suggests, first, the lack of accountability of the implementer, and second, the dependence on a third party. In the first case we find a lack of accountability in the disregard of users' rights, while the second raises a different set of questions, such as: what if the person acting as human-in-the-loop does not have time to cover everything? As such, technologists must assume that the system will be used as a prescriptive system and implement the adequate measures assuming that the participation of the human-in-the-loop will be minimal or even negligible. Thus, the system could be self-sufficient in adjusting human involvement in the loop according to the uncertainty derived from the

prescription, also known as *adaptive automation*. Whatever the case may be, the system needs to offer maximum transparency to enable the human-in-the-loop to fully assess the system, from its technical implementation to its compliance with respect for human rights, freedoms, and principles.

There is but another important assumption when a prediction system is used as a decision-making mechanism in the context that it interacts with other humans. While ML leverages the power of statistical, probabilistic, or even causal prediction theory, this has little to do with actual human decision or choice. Disciplines such as decision theory or rational choice theory study different aspects of the process of making decisions [39]. Decision theory is concerned with the challenge of optimal decision-making under certain consistent conditions. On the other hand, rational choice theory focuses on the behavioral aspect of individuals' judgements. In this context, ML predictions are just an input to a decision system that has to leverage more contextual aspects in order to postulate the optimal options to the user. The most widely used underlying assumption that bridges the gap between both is the concept of maximization of expected utility [40], which stands for the search for the decision that entails the highest utility value. This assumes that the decision agents are fully rational and are able to assign a utility value to the different choices at their disposal. And this is a perfect assumption when considering the ML system in isolation. However, when these systems interact with human beings other contextual factors should be considered to avoid the rejection of the proposed decision. For example, a prescription method should be aware of the context under which the decision is being made when the decision involves some kind of risk. The chance of accepting a prescription of a decision support system may depend on the individual's attitude towards risk or even to the degree of risk the individual assigns to the situation. For example, if users have a good economic status, they may be more keen to accept more risky gambles concerning gaining or losing wealth even if the expected utility is lower than that of another option. This is known as framing effect. Also, as the name itself indicates, expected utility considers expectations. But, many times we are in face of a single unique decision. In this particular situation would knowing that this is the best available option on average satisfy our needs? Thus, the rationality of the decision system as well as the utility theory implicit in the development of ML automatic systems are assumptions that deserve further development [41] by including aspects of human psychology and choice preferences such as those proposed in prospect theory [42]. Thus, the construction of a context that allows the fit between humans and automated decision mechanisms requires of further thought.

Equally important is the early development of mechanisms to be executed in case the prescription is rejected. In order to ensure people's decision-making capacity, it is important to offer alternatives to the automated procedure. As stated by Parasuraman and Riley [7], humans are still vital after all these years of automation. In the case of recommendation systems, we may count on a never-ending list of suggestions. In the event that ML solutions fail, there must be alternatives to achieve the same objective, even if the redress mechanism is not as efficient as the algorithmic one. For instance, misrecognition may limit the access to goods and resources. Misrecognition in facial recognition systems illustrates this point clearly. An access control system not correctly identifying users, in situations such as building [43] or border [44] access, will be limiting people without justification. Being able to opt-out from the application of these systems would help to further protect people and their autonomy. There are a number

of new rights, such as the right to be forgotten or the right not to be subject to automated decision-making, developed to protect us in the changing algorithmic world. Even so, there are still unresolved challenges, such as those related to the principle of proportionality that should place value on finding the right balance between benefit and non-maleficence.

4.2. Stakeholders' challenges

But the impact of ML systems also depends on the usage we do of them. For some time now there has been a trend towards the criminalisation of algorithms and process automation from non-technical stakeholders². But algorithms are just tools designed to make our lives easier, they are not meant to act on our behalf. As stakeholders we must know which party is responsible for what, and taking our share of responsibility as stakeholders it is necessary to build a reliable framework for ML.

4.2.1. Being well-informed

ML models build their reasoning extracting knowledge from data. Thus, data enforces the limits of models' capabilities. Data may be collected or created. In the former case, data aims to capture a specific reality, but its scope is restricted to the given context and the collection methods. Thus, data must be interpreted as a crystallization of a specific context at a particular moment in time, which is the result of the succession of many previous events. ML algorithms process this information to extract patterns and abstract general rules of behaviour. And to use them correctly, we must understand their limitations, such as the historical and societal influence that affects the given piece of data, the biases that may affect data collection or data engineering, and even cognitive biases that could potentially impact the ML pipeline. ML models respond by following their learned reasoning based on what they know, but they do not know how to act given information that does not resemble what they have seen so far. For example, if a model trained to discern between images of blond and brunet people receives an image of a person with red hair, it will give a response that escapes its logic. Consequently, we must understand ML for what it is: a tool for solving specific tasks, not for creating globally applicable reasoning using only local knowledge. The misuse and misunderstanding of ML systems will result in the misinterpretation of the proposed solutions (read [48] to know more about the factors associated with misuse).

Human-created data responds to the necessity of articulating concepts in a machine-readable format, as in the case of taxonomies. Likewise all human creations, data generated by humans also inherits humans' subjectivity. Let us analyze a particular example, such as the image database of ImageNet, based on the lexical English database of WordNet. ImageNet contains about one thousand images to illustrate each meaningful concept of WordNet. These images are quality-controlled and human-annotated. Prior to 2020, ImageNet contained more than three thousand concepts that characterized people, ranging from race to professions. Due to the impossibility to visually characterize many of those traits, ImageNet opted for leaving

²The following excerpts from newspaper's headlines exemplify the aforesaid criminalisation of algorithms: "algorithms are biased" [45] or "this is how the algorithms that decide who gets fired from jobs work" [46] from the Spanish newspaper *El País*, or the characterization of "creepy algorithm" from the US magazine *Wired* [47].

them out³ [49]. Identity is a social construct which is derived from the self-determination of each individual. Thus, ignoring people's self-determined identity results in non-consensual classifications. Understanding who defines human-created data permit us to understand under what perspective we are analysing the information [50].

ML carries the patina of objectivity and neutrality because of the misconception related to ML reasoning capabilities: ML decreases variance and noise in judgements, but this does not imply an increase in neutrality. ML reasons in a deterministic manner, which does not entail an objective reasoning. However, this patina of objectivity and neutrality makes some people place machines "opinion" before theirs [51]. We, as users of ML systems, must understand these solutions for what they are: tools that allow us to analyse information more robustly, eliminating variance in decision-making, but which are limited by the restrictions of data.

Another challenge we face is to respond to people who are not *de facto* autonomous. Adequately informing people who have certain limitations to be so, for example due to medical conditions or age, is a challenge to which ML may be able to make some proposals for improvement.

4.2.2. Being able to make free decisions

We have a tendency to anthropomorphize the non-human, and relating the terms "AI" and "trustworthy" exemplifies it. But "AI is not a thing to be trusted. It is a set of software development techniques by which we should be increasing the trustworthiness of our institutions and ourselves" [52]. Then, we should avoid the usage of this term and talk about *reliability* instead [53]. This has already been shown to be a problem by Robinette et al., who executed an experiment where humans *over-trusted* robots in situations of emergency evacuation [54]. The conjecture that ML-based systems are more reliable than humans may lead to the conclusion that they should be allowed to make decisions of paramount importance [50], escaping the need to question the rationale behind the proposed solutions. The over-reliance on ML-powered solutions decreases our capacity to make decisions.

People who are not able to make free decisions for whatever reason, for example because of their socio-economic context, suffer a reduction in their autonomy. Investigating how ML can contribute to strengthening the autonomy of people who are not *de facto* autonomous is an open research area that deserves attention.

ML-solutions have rapidly permeated our lives impacting our ways of doing and deciding. The AI Act covers many cases, but at what scale? Is the usage of all platforms that manipulate us forbidden? Would that imply that all systems of recommendation ought to be shut down? May e-commerce platforms use their recommendation systems to alter individual perceptions of reality? Or to push us towards decisions we do not foresee? May streaming platforms use

³The ImageNet database contains pictures that characterise concepts like "Persian cat", "police wagon", or "towel". However, previous to 2020, it also included the visual representation of terms like "orphan", "separatist", or "parricide". The first set of examples represent concepts whose definition responds to the fulfilment of certain objective requirements. For example, a police wagon is recognisable as a vehicle with a police identification, such as the logo of the police force to which it belongs. In contrast, the second set of terms cannot be defined by visually recognisable attributes, but are based on facts that are not recognisable by sight. Assuming that all concepts can be represented visually is a mistake that can perpetuate stereotypes based on physical attributes, such as appearance or skin colour.

their suggestion power to influence our collective imaginary? Note that people's opinion is what matters in market economy and societies structure and many technological initiatives have the power and space to affect our judgement. But we tend only to realize about harms once it is too late. The protection of people's autonomy and respect for it is a collective issue that needs not only a view from the present but also towards the future. We do have an obligation to promote the interests of present and future generations. Securing and preserving people's autonomy is a collective matter.

5. Discussion and conclusions

Technology and ethics cannot advance independently of each other. In a democratic society, progress must go hand in hand with both, involving all stakeholders to find the optimal balance. Separating the ethical framework that governs our society from any technological development moves us away from a reliable AI. This kind of solutions should not have an ethics module built in, but they should nourish from the ethical considerations that derive from each phase of the life cycle of ML solutions.

ML developers must take responsibility for the creation of ML systems and pre-emptively mitigate any potential misuse. A proper understanding of the social and ethical impact of ML systems that may result from each phase of the ML system's life cycle is necessary. In the development of ML-based solutions, an appropriate balance must be found between the exploration and exploitation capabilities of the algorithms to avoid the creation of information bubbles, epistemic bubbles, or echo chambers. Systems must be able to provide all contextual information to foster respect for people's autonomy, from obtaining relevant and complete information to their ability to make decisions. A system that limits people's autonomy will discriminate against them, and if these discriminations are not addressed, they will not be detected nor made visible. Such a system must ensure respect for the autonomy of people not only technically but also socially, i.e., the deployment of surveillance systems ought to be carefully analysed since they indirectly shape the behavior of people, regulating, thus, their degree of autonomy.

The current ethical framework has several shortcomings, of particular relevance being the lack of familiarity of ML developers with this framework. We find ourselves with a legal response that marks a strategic line based on the respect and defence of human rights, but, to date, does not provide a complete and explicit response on how to respect, in particular, the principle of autonomy in ML solutions. This gap opens the door to reflect on what step should be taken now: to seek a consensual ethical basis at the international level or to demand a more complete legal response? In our opinion, both are compatible and necessary but, first, we need to work on a common language based on the respect for human rights that allows us to operationalize the application of the rights, freedoms, and principles that govern our society. Only by respecting and defending human rights will we be able to progress towards a reliable IA. One of the barriers we face is the current tendency to commodify subjects instead of considering this as a collective value, a collective enterprise, etc.

Constant change obliges us to adapt permanently. And we do not want to be left behind because the old has negative connotations. The technological breakthrough of recent years

has changed the way we understand the world around us. Everything happens so fast that it is impossible for us to process all the stimuli around us. We have so much information within reach that we are overwhelmed. As a result, we have become dependent on technology. Since technology does not forget, we do not longer need to dedicate ourselves at 100% to process everything that surrounds us. The immediacy demanded by the present strains us in such a way that we give up part of our autonomy. We must be aware of the implications of losing the balance between autonomy and automation. The current imbalance diminishes our ability to make decisions. Part of this mismatch comes from the misinterpretation that ML is “intelligent enough” to replicate our reasoning. But this does not empower us, on the contrary, it diminishes our autonomy. Our overconfidence in these solutions makes us misunderstand them. Being better informed about how ML solutions work (from how they are created to how they should be used) will enable us to treat them as what they are, mere tools. We must also be aware of what the human-ML interactions are in order to autonomously decide how to use them.

There is always someone who benefits from global imbalances. Being aware of who stands to gain makes us well-informed and enables us to respond if we disagree with the current course of events. Be that as it may, we should not think that algorithms are evil. Algorithms are mechanisms that allow us to automate processes, but what is the purpose of automating processes? To allow us to spend our time on issues that we think deserve more of our attention. So that is how we should use them. Currently, systems are evaluated at the design phase and then they “act freely”. This new way of working entails many risks that we should not normalize nor accept. We should control ML systems, and not allow that ML systems control us. As individuals in a democratic society we have the power to decide how we want to be governed. For this reason, we should have a say on the place of ML in the world and the way we want to interact with ML systems. For this, however, we need to become a literate and informed society capable of making free decisions, i.e., a society formed by autonomous people.

We must address this issue with a transdisciplinary approach that allows us to properly determine how to translate abstract concepts such as autonomy or justice into computer language. We will find that there is never a one-size-fits-all answer, so what is at stake and what is involved in prioritising one thing over another must be carefully studied. In this case, the question that sums up our work is that of finding the right balance between autonomy and automation. In trying to move from abstract to mathematical language, we may be tempted to solve the problem by means of a checklist that simplifies and lists important issues. But ethics cannot be assessed by checklists alone, ethics should not be translated as a mere add-on.

Leaving aside human sovereignty in ML developments can diminish our power. Taken to the extreme, losing our autonomy could cause democracies to falter and eventually collapse. Autonomy, as well as all other rights, freedoms, and ethical principles, are matters of collective interest and should be treated as such.

This research has yielded a number of reflections that open up several lines of work in order to move towards effective respect for the autonomy of people in ML. It is necessary to continue working on the issues expounded in Section 4 in order to eventually produce a fully operational formulation capable of being translated to computer language.

This paper focuses on the principle of autonomy because it is one of the most neglected ethical principles in the eyes of ML, but it is not the only one. Our future work will focus on other principles that have still not been adequately replicated or protected in ML and propose a

formulation of how to operatize them in a technical way.

Acknowledgments

This work is partially supported by MCIN/AEI/10.13039/501100011033 under project PID2019-105093GB-I00.

References

- [1] United Nations, Universal Declaration of Human Rights, 1948.
- [2] S. A. Friedler, C. Scheidegger, S. Venkatasubramanian, The (im)possibility of fairness: Different value systems require different mechanisms for fair decision making, *Communications of the ACM* 64 (2021) 136–143.
- [3] W. Dieterich, C. Mendoza, T. Brennan, COMPAS risk scales: Demonstrating accuracy equity and predictive parity, Northpointe Inc (2016). URL: <https://www.documentcloud.org/documents/2998391-ProPublica-Commentary-Final-070616.html>.
- [4] J. Larson, J. Angwin, Technical response to Northpointe, *ProPublica* 29 (2016). URL: <https://www.propublica.org/article/technical-response-to-northpointe>.
- [5] D. Kahneman, O. Sibony, C. R. Sunstein, *Noise: A flaw in human judgment*, William Collins, Dublin, 2021.
- [6] P. A. Hancock, R. J. Jagacinski, R. Parasuraman, C. D. Wickens, G. F. Wilson, D. B. Kaber, Human-automation interaction research: Past, present, and future, *Ergonomics in Design* 21 (2013) 9–14. URL: <http://erg.sagepub.com/cgi/alerts>. doi:10.1177/1064804613477099.
- [7] R. Parasuraman, C. D. Wickens, Humans: still vital after all these years of automation, *Human factors* 50 (2008) 511–520. URL: <https://pubmed.ncbi.nlm.nih.gov/18689061/>. doi:10.1518/001872008X312198.
- [8] A. Laitinen, O. Sahlgren, AI Systems and Respect for Human Autonomy, *Frontiers in Artificial Intelligence* 4 (2021) 151. doi:10.3389/FRAI.2021.705164/BIBTEX.
- [9] N. Diakopoulos, S. Friedler, M. Arenas, S. Barocas, M. Hay, B. Howe, H. V. Jagadish, K. Unsworth, A. Sahuguet, S. Venkatasubramanian, C. Wilson, C. Yu, B. Zvenbergen, Principles for Accountable Algorithms and a Social Impact Statement for Algorithms, Technical Report, FAT/ML, 2017. URL: <https://www.fatml.org/resources/principles-for-accountable-algorithms>.
- [10] J. Gumbis, V. Bacienskaite, J. Randakeviciute, Do Human Rights Guarantee Autonomy?, *Cuadernos Constitucionales de la Cátedra Fadrique Furió Ceriol* 62 (2008) 77–93. URL: www.un.org.
- [11] M. Foucault, *Discipline & Punish: The Birth of the Prison*, 1975.
- [12] Z. Bauman, *Liquid life, Polity*, 2005.
- [13] E. Morozov, To save everything, click here: The folly of technological solutionism, *Public Affairs*, 2013.
- [14] Y. N. Harari, *21 Lessons for the 21st Century*, Random House, 2018.
- [15] Y. N. Harari, Rebellion of the Hackable Animals, *The Wall Street Journal* (2020).

- [16] I. de Lecuona, La tendencia a la mercantilización de partes del cuerpo humano y de la intimidad en investigación con muestras biológicas y datos (pequeños y masivos), in: Editorial Fontamara (Ed.), De la Solidaridad al Mercado, Edicions de la Universitat de Barcelona, 2016, pp. 267–296. URL: www.bioeticayderecho.ub.edu.
- [17] I. de Lecuona Ramírez, M. Villalobos-Quesada, The value of personal data in the digital society, in: El cuerpo diseminado: estatuto, uso y disposición de los biomateriales humanos, Aranzadi, 2018, pp. 171–191. URL: <https://dialnet.unirioja.es/servlet/articulo?codigo=6499584>.
- [18] I. de Lecuona Ramírez, Ethical, legal and societal issues of the use of artificial intelligence and big data applied to healthcare in a pandemic, Revista Internacional de Pensamiento Político (2020) 139–166. URL: <https://dialnet.unirioja.es/servlet/articulo?codigo=7736125>.
- [19] M. Casado, Los derechos humanos como marco para el Bioderecho y la Bioética, in: Derecho biomédico y bioética, Comares, 1998, pp. 113–136. URL: <https://dialnet-unirioja-es.sire.ub.edu/servlet/articulo?codigo=568994>.
- [20] General Conference of UNESCO, Universal Declaration on Bioethics and Human Rights, 2005. URL: <https://en.unesco.org/themes/ethics-science-and-technology/bioethics-and-human-rights>.
- [21] News European Parliament, Artificial intelligence: the EU needs to act as a global standard-setter, 2022. URL: <https://www.europarl.europa.eu/news/en/press-room/20220318IPR25801/artificial-intelligence-the-eu-needs-to-act-as-a-global-standard-setter>.
- [22] Directorate general of human rights and rule of law, Consultative committee of the convention for the protection of individuals with regard to automatic processing of personal data (Convention 108). Guidelines on artificial intelligence and data protection (2019) 1–4. URL: <https://www.coe.int/en/web/human-rights-rule-of-law/artificial-intelligence/glossary>.
- [23] Council of Europe - Commissioner for Human Rights, Unboxing Artificial Intelligence: 10 steps to protect Human Rights, Technical Report, Council of Europe, 2019.
- [24] COMEST, Preliminary study on the Ethics of Artificial Intelligence, Technical Report, 2019. URL: <https://unesdoc.unesco.org/ark:/48223/pf0000367823>.
- [25] European Union Agency for Fundamental Rights, Getting the future right – Artificial intelligence and fundamental rights, Technical Report, European Union Agency for Fundamental Rights, Luxembourg, 2020. doi:10.2811/58563.
- [26] AI HLEG, Ethics Guidelines for Trustworthy AI, Technical Report, High-Level Expert Group on Artificial Intelligence, Brussels, 2019. URL: <https://ec.europa.eu/futurium/en/ai-alliance-consultation.1.html>.
- [27] T. Metzinger, Ethics washing made in Europe, Der Tagesspiegel (2019). URL: <https://www.tagesspiegel.de/politik/eu-guidelines-ethics-washing-made-in-europe/24195496.html>.
- [28] I. Ben-Israel, J. Cerdio, A. Ema, L. Friedman, M. Ienca, A. Mantelero, E. Matania, C. Muller, H. Shiroyama, E. Vayena, Towards regulation of AI systems. Global perspectives on the development of a legal framework on Artificial Intelligence systems based on the Council of Europe’s standards on human rights, democracy and the rule of law, Technical Report, Council of Europe, Strasbourg Cedex, 2020. URL: <https://rm.coe.int/prems-107320-gbr-2018-compli-cahai-couv-texte-a4-bat-web/1680a0c17a>.

- [29] I. de Lecuona, M. J. Bertrán, B. Bórquez, L. Cabré, M. Casado, M. Corcoy, M. Dobernig, F. Estévez, F. G. López, B. Gómez, C. Humet, L. Jaume-Palasi, E. Lamm, F. Leyton, M. J. L. Baroni, R. L. d. Mántaras, F. Luna, G. Marfany, J. Martínez-Montauti, M. Mautone, I. Melamed, M. Méndez, M. Navarro-Michel, M. J. Plana, N. Riba, G. Rodríguez, R. Rubió, J. Santaló, P. Subías-Beltrán, Guidelines for reviewing health research and innovation projects that use emergent technologies and personal data, *Physical Education and Sport for Children and Youth with Special Needs Researches – Best Practices – Situation* (2020) 343–354. URL: <https://research.wur.nl/en/publications/guidelines-for-reviewing-health-research-and-innovation-projects->. doi:10.2/JQUERY.MIN.JS.
- [30] European Commission, Proposal for a Regulation of the European Parliament and of the Council Laying Down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act) and Amending Certain Union Legislative Acts, Technical Report, European Commission, Brussels, 2021.
- [31] L. Edwards, Regulating AI in Europe: four problems and four solutions, Technical Report, Ada Lovelace Institute, 2022. URL: <https://www.adalovelaceinstitute.org/report/regulating-ai-in-europe>.
- [32] European Digital Rights, Access Now, Panoptykon Foundation, epicenter.works, AlgorithmWatch, European Disability Forum, Bits of Freedom, Fair Trials, PICUM, ANEC, ANEC, An EU Artificial Intelligence Act for Fundamental Rights. A Civil Society Statement, Technical Report, European Digital Rights, 2021.
- [33] Autoriteit Persoonsgegevens, Boete Belastingdienst voor zwarte lijst FSVPersoonsgegevens, Technical Report, Autoriteit Persoonsgegevens, 2022. URL: <https://autoriteitpersoonsgegevens.nl/nl/nieuws/boete-belastingdienst-voor-zwarte-lijst-fsv>.
- [34] Commission Nationale de l'Informatique et des Libertés, Cookies: GOOGLE fined 150 million euros, Technical Report, Commission Nationale de l'Informatique et des Libertés, Paris, 2022. URL: <https://www.cnil.fr/en/cookies-google-fined-150-million-euros>.
- [35] Garante per la protezione dei dati personali, Ordinanza ingiunzione nei confronti di Enel Energia S.p.a. - 16 dicembre 2021 [9735672], Technical Report, Garante per la protezione dei dati personali, Rome, 2021. URL: <https://www.garanteprivacy.it/web/guest/home/docweb/-/docweb-display/docweb/9735672>.
- [36] L. Floridi, Soft ethics, the governance of the digital and the General Data Protection Regulation, *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* 376 (2018). doi:10.1098/rsta.2018.0081.
- [37] E. Pariser, *The Filter Bubble: What the Internet Is Hiding from You*, Penguin Press, New York, 2011.
- [38] A. Tversky, D. Kahneman, Judgment under Uncertainty: Heuristics and Biases, *Science* 185 (1974) 1124–1131. URL: <https://pubmed.ncbi.nlm.nih.gov/17835457/>. doi:10.1126/SCIENCE.185.4157.1124.
- [39] R. Boudon, Beyond rational choice theory, *Annual review of sociology* 29 (2003) 1–21. URL: <https://www.annualreviews.org/doi/abs/10.1146/annurev.soc.29.010202.100213>. doi:10.1146/ANNUREV.SOC.29.010202.100213.
- [40] J. von Neumann, O. Morgenstern, *Theory of games and economic behavior*, Princeton University Press, 2007. doi:10.1515/9781400829460.

- [41] A. Tversky, A critique of expected utility theory: Descriptive and normative considerations, *Erkenntnis* (1975) 163–173. URL: <https://www.jstor.org/stable/20010465>.
- [42] D. Kahneman, A. Tversky, Prospect theory: An analysis of decision under risk, *Econometrica* 47 (1979) 263–292. doi:10.2307/1914185.
- [43] T. Misra, The Tenants Fighting Back Against Facial Recognition Technology, Bloomberg CityLab (2019). URL: <https://www.bloomberg.com/news/articles/2019-05-07/when-facial-recognition-tech-comes-to-housing>.
- [44] J. Sánchez-Monedero, L. Dencik, The politics of deceptive borders: 'biomarkers of deceit' and the case of iBorderCtrl, *Information, Communication & Society* (2020) 1–18. URL: https://www.researchgate.net/publication/337438212_The_politics_of_deceptive_borders_%27biomarkers_of_deceit%27_and_the_case_of_iBorderCtrl.
- [45] R. Gimeno, Los algoritmos tienen prejuicios: ellos son informáticos y ellas, amas de casa, 2017. URL: https://elpais.com/retina/2017/05/12/tendencias/1494612619_910023.html?rel=buscador_noticias.
- [46] M. Echarri, 150 despidos en un segundo: así funcionan los algoritmos que deciden a quién echar del trabajo, 2021. URL: <https://elpais.com/>.
- [47] D. Jemio, A. Hagerty, F. Aranda, The Case of the Creepy Algorithm That 'Predicted' Teen Pregnancy, *Wired* (2022). URL: <https://www.wired.com/story/argentina-algorithms-pregnancy-prediction/>.
- [48] R. Parasuraman, V. Riley, Humans and automation: Use, misuse, disuse, abuse, *Human factors* 39 (1997) 230–253. URL: <https://journals.sagepub.com/doi/abs/10.1518/001872097778543886>. doi:10.1518/001872097778543886.
- [49] K. Yang, K. Qinami, L. Fei-Fei, J. Deng, O. Russakovsky, Towards fairer datasets: Filtering and balancing the distribution of the people subtree in the imagenet hierarchy, in: *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 2020, pp. 547–558. URL: <https://www.image-net.org/filtering-and-balancing/>.
- [50] K. Crawford, Atlas of AI, 2021. URL: <https://www.katecrawford.net/index.html>.
- [51] V. Eubanks, Automating inequality: How high-tech tools profile, police, and punish the poor, St. Martin's Press, 2018.
- [52] J. Bryson, AI & global governance: no one should trust AI, *United Nations University* (2018). URL: <https://cpr.unu.edu/publications/articles/ai-global-governance-no-one-should-trust-ai.html>.
- [53] M. Ryan, In AI we trust: ethics, artificial intelligence, and reliability, *Science and Engineering Ethics* 26 (2020) 2749–2767. doi:10.1007/S11948-020-00228-Y.
- [54] P. Robinette, W. Li, R. Allen, A. M. Howard, A. R. Wagner, Overtrust of Robots in Emergency Evacuation Scenarios, in: *2016 11th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, IEEE, 2016, pp. 101–108.