Societal Implications of Recommendation Systems: A Technical Perspective

Joana Gonçalves-Sá and Flávio Pinheiro

Abstract One of the most popular applications of artificial intelligence algorithms is in recommendation systems (RS). These take advantage of large amounts of user data to learn from the past to help us identify patterns, segment user profiles, predict users' behaviors and preferences. The algorithmic architecture of RS has been so successful that it has been co-opted in many contexts, from human resources teams, trying to select top candidates, to medical researchers, wanting to identify drug targets. Although the increasing use of AI can provide great benefits, it represents a shift in our interaction with data and machines that also entails fundamental social threats. These can derive from technological or implementation mistakes but also from profound changes in decision-making.

Here, we overview some of those risks including ethical and privacy challenges from a technical perspective. We discuss two particularly relevant cases: (1) RS that fail to work as intended and its possible unwanted consequences; (2) RS that work but at the possible expense of threats to individuals and even to democratic societies. Finally, we propose a way forward through a simple checklist that can be used to improve the transparency and accountability of AI algorithms.

1 Introduction

Much like the previous Industrial Revolutions, the Digital Revolution is sure to have enormous impact on society, at many different levels. By learning from the unparalleled amounts of individual-level data that is currently shared and collected, machines will be increasingly able to identify patterns, create profiles, predict

J. Gonçalves-Sá (✉)
LIP - Laboratório de Instrumentação e Física Experimental de Partículas, Lisbon, Portugal
e-mail: joanagsa@lip.pt

F. Pinheiro
IMS - Information Management School, Universidade Nova de Lisboa, Lisbon, Portugal
e-mail: fpinheiro@novaims.unl.pt

H. Sousa Antunes et al. (eds.), *Multidisciplinary Perspectives on Artificial Intelligence and the Law*, Law, Governance and Technology Series 58,
https://doi.org/10.1007/978-3-031-41264-6_3

behaviors, and make decisions. Therefore, it is fundamental to understand the limitations of these tools to anticipate and minimize negative consequences.

In this chapter we focus on machine-learning (ML) models, particularly recommendation (or recommender) systems (RS), and how their use in decision-making processes can offer better services but also create important risks. Typically, RS refer to algorithms that recommend some item X to a user A, very often consumption goods (such as recommending a book the algorithm identifies as matching our interests) or in the context of social networks (a new friend or post); however, here we use this term in a broader sense, to refer to any algorithm that uses large datasets on people to identify similarity matches and recommend decisions, in many different contexts. First, we describe how RS work and their unavoidable limitations. Second, we focus on RS that work as intended and discuss how the creation of individual profiles can lead to abusive targeted advertisement and even to threats to democracy, from disinformation to state surveillance. Third, we describe what happens when these systems are faulty, but are still used to make probabilistic generalizations and aid in AI-based decision making. We will offer specific examples of how mistakes in data selection or coding might lead to discrimination and injustice. In the last section, we summarize some ideas on how to make AI more accountable and transparent and argue that the important decisions ahead should not be made by a limited group of non-elected AI leaders, but it should be the role of AI experts to raise awareness of such threats, paving the way for important regulatory decisions.

2 Recommendation Systems

The general goal of a recommendation systems is to predict, as accurately as possible, a new item to a user while optimizing for the rate of acceptance (Resnick and Varian 1997). These systems leverage information on users (demographic, past choices) and/or on items (for example, movies) to find accurate matches between them (ex: if you liked movie X you might like movie Y, or people "like you" have enjoyed book Z). At the core of RS is the assumption that items and/or users of a service can be mapped in terms of their similarities and that person A (or item X) can serve as proxy for person B (item Y). In that sense, a recommendation system suggests items that are closer in such similarity space to a user's past choices or revealed preferences and is only as good as it can provide the most accurate recommendation to a user (person A will actually enjoy movie Y and book Z).

Over the past decades we have seen an increase in use of ML/AI techniques to support the development and implementation of faster, more reliable, and more capable RS (Fayyaz et al. 2020). These are possible because of (a) large accumulation of data about users' past choices; (b) large datasets on details about items, and (c) increasingly sophisticated algorithms that take advantage of such data and take value from growing numbers of features and instances.

In general, recommendation systems can be divided into three big families: collaborative-base filtering (that tries to predict whether person A will like product X based on the preferences of "similar-person" B); content-based filtering (that tries to predict whether person A will like an item X based on person A's past revealed preferences for similar items); and hybrid systems, that combine both. In terms of the algorithms used, these are further divided on whether they are supported by heuristic- or model-based approaches.

Collaborative filtering (Herlocker et al. 2000) recommendation systems rely on the similarity between users to perform recommendations. That is, if user A and B are similar, then the past choices of B can shed light on what to recommend to user A. Hence, the core technical challenge is to estimate similarities between users or items from data on the revealed past preferences of users (ex. past favorite movies). This approach has been widely popular on web-based portals such as Netflix and Reddit where users' characteristics and up or down votes are used to estimate similarities. Popular algorithms range from Graph models of social networks (Bellogin and Parapar 2012) of similarity between users and Nearest Neighbor to the use of Linear regressions, Clustering techniques (Ungar and Foster 1998), Artificial Neural Networks (He et al. 2017) and Bayesian networks.

Often, auxiliary data is used to either improve collaborative filtering systems, or to overcome some of its limitations. Context information (e.g., location or time) can help systems achieve higher success and, in scenarios that have more users than items, recommendations are often done through an item-item similarity. Moreover, when past information on user activity is scarce (e.g., in the case of a new user of a new service), users' information about their social relationships and characteristics (e.g., gender, age, income, location, employment, etc.) can help these systems establish a similarity even without specific historical activity.

Content-based filtering (Lops et al. 2011; Aggarwal 2016) does not require information about users, instead it maps similarity between items to perform recommendations. In other words, users are recommended items similar to their past choices (person A likes vanilla milkshakes, thus might also like vanilla ice-cream). Algorithmically, these problems are approached using techniques that range from TF-IDF (Rigutini and Maggini 2004) and clustering for topic modelling and inference, but also using classification models based on Bayesian classifiers, Decision trees, and Artificial Neural Networks. A traditional application of these techniques is in book recommendation engines that measure content (Mooney and Roy 2000) similarity between books.

Hybrid solutions (Burke 2007) combine aspects of both content and collaborative filtering. They arise in situations where it is practical and beneficial to develop meta-algorithms to balance the recommendation stemming from a collaborative- and content-based systems. These increasingly use complex deep learning algorithms and are common in social media recommendations, including newsfeed content, advertisements, and friends (Naumov et al. 2019).

There are several limitations and problems associated with the development of RS. From the technical perspective, problems can arise at two extremes of the spectrum. First, lack of initial data can lead to a "cold start", that prevents

the setup of the entire recommender system (ex. new movies that have not yet been rated by anyone or are from little known studios or directors) or limit the recommendations that can be given to new users. Second, when there is a "sparsity problem" and the number of items to be recommended is very large, the algorithm might lack scalability and users keep seeing the same few recommendations, either because they are the few most rated or because the individual users only rated a few (Adomavicius and Tuzhilin 2005). Third, and more conceptually, the implicit assumption in ML/AI solutions that the future can be predicted from past actions, renders them awkwardly unable to perform under novelty (e.g., expanding a service to a new cultural setting). Fourth, some models described above can learn from past mistakes (user A hated book Z after all) and, therefore, improve continuously, but this is not the case in several other examples of RS, sometimes with dire consequences. Important examples of the impact of the listed limitations will be discussed in more detail in the following sections.

3 When Recommendation Systems Work

3.1 Implications for Consumption

Although stemming from a seemingly intuitive and simple problem, recommendation systems have matured to highly complex algorithmic solutions that are able to leverage a multitude of data sources to improve services that underlie the success of some of the largest companies in the world. The success of RS, and their ubiquity, stems from their capability to enhance user retention and to seemingly help users find the relevant content for their profile. Moreover, it is already possible to extract information and patterns from both structured information (e.g. online shopping basket) and unstructured information (e.g., free text, images, and videos). As such, RS are improving faster and will offer more gains to content providers.

Naturally, the specifics of each system are largely dependent on their application context and goals. Take, for instance, Amazon which started using item-to-item approaches but currently leverages information from users' past orders, profile, and activity, to offer different types of personalized recommendations, from targeted e-mails to shopping recommendations (Smith and Linden 2017). In 2018, the consultancy company McKinsey estimated that Amazon's RS was responsible for 35% of its sales (MacKenzie et al. 2013). Netflix gained international fame among engineers and enthusiasts with the release, in 2006, of a dataset of 100 million users' movie ratings and offered a 1Million USD reward to the team that could develop the best RS. Ten years later, Netflix RS was estimated to be worth up to 1 billion USD (McAlone 2016) and to drive 75% of users' viewing choices (Vanderbilt 2013).

However, these large companies depend on using data freely and often willingly shared by their users, who might give away control of their privacy and decisions in exchange for convenience and productivity. In fact, we all know that our data

is being used, but we may not know the extent to which this is happening or the problems it could pose (Englehardt and Narayanan 2016). First, and although these processes involve consent, terms of service are often unintelligible, sharing is not always voluntary, and might be a requirement to access content or services (Solomos et al. 2019; Urban et al. 2020). Second, and even when it is voluntary, it can have unexpected implications. In 2012, the New York Times reported a case in which the US-based chain Target generated predictions about the pregnancy of its customers, precisely by analyzing shopping profiles (Duhigg 2012). One such store was visited by a father, outraged that his teenage daughter had received promotions for baby products; later, when the store manager called to apologize, the man embarrassedly replied that his daughter was indeed pregnant: the supermarket chain knew before the family. Such anecdotical situations, corroborate our increasing reliance on such systems, which also makes us vulnerable to manipulation. For instance, nothing keeps online stores from showing more expensive products to people who did not previously compare prices online (Mikians et al. 2012). Indeed, different webservices commonly trade user information for marketing purposes, and it is common for a user that searches jeans on Amazon to be immediately targeted with jeans' ads on Facebook[1] or Google.[2] Importantly, these "surveillance systems" are so prevalent and increasingly sophisticated that even when you use caution when publishing online, that caution itself can be informative (Zuboff 2019).

As mentioned, the traditional application of RS is to drive the consumption of content and products and, as such, it represents the most common development of such algorithms (similarity identification, reliance on proxies, prediction of future outcomes). However, they also find application in other types of algorithmic decision-making (e.g., credit score, or financial trading) and we will use them in a broader context to further discuss their current implications.

3.2 Implications for Democracy

As described, RS can be very useful to direct people to products that interest them, be it movies or diapers. But there is a thin line between informing and manipulating, and this is particularly relevant when the promoted "goods" are news or ideas. Social networks, such as Facebook or Instagram, have been long known to promote addictive attention, even if at the cost of spreading disinformation (Del Vicario et al. 2016; Vosoughi et al. 2018), creating echo chambers (Nikolov et al. 2015; Quattrociocchi et al. 2016), increasing polarization (Flaxman et al. 2016), and

[1] See for example: "Help your ads reach the people who will love your business", by Facebook, 2021. https://pt-pt.facebook.com/business/ads/ad-targeting.

[2] See for example: "What are retargeting ads?", by Google Ads, 2021. https://ads.google.com/intl/en_uk/home/resources/retargeting-ads/.

threatening the user's mental health.[3] From researchers to data protection advocates, many have voiced concerns about the data that large platforms collect and how their recommendation systems can manipulate the information individuals are exposed to, be them prioritizing posts, or search engines displaying sponsored adds. In fact, Facebook offers any interested add-placer the possibility of selecting over tenths of individual characteristics, including (estimated) age or gender, level of education and in which subject, income-level, hobbies, political orientation (if from the US), travel profile, and even whether targets are away for the weekend with family or friends (Haidt and Twenge 2021).

The Cambridge Analytica case, in early 2018, brought to the public spotlight how, through refined individual profiling, political campaigns could influence the voting of target individuals or constituencies.[4] Political scientists have argued that the use of modern data science approaches to politics represented a significant shift from classical strategies: marketing techniques have been used in politics since at least the 1930s (O'Shaughnessy 1990), but the speed and increasing precision of AI tools means that political messages no longer need to be general and appeal to a broad constituency; instead, they can send highly personalized messages based on individual profiles, saying one thing to one demographic and the opposite to another, with very little scrutiny (Aldrich et al. 2015; Ribeiro et al. 2019; Silva et al. 2020). That some of these messages can include untrue information is even more worrisome. Naturally, the political use of misleading and even outright false information is nothing new, but the surge in online activity, coupled with poor digital literacy, and individual-level consumer profiling, has all the ingredients of a perfect storm. Disinformation spreading has found fertile ground on social networks, often through emotion manipulation, first shown to occur by Facebook itself (Kramer et al. 2014), and to work not only in the targeted individuals but also to contaminate their friends (Coviello et al. 2014). There is also increasing evidence that some individuals might be more susceptible to political disinformation than others (Pennycook and Rand 2021), with specific cognitive bias playing important roles.

In fact, personalized algorithms on search engines and social networks feeds might strengthen these already existing biases in at least three different ways: (1) as information is filtered based on past history (potentially magnifying availability biases, in which individuals tend to rate as more important things that they can more easily recall (Abbey 2018), and confirmatory tendencies, in which individuals seek or particularly trust in information that re-enforces or confirms their beliefs (Burtt 1939)); (2) as humans tend to associate with others similar to them and to favor people in one's own group, over people identified as belonging to outgroups

[3] This is a fast-growing field, with exposés becoming increasingly frequent. On mental health impacts there is an excellent ongoing open-source literature review posted and curated by Jonathan Haidt (NYU-Stern) and Jean Twenge (San Diego State U), Haidt and Twenge (2021); (Zuboff 2019).

[4] Cadwalladr and Graham-Harrison (2018).

(ingroup bias) (Nelson 1989); and (3) with beliefs, biases, and even disinformation (McPherson et al. 2001) amplified and reinforced by this closed, homophilic communities, leading to the already mentioned echo chambers (Barberá et al. 2015; Flaxman et al. 2016; Quattrociocchi et al. 2016) and to increased online hostility and polarization (Yardi and Boyd 2010; Conover et al. 2021).

But political (mis)information is not the only kind to heavily impact on society and democracy. In April 2020, Facebook acknowledged that millions of its users saw false COVID19-related information on this platform (Ricard and Medeiros 2020); On Twitter, according to Yang et al. (2020), "the combined volume of tweets linking to low-credibility information was comparable to the volume of New York Times articles and CDC links"; by August 2021, YouTube had removed 1 million videos that included dangerous COVID-19 misinformation. Importantly, there is evidence that such misinformation impacted vaccination hesitancy (Loomba et al. 2021) and compliance with control measures (Roozenbeek et al. 2020), in line with the notion that misinformation often serves the goal of creating divisive content and leading to social unrest (Emmott 2020; Ricard and Medeiros 2020; Barnard et al. 2021; Silva and Benevenuto 2021). For much of 2020 and 2021, the world was fighting two pandemics in parallel: one caused by a virus, and another caused by fake news, supported by human bias and attention maximizing algorithms (Goncalves-Sa 2020).

Another very relevant risk comes from societal control. As described, politicians can use social networks and AI systems to target possible voters, but several leaders have also realized the much broader potential of AI, from improving public administration, to creating war robots. According to Vladimir Putin: "Artificial intelligence is the future, not only of Russia, but of all of mankind (. . .) Whoever becomes the leader in this sphere will become the ruler of the world." (Allen 2017) China hopes to be the leader by 2030 (Department of International Cooperation Ministry of Science and Technology (MOST) 2017) and is designing and implementing a large-scale social experiment, which involves using RS to classifying citizens according to their social behavior, only possible thanks to AI-driven facial recognition technology (Liang et al. 2018). These models have been increasingly used around the world,[5] often with security purposes. In 2020, the Israeli and US armies used AI to track and assassinate an Iranian physicist (Bergman and Fassihi 2021).

All these examples describe situations in which the RS are worrying because they work as intended, be it to improve consumption or to target voters. In the next chapter, we will focus on situations in which they fail and how that can have consequences for individuals and societies.

[5] For a collection of countries that legalized or are using facial recognition tools see for example https://surfshark.com/facial-recognition-map.

4 When Recommendation Systems Fail

The described recommendation systems use fine-grained information to train AI models to target specific individuals. Typically, what these systems do is output probabilities of a certain event and aid in decision-making. Examples can range from algorithms that calculate a risk score for depression (Reece et al. 2017; Eichstaedt et al. 2018), try to identify the best candidate for a given position (Paparrizos et al. 2011), or that recommend a movie based on previous choices (Bennett and Lanning 2007). These algorithms are trained on large training datasets, of variable quality, and "learn" by trial and error, with subjective definitions of error (for example, what "best candidate" means, must be represented as a mathematical object, when often "best" cannot be easily quantified). This means that there is no real distinction between the model, the data used to train it, and the assumptions that the coder made: if the data or the target are biased, the model will be biased. These bias might appear at different steps and have different consequences, but it is important to realize that: (1) it is virtually impossible to have a complete dataset and all datasets are samples, biased by the sampling process; (2) there are human decisions involved in defining targets; (3) targets often rely on proxies, (4) the predictions might turn into self-fulfilling prophecies because they frequently impact outcomes and it is often very difficult to have external validation. Again, this might be of little importance in the case of a user who never gets to seem movie X because it is not suggested, but very serious in the case of someone who gets a credit request denied and, consequentially, defaults on another payment: the system might find confirmation that the credit refusal was the best decision when indeed, it was what caused the default.

Consequently, there should be no illusions of "model neutrality". All models have problems, and acknowledging it is a fundamental and essential step to design mitigation strategies. In this section, we describe how biased data leads to biased algorithms, how biased algorithms can lead to discriminatory policies, and offer some examples from both the private and public sectors.

4.1 Learning from Biased Data: Implications for Individuals

As there is no perfect dataset, it is important to understand its limitations when training any algorithm with it. Let us think of a model to identify the best candidates to enter engineering school. One would start by collecting vast amounts of data, including grades, happiness scores, time to degree completion, previous education, future career, etc., on all students who have ever gone through a given university. This dataset would still have no information on how good the rejected candidates could have become (sampling bias) or on how many of them eventually suffered from burnout (limitations and subjectivity in feature selection): this means, that if the system systematically rejected promising candidates in the past, the algorithm is

very likely to continue doing so in the future; and that if, for example, it values prizes over creating a safe work environment, it might pave the way for more accidents in the future. It is also easy to anticipate that such a dataset would be unbalanced in terms of gender and likely also age, ethnicity, nationality, and probability of wearing glasses, and so would the model predictions. In fact, several previous attempts at training such algorithms to select applicants, for schools or jobs, have led to discriminatory practices, stirring large discussions (O'Neil 2016). It is important to note that, very often, these algorithms are created not just for speeding up and automating processes, but also because we know that human-based systems are biased: the assumption is that models would be blind to color or gender and, thus, fairer. However, RS trained with biased data will generally be biased as well (Garcia 2016), and this is true even if models are trained on very large datasets. For example, the increasingly popular Chat GPT application was trained using 570 Gb of data, but most of this data was obtained through the internet, which is known to have an overrepresentation of some countries and age groups (Sheng et al. 2021).

An area in which such discrimination can have dire consequences is health. Kadambi (2021) have crucial sources of bias in medical devices, including computational bias, which happens when datasets used in clinical trials or when training algorithms to select candidates for such clinical trials, are biased. Historically, this has been the case for specific ethnical groups and women, often underrepresented in health datasets (and even in experimental protocols).

Biased datasets have also been shown to play important roles in classification and facial recognition (Buolamwini and Gebru 2018; Barlas et al. 2019). For example, Twitter dropped its picture cropping algorithm after suspicions of racial bias (Agrawal and Davis 2021) and both Flickr and Google algorithms tagged photos of black individuals as apes (Zhang 2015).

As disastrous as these examples are, it can be argued that they are the price to pay for the learning process: they are precautionary tales, reminding us that we are still at the infancy of machine decision-making and many other mistakes will be made before we can rely on algorithms. Unfortunately, and despite their current limitations, many are already being deployed, including in punitive environments, as described in the next section.

4.2 From Bad Algorithms to Discriminatory Policies

The individual consequences of a faulty Netflix algorithm are probably easy to minimize; models that select candidates for a given job can have much worse consequences, but nothing compares to when such algorithms are deployed in a large-scale punitive context. We already mentioned how the Chinese government is using facial recognition and other AI tools to evaluate citizens according to their behavior. If the system is faulty, the consequences for the individuals can be tremendous.

Another very debated example, that relies on proxies, is COMPAS, a proprietary algorithm that helps US judges set bails based on estimated risk scores of future offenses. In 2016, COMPAS was analyzed by ProPublica (Larson et al. 2016) and revealed to discriminate individuals based on their race: for similar offenses and crimes, black defendants were more likely to be given higher risk scores. Importantly, the datasets that were used to train the model did not include information on race: the model was possibly using zip code as a proxy for risk, thus picking it as a proxy the correlation between the ethnicity and economic status in US society (this analysis is disputed by the owning company and the extent of the discrimination is still being debated (Spielkamp 2017)).

Despite so many notable failures, governments around the world have been sponsoring the development of algorithms for use in public administration, in a variety of areas. These algorithms are often proprietary and function as a blackbox: not even the government officials know how they work and what justifies their recommendations, or risk scores. This leaves very little room for people to complain or even understand their "evaluation", raising fundamental legal questions. The Dutch government used such an algorithm, SyRI, from 2014 to 2020, when the Court of the Hague halted its use (Amnesty International 2021). It aimed at identifying social welfare fraud, was trained on large governmental datasets, and included information on virtually all inhabitants of The Netherlands. It would generate risks-scores and, if these were high, trigger an investigation. However, it was shown that the algorithm disproportionately and unfairly targeted poor and minority communities (Xenophobic Machines 2021), with consequences so dire that it led to the resignation of the Dutch government.

Such faulty algorithms have been increasingly revealed (Bandy 2021) but, naturally, it can be argued that they only reveal past and pre-existing bias, hidden in the data, and that human decision-making is equally discriminatory. While the first contention is very likely true, it still raises the important question of whether it is acceptable to perpetuate such discriminatory practices under an illusion of mathematical neutrality. The second is more interesting, as it is difficult to quantify whether humans or current algorithms are more discriminatory (Dressel and Farid 2018), but at least in the case of the examples described here, there are at least two good arguments in favor of the later. One is technical, as AI models identify dominant patterns and are more likely to exclude relevant outsiders (for example, the brilliant candidate from a very poor, black neighborhood). The other is scale, as human panels might have their own biases, but these might be different from panel to panel and there are human limits to how many applicants a panel can see; obviously, these limits and natural variation do not necessarily apply to machine decision-making (O'Neil 2016). A third, less studied possibility, is that algorithms, including commercial-like cookies, might be used to mask deliberate targeting of individuals by state actors, as in the case of the identification of minorities (Borgesius 2018). Therefore, there are serious concerns that algorithms trained on biased datasets will not only make biased decisions, they will also amplify existing societal discrimination and unfairness.

5 A Way Forward

There is much room for improvement of current and future RS (and AI in general), and we propose six steps, summarized in a simple mnemonic: (ATI) (Fayyaz et al. 2020). The first is recognizing that they are not neutral and can be very prone to bias. This **Acceptance** should be obvious, but it is still disputed by several in the field, typically contesting that they (a) are not biased as algorithms are blind to individuals, (b) are not more biased than non-algorithmic systems, or (c) that this a problem for social scientists and that engineers and programmers should not be concerned with such issues. In fact, most Data Science and Artificial Intelligence graduate programs still do not include Ethics or even Algorithmic Fairness courses, effectively training generations of students to ignore fundamental problems with datasets, algorithms and, consequentially, recommendation and decision systems they design and often implement. Such content should be compulsory in all formal AI education, taking us to the second step—**Training**.

Another fundamental issue is lack of **Inclusion** and Diversity. This is observed not only on the training datasets, as already discussed, but also in the coding teams. In "Racist in the Machine" (Reece et al. 2017), Megan Garcia describes some grave consequences of design blind spots and gives the example of four smartphone personal assistants (Siri, Google Now, Cortana, and S Voice), increasingly used for help in health and emergency situations, that could not recognize "I am being abused" or "I was beaten up by my husband". ML teams should be diverse and bring together people that work on different disciplines and that can contribute to both the technical and social components of algorithm design. Moreover, that algorithms try to find similarities can lead to polarization and homophily, but also to uniformization. As Thomas Homer-Dixon put it, "a simplified, uniform global culture will inevitably have less diversity of ideas and ingenuity that can help us cope with the great challenges we're facing" (Homer-Dixon 2001). Diversity should be a value at many different levels.

RS pipelines should also include streamlined Data **Auditing** and Debiasing: accepting it as an integral part of data processing pipelines, recognizes its importance for fair and effective AI, while reducing dataset bias. One of the first such efforts was developed by Pedro Saleiro and Rayid Ghani, through a data auditing algorithm, *Aequitas*, that inspects datasets for different types of demographic unbalances, including age, gender, and race (Saleiro et al. 2018). In all three datasets analyzed, they found important bias that affects the models results. These are excellent first efforts, but it is important to note that (a) we can only audit data in a very limited number of instances, and (b) that debiasing is even more challenging. For example, we can check gender-classified datasets for unbalances in gender (as in the hiring example described above), but this might be impossible to do in fully anonymized datasets or in datasets that simply do not include possibly relevant data as is often the case for ethnicity or physical disabilities. Even more critical, we cannot identify biases that we do not know we have, as a society: it might be the case that people with glasses are perceived as more competent for

some jobs; as we are unaware of it, we would not include "having glasses" as a label and, even if we did, we would most likely not audit our algorithm for possible discrimination. But let us assume that complete auditing was possible and that all possible discriminatory imbalances in our dataset had been identified: we would still have important decisions to make regarding how to de-bias them. Continuing with the college admission example, it should be possible to understand that the model is being trained on a gender unbalanced dataset and that correcting for it would now lead to more women candidates being selected. But how big should the correction be? Should it reflect past ratios of engineering school admissions, perpetuating existing imbalances or should it aim for the same ratios observed in the population, effectively imposing a 50% gender quota? These and other example illustrate how many of these decisions can and are effectively being made, often implicitly, and how ill-informed attempts to correct bias might generate new forms of unfairness.

These decisions are fundamentally moral, helping to create a society by design. Thus, the final step should be **Transparency**. As Rhema Vaithianathan put it, "If you can't be right, be honest" (Courtland 2018). Blackbox algorithms, in which the process and features used to reach a decision are unknown or proprietary, should be avoided. However, they are increasingly used for two main reasons: first, it can be argued that if the decision process is known, individuals and companies could abuse and even rig the system on their behalf; second, the more complex the algorithms, as is the case with deep learning, the more difficult it is to understand the decision-making process. Therefore, it has been argued that such algorithms should only be used in positive environments and when they significant outperform traditional processes (e.g. for medical diagnosis), and never in punitive contexts (such as in the COMPAS and SyRI examples). In any case, individuals should always have the right to access, verify, correct errors, and appeal from algorithmic decisions. As these processes are often very complex, this generates an important tension, extensively noticed by the thinkers of the so-called "Risk Society" (Beck 1992), in which technical expertise is fundamental to design and control such systems, but this control should be put into effect by the, often lay, society. Therefore, the ones who understand the problems should also accept their political and social responsibility and engage in active **Interaction** with communities and decision-makers.

6 Conclusions

It should be increasingly obvious that using machine-based decisions is far from neutral, and that its problems have important societal implications. In this chapter, we summarized some limitations of recommendations systems, from both technical and conceptual perspectives, and offered examples of its past, ongoing, and possible future negative impacts. Overall, we argue that these risks should be understood

by the general population, and we offer specific guidelines for improving RS and societal oversight.[6]

Acknowledgements We thank members of the Social Physics and Complexity Lab (SPAC-LIP) for comments and suggestions. Some of the examples described here had previously been presented in a series of 10 articles edited by one of the authors for the Portuguese newspaper Público: https://www.publico.pt/os-riscos-da-revolucao-digital. This work was partially supported by ERC-STG 853566 – FARE to JGS.

References

Abbey R (2018) #Republic: divided democracy in the age of social media, by Cass R. Sunstein. Am Polit Thought 7:370–373

Adomavicius G, Tuzhilin A (2005) Toward the next generation of recommender systems: a survey of the state-of-the-art and possible extensions. IEEE Trans Knowl Data Eng 17:734–749

Aggarwal CC (2016) Content-based recommender systems. In: Aggarwal CC (ed) Recommender systems: the textbook. Springer International Publishing, Cham, pp 139–166

Agrawal P, Davis D (2021) Transparency around image cropping and changes to come. https://blog.twitter.com/en_us/topics/product/2020/transparency-image-cropping. Accessed 28 Jan 2022

Aldrich JH, Gibson RK, Cantijoch M, Konitzer T (2015) Getting out the vote in the social media era: are digital tools changing the extent, nature and impact of party contacting in elections? Party Polit 22:165–178

Allen GC (2017) Putin and Musk are right: Whoever masters AI will run the world. https://www.cnn.com/2017/09/05/opinions/russia-weaponize-ai-opinion-allen/index.html. Accessed 28 Jan 2022

Amnesty International (2021) Discrimination through unregulated use of algorithms in the Dutch childcare benefits scandal. https://www.amnesty.org/en/wp-content/uploads/2021/10/EUR3546862021ENGLISH.pdf. Accessed 28 Jan 2022

Bandy J (2021) Problematic machine behavior: a systematic literature review of algorithm audits. Proc ACM Hum-Comput Interact 5:Article 74

Barberá P, Jost JT, Nagler J, Tucker JA, Bonneau R (2015) Tweeting from left to right: is online political communication more than an echo chamber? Psychol Sci 26:1531–1542

Barlas P, Kyriakou K, Kleanthous S, Otterbacher J (2019) Social B(eye)as: human and machine descriptions of people images. In: Dataset Papers of the Thirteen International AAAI conference on web and social media, Munich, Germany, 11—14 June 2019

[6] *See* generally, on the different applications of Machine Learning and AI, in this book A Oliveira and M A T Figueiredo - Artificial intelligence - historical context and state of the art; I Trancoso, N Mamede, B Martins, H S Pinto and R Ribeiro - The impact of language technologies in the legal domain; A T Freitas - Data-driven approaches in healthcare - challenges and emerging trends; M Correia and L Rodrigues - Security and Privacy; E Magrani and P G F Silva - The Ethical and Legal Challenges of Recommender Systems Driven by Artificial Intelligence; M Lanz and S Mijic - Risks associated with the use of natural language generation - Swiss civil liability law perspective; M S Fernandes and J R Goldim - Artificial Intelligence and Decision Making in Health - Risks and Opportunities; W Gravett - Judicial Decision-making in the Age of Artificial Intelligence; and D Durães, P M Freitas and P Novais - The Relevance of Deepfakes in the Administration of Criminal Justice.

Barnard M, Iyer R, Del Valle SY, Daughton AR (2021) Impact of COVID-19 policies and misinformation on social unrest. arXiv preprint arXiv:2110.09234
Beck PU (1992) Risk society: towards a new modernity. Sage Publications, London
Bellogin A, Parapar J (2012) Using graph partitioning techniques for neighbour selection in user-based collaborative filtering. In: Proceedings of the sixth ACM conference on recommender systems. Association for Computing Machinery, Dublin, Ireland, pp 213–216
Bennett J, Lanning S (2007) The netflix prize. In: Proceedings of KDD Cup and Workshop 2007. ACM, San Jose, CA, pp 3–6
Bergman R, Fassihi F (2021) The scientist and the A.I.-assisted, Remote-Control killing machine, The New York Times. Accessed 28 Jan 2022
Borgesius FJZ (2018) Discrimination, artificial intelligence, and algorithmic decision-making. Directorate General of Democracy, Council of Europe, pp 1–49
Buolamwini J, Gebru T (2018) Gender shades: intersectional accuracy disparities in commercial gender classification. In: 1st conference on fairness, accountability and transparency, Proceedings of Machine Learning Research, 4 February 2018
Burke R (2007) Hybrid web recommender systems. In: Brusilovsky P, Kobsa A, Nejdl W (eds) The adaptive web: methods and strategies of web personalization. Springer, Berlin, pp 377–408
Burtt EA (1939) The English philosophers: from bacon to mill. Modern Library, New York
Cadwalladr C, Graham-Harrison (2018) 50 million Facebook profiles harvested for Cambridge Analytica in major data breach. The Guardian. https://www.theguardian.com/news/2018/mar/17/cambridge-analytica-facebook-influence-us-election
China State Council (2017) Next generation artificial intelligence development plan. State Department for International Science and Technology Cooperation, China State Council, edited and translated by Rogier Creemers https://chinacopyrightandmedia.wordpress.com/2017/07/20/a-next-generation-artificial-intelligence-development-plan/. Accessed 28 Jan 2022
Conover M, Ratkiewicz J, Francisco M, Goncalves B, Menczer F, Flammini A (2021) Political polarization on Twitter. In: Full papers of the 5th International AAAI conference on weblogs and social media, Barcelona, Spain, 17—21 July 2011
Courtland R (2018) The bias detectives. Nature 558:357–360
Coviello L, Sohn Y, Kramer AD, Marlow C, Franceschetti M, Christakis NA, Fowler JH (2014) Detecting emotional contagion in massive social networks. PLoS One 9:e90315
Del Vicario M, Bessi A, Zollo F, Petroni F, Scala A, Caldarelli G, Stanley HE, Quattrociocchi W (2016) The spreading of misinformation online. Proc Natl Acad Sci USA 113:554–559
Dressel J, Farid H (2018) The accuracy, fairness, and limits of predicting recidivism. Sci Adv 4:eaao5580
Duhigg C (2012) How companies learn your secrets. https://www.nytimes.com/2012/02/19/magazine/shopping-habits.html. Accessed 28 Jan 2022
Eichstaedt JC, Smith RJ, Merchant RM, Ungar LH, Crutchley P, Preoţiuc-Pietro D, Asch DA, Schwartz HA (2018) Facebook language predicts depression in medical records. Proc Natl Acad Sci U S A 115:11203–11208
Emmott R (2020) Russia deploying coronavirus disinformation to sow panic in West, EU document says. https://www.reuters.com/article/us-health-coronavirus-disinformation-idUSKBN21518F. Accessed 28 Jan 2022
Englehardt S, Narayanan A (2016) Online tracking: a 1-million-site measurement and analysis. In: Proceedings of the 2016 ACM SIGSAC conference on computer and communications security, Association for Computing Machinery, Vienna, Austria, pp 1388–1401
Fayyaz Z, Ebrahimian M, Nawara D, Ibrahim A, Kashef R (2020) Recommendation systems: algorithms, challenges, metrics, and business opportunities. Appl Sci 10:7748
Flaxman S, Goel S, Rao JM (2016) Filter bubbles, echo chambers, and online news consumption. Public Opin Q 80:298–320
Garcia M (2016) Racist in the machine: the disturbing implications of algorithmic bias. World Policy J 33:111–117
Goncalves-Sa J (2020) In the fight against the new coronavirus outbreak, we must also struggle with human bias. Nat Med 26:305

Haidt J, Twenge J (2021) Social media use and mental health: a review. https://docs.google.com/document/d/1w-HOfseF2wF9YIpXwUUtP65-olnkPyWcgF5BiAtBEy0/mobilebasic#h.xi8mrj7rpf37. Accessed 28 Jan 2022

He X, Liao L, Zhang H, Nie L, Hu X, Chua T-S (2017) Neural collaborative filtering. In: Proceedings of the 26th International conference on world wide web, International World Wide Web Conferences Steering Committee, Perth, Australia, pp 137–182

Herlocker JL, Konstan JA, Riedl J (2000) Explaining collaborative filtering recommendations. In: Proceedings of the 2000 ACM conference on computer supported cooperative work, Association for Computing Machinery, Philadelphia, Pennsylvania, USA, pp 241–250

Homer-Dixon T (2001) We need a forest of tongues. https://homerdixon.com/we-need-a-forest-of-tongues/. Accessed 28 Jan 2022

Kadambi A (2021) Achieving fairness in medical devices. Science 372:30–31

Kramer ADI, Guillory JE, Hancock JT (2014) Experimental evidence of massive-scale emotional contagion through social networks. Proc Natl Acad Sci U S A 111:878–8790

Larson J, Mattu S, Kirchner L, Angwin J (2016) How we analyzed the COMPAS recidivism algorithm. https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm?token=XqwQ3rgbDdgxLwZrgdO5MED4b-chsjSu. Accessed 28 Jan 2022

Liang F, Das V, Kostyuk N, Hussain MM (2018) Constructing a data-driven society: China's social credit system as a state surveillance infrastructure. Policy Internet 10:415–453

Loomba S, de Figueiredo A, Piatek SJ, de Graaf K, Larson HJ (2021) Measuring the impact of COVID-19 vaccine misinformation on vaccination intent in the UK and USA. Nat Hum Behav 5:337–348

Lops P, de Gemmis M, Semeraro G (2011) Content-based recommender systems: state of the art and trends. In: Ricci F, Rokach L, Shapira B, Kantor PB (eds) Recommender systems handbook. Springer, Boston, pp 73–105

MacKenzie I, Meyer C, Noble S (2013) How retailers can keep up with consumers. https://www.mckinsey.com/ch/~/media/McKinsey/Industries/Retail/Our%20Insights/How%20retailers%20can%20keep%20up%20with%20consumers/How_retailers_can_keep_up_with_consumers_V2.pdf. Accessed 28 Jan 2022

McAlone N (2016) Why Netflix thinks its personalized recommendation engine is worth $1 billion per year. https://www.businessinsider.com/netflix-recommendation-engine-worth-1-billion-per-year-2016-6. Accessed 28 Jan 2022

McPherson M, Smith-Lovin L, Cook JM (2001) Birds of a feather: homophily in social networks. Annu Rev Sociol 27:415–444

Mikians J, Gyarmati L, Erramilli V, Laoutaris N (2012) Detecting price and search discrimination on the internet. In: Proceedings of the 11th ACM workshop on hot topics in networks, Association for Computing Machinery, Redmond, Washington, pp 79–84

Mooney RJ, Roy L (2000) Content-based book recommending using learning for text categorization. In: Proceedings of the fifth ACM conference on digital libraries, Association for Computing Machinery, San Antonio, Texas, USA, pp 195–204

Naumov M, Mudigere D, Shi H-JM, Huang J, Sundaraman N, Park J, Wang X, Gupta U, Wu C-J, Azzolini AG (2019) Deep learning recommendation model for personalization and recommendation systems. arXiv preprint arXiv:1906.00091

Nelson RE (1989) The strength of strong ties: social networks and intergroup conflict in organizations. Acad Manag J 32:377–401

Nikolov D, Oliveira DFM, Flammini A, Menczer F (2015) Measuring online social bubbles. PeerJ Comput Sci 1:e38

O'Neil C (2016) Weapons of math destruction: how big data increases inequality and threatens democracy. Crown Publishers, New York

O'Shaughnessy NJ (1990). Big lies, little lies: The story of propaganda. In *The Phenomenon of Political Marketing*. Palgrave Macmillan, London, UK, pp. 17–29

Paparrizos I, Cambazoglu BB, Gionis A (2011) Machine learned job recommendation. In: Proceedings of the fifth ACM conference on recommender systems, Association for Computing Machinery, Chicago, pp 325–328

Pennycook G, Rand DG (2021) The psychology of fake news. Trends Cogn Sci 25:388–402
Quattrociocchi W, Scala A, Sunstein CR (2016) Echo chambers on Facebook. SSRN
Reece AG, Reagan AJ, Lix KLM, Dodds PS, Danforth CM, Langer EJ (2017) Forecasting the onset and course of mental illness with Twitter data. Sci Rep 7:13006
Resnick P, Varian HR (1997) Recommender systems. Commun ACM 40:56–58
Ribeiro FN, Saha K, Babaei M, Henrique L, Messias J, Benevenuto F, Goga O, Gummadi KP, Redmiles EM (2019) On microtargeting socially divisive Ads: a case study of Russia-Linked Ad Campaigns on Facebook. In: Proceedings of the conference on fairness, Accountability, and Transparency. Association for Computing Machinery, Atlanta, pp 140–149
Ricard J, Medeiros J (2020) Using misinformation as a political weapon: Covid-19 and Bolsonaro in Brazil. Harv Kennedy School Misinform Rev 1(2)
Rigutini L, Maggini M (2004) Automatic text processing: machine learning techniques, Diss. PhD. thesis, University of Siena. https://www.researchgate.net/publication/236667720_ AUTOMATIC_TEXT_PROCESSING_MACHINE_LEARNING_TECHNIQUES Accessed 15 Jan 2022
Roozenbeek J, Schneider CR, Dryhurst S, Kerr J, Freeman ALJ, Recchia G, van der Bles AM, van der Linden S (2020) Susceptibility to misinformation about COVID-19 around the world. R Soc Open Sci 7:201199
Saleiro P, Kuester B, Hinkson L, London J, Stevens A, Anisfeld A, Rodolfa KT, Ghani R (2018) Aequitas: a bias and fairness audit toolkit. arXiv preprint arXiv:1811.05577
Sheng E, Chang KW, Natarajan P, Peng N (2021) Societal biases in language generation: Progress and challenges. arXiv preprint arXiv:2105.04054
Silva M, Benevenuto F (2021) COVID-19 ads as political weapon. In: Proceedings of the 36th Annual ACM symposium on applied computing, Association for Computing Machinery, Virtual Event, Republic of Korea, pp 1705–1710
Silva M, Oliveira LSD, Andreou A, Melo POVD, Goga O, Benevenuto F (2020) Facebook Ads Monitor: An independent auditing system for political ads on Facebook. In: WWW'20 Proceedings of The Web Conference, Taipei, Taiwan, 20–24 April 2020
Smith B, Linden G (2017) Two decades of recommender systems at Amazon.com. IEEE Internet Comput 21:12–18
Solomos K, Ilia P, Ioannidis S, Kourtellis N (2019) Clash of the trackers: measuring the evolution of the online tracking ecosystem. arXiv preprint arXiv:1907.12860
Spielkamp M (2017) Inspecting algorithms for bias. https://www.technologyreview.com/2017/06/12/105804/inspecting-algorithms-for-bias/. Accessed 28 Jan 2022
Ungar LH, Foster DP (1998) Clustering methods for collaborative filtering. In: AAAI workshop on recommendation systems, Madison, Wisconsin, 26–27, 31 July 1998
Urban T, Tatang D, Degeling M, Holz T, Pohlmann N (2020) Measuring the impact of the GDPR on data sharing in ad networks. In: Proceedings of the 15th ACM Asia conference on computer and communications security. Association for Computing Machinery, Taipei, Taiwan, pp 222–235
Vanderbilt T (2013) The science behind the Netflix algorithms that decide what you'll watch next. https://www.wired.com/2013/08/qq-netflix-algorithm/. Accessed 28 Jan 2022
Vosoughi S, Roy D, Aral S (2018) The spread of true and false news online. Science 359:1146–1151
Yang K-C, Torres-Lugo C, Menczer F (2020) Prevalence of low-credibility information on twitter during the covid-19 outbreak. arXiv preprint arXiv:2004.14484
Yardi S, Boyd D (2010) Dynamic debates: an analysis of group polarization over time on Twitter. Bull Sci Technol Soc 30:316–327
Zhang M (2015) Google Photos Tags Two African-Americans as Gorillas Through Facial Recognition Software. https://www.forbes.com/sites/mzhang/2015/07/01/google-photos-tags-two-african-americans-as-gorillas-through-facial-recognition-software/. Accessed 28 Jan 2022
Zuboff S (2019) The age of surveillance capitalism: the fight for a human future at the new frontier of power. Public Aff, New York