Recommender Systems and their Ethical Challenges

Silvia Milano[1]*, Mariarosaria Taddeo[1,2], Luciano Floridi[1,2]

[1]Oxford Internet Institute, University of Oxford, 1 St Giles, Oxford, OX1 3JS, United Kingdom
[2]The Alan Turing Institute, 96 Euston Road, London, NW1 2DB, United Kingdom.

* Corresponding author. Email: silvia.milano@oii.ox.ac.uk

Abstract

This article presents the first, systematic analysis of the ethical challenges posed by recommender systems. Through a literature review, the article identifies six areas of concern, and maps them onto a proposed taxonomy of different kinds of ethical impact. The analysis uncovers a gap in the literature: currently user-centred approaches do not consider the interests of a variety of other stakeholders—as opposed to just the receivers of a recommendation—in assessing the ethical impacts of a recommender system.

Keywords

Algorithms; Artificial Intelligence; Digital Ethics; Ethical Trade-offs; Ethics of Recommendation; Machine Learning; Recommender Systems.

Funding

This work was supported by Privacy and Trust Stream - Social lead of the PETRAS Internet of Things research hub. PETRAS is funded by the Engineering and Physical Sciences Research Council (EPSRC), grant agreement no. EP/N023013/1; and Google UK Limited.

1. Introduction

We interact with recommender (or recommendation) systems (RS) on a regular basis, when we use digital services and apps, from Amazon to Netflix and news aggregators. They are algorithms that make suggestions about what a user may like, such as a specific movie. Slightly more formally, they are functions that take information about a user's preferences (e.g. about movies) as an input, and output a prediction about the rating that a user would give of the items under evaluation (e.g., new movies available). We shall say more about the nature of recommender systems in the following pages, but even this general description suffices to clarify that, in order to work effectively and efficiently, recommender systems collect, curate, and act upon vast amounts of personal data. Inevitably, they end up shaping individual experience of digital environments and social interactions (Burr, Cristianini, & Ladyman, 2018; de Vries, 2010; Karimi, Jannach, & Jugovac, 2018).

RS are ubiquitous and there is already much technical research about how to develop ever more efficient systems (Adomavicius & Tuzhilin, 2005; Jannach & Adomavicius, 2016; Ricci, Rokach, & Shapira, 2015). In the past 20 years, RS have been developed focusing mostly on business applications, and the emphasis has tended to be on commercial objectives. But RS have a wider impact on users and on society more broadly. After all, they shape user preferences and guide choices, both individually and socially. This impact is significant and deserves ethical scrutiny, not least because RS can also be deployed in contexts that are morally loaded, such as health care, lifestyle, insurance, and the labour market. Clearly, whatever the ethical issues may be, they need to be understood and addressed by evaluating the *design*, *deployment* and *use* of the recommender systems, and the trade-offs between the different interests at stake. A failure to do so may lead to opportunity costs as well as problems that could otherwise be mitigated or avoided altogether, and, in turn, to public distrust and backlash against the use of RS in general (Koene et al., 2015).

Research into the ethical issues posed by RS is still in its infancy. The debate is also fragmented across different scientific communities, as it tends to focus on specific aspects and applications of these systems in a variety of contexts. The current fragmentation of the debate may be due to two main factors: the relative newness of the technology, which took off with the spread of internet-based services and the introduction of collaborative filtering techniques in the 1990s (Adomavicius & Tuzhilin, 2005; Pennock, Horvitz, & Giles, 2000); and the proprietary and privacy issues involved in the design and deployment of this class of algorithms. The details of RS currently in operation are treated as highly guarded industrial secrets. This makes it difficult for independent researchers to access information about their internal operations, and hence provide any evidence-

based assessment. In the same vein, due to privacy concerns, providers of recommendation systems may be reluctant to share information that could compromise their users' personal data (Paraschakis, 2018).

Against this background, this article addresses both problems (infancy and fragmentation), by providing a survey of the current state of the literature, and by proposing an overarching framework to situate the contributions to the debate. The overall goal is to reconstruct the whole debate, understand its main issues, and hence offer a starting point for better ways of designing RS and regulating their use.

2. A Working Definition of Recommender Systems

The task of a recommendation system – i.e. what we shall call the *recommendation problem* – is often summarized as that of *finding good items* (Jannach & Adomavicius, 2016). This description is common and popular among practitioners, especially in the context of e-commerce applications. However, it is too broad and not very helpful for research purposes. To make it operational one needs to specify, among other things, three parameters:

a) what the space of *options* is;
b) what counts as a *good* recommendation; and, importantly
c) how the RS's performance can be *evaluated*.

Specifying these parameter choices is highly dependent on the domain of application and the level of abstraction (LoAs, see (Floridi, 2016))[1] from which the problem is considered (Jannach, Zanker, Ge, & Gröning, 2012). Typically, the literature implements three LoAs: catalogue-based, decision support, and multi-stakeholder environment. Let us consider each of these in turn.

In e-commerce applications, the space of options (that is, the observables selected by the LoA) may be the items in the catalogue, while a good recommendation may be specified as one which ultimately results in a purchase. To evaluate the system performance, one may compare the RS's predictions to the actual user behaviour after a recommendation is made. In the domain of news recommendations, a good recommendation may be defined as a news item that is *relevant* to the user (Floridi, 2008), and one may use click-through rates as a proxy to evaluate the accuracy of the system's recommendations. Similar RS are designed to develop a model of individual users and to use it to predict the users' feedback on the system's recommendation, which is essentially a prediction problem.

[1] A level of abstraction can be imagined as an interface that enables one to observe some aspects of a s system analysed, while making other aspects opaque or indeed invisible. For example, one may analyse a house at the LoA of a buyer, of an architect, of a city planner, of a plumber, and so on. LoAs are common in computer science, where systems are described at different LoAs (computational, hardware, user-centred etc.). LoAs can be combined in more complex sets, and can be, but are not necessarily always, hierarchical.

Taking a different LoA, RS may also be considered to provide *decision support* to their users. For example, an online booking RS may be designed to facilitate the user's choice of hotel options. In this case, defining what counts as a good recommendation is more complex, because it involves appreciation of the user's goals and decision-making abilities. Evaluating the system's performance as a decision support requires more elaborate metrics. For example, (Jameson et al., 2015) consider six strategies for generating recommendations, which track different choice patterns based on either of the following features: (1) the attributes of the options; (2) the expected consequences of choosing an option; (3) prior experience with similar options; (4) social pressure or social information about the options; (5) following a specific policy; (6) trial-and-error based choice.

More recently, (Abdollahpouri, Burke, & Mobasher, 2017) have proposed a different kind of LoA (our terminology), defining RS in terms of multi-stakeholder environments (what we would call the LoA's observables), where multiple parties (including users, providers, and system administrators) can derive different utilities from recommendations. Epistemologically, this approach is helpful because it enables one to conceptualise explicitly the impact that RS have at different levels, both on the individual users interacting with them, and on society more broadly, making it possible to articulate what ethical trade-offs could be made between these different, possibly competing interests.

In view of the previous LoAs, and for the purposes of this article, we take recommender systems to be a class of algorithms that address the *recommendation problem* using a content-based or collaborative filtering approach, or a combination thereof. This choice has three advantages. It is compatible with the most common LoAs we have listed above. By focusing on the algorithmic nature of recommender systems, it also singles out one of the fastest growing areas of research and applications for machine learning. And it enables us to narrow down the scope of the study, as we shall not consider systems that approach the recommendation problem using different techniques, such as, for instance, expert systems like IBM Watson. With these advantages in mind, in the next section we propose a general taxonomy to identify the ethical challenges of RS. In section 4 we review the current literature, structured around six areas of concern. We conclude in section 5, by mapping the discussion onto our ethical taxonomy and indicating the direction of our further work in the area.

3. How to Map the Ethical Challenges Posed by Recommender Systems

In order to identify what is ethically at stake in the design and deployment of a RS, let us start with a formal taxonomy. This is how we propose to design it.

The question about which moral principles may be correct is deeply contentious and debated in philosophy. Fortunately, in this article we do not have to take a side because all we need is a distinction about which there is a general consensus: there are at least two classes of variables that are morally relevant, *actions* and *consequences*. Of course, other things could also be morally relevant, in particular *intentions*. However, for our purposes, the aforementioned distinction is all we need, so we shall assume that a recommender system's *behaviour* and *impact* will suffice to provide a clear understanding of what is ethically at stake.

The value of some consequences is often measured in terms of the *utility* they contain. So, it is reasonable to assume that any aspect of a RS that could impact negatively the utility of any of its stakeholders, or risk imposing such negative impacts, constitutes a feature that is ethically relevant.

While the concept of utility can be made operational using *quantifiable* metrics, rights are usually taken to provide *qualitative* constraints on actions. Thinking in terms of actions and consequences, we can identify two ways in which a recommender system can have ethical impacts. First, its operations can

a) impact (negatively) the utility of any of its stakeholders; and/or
b) violate their rights.

Second, these two kinds of ethical impact may be *immediate*—for example, a recommendation may be inaccurate, leading to a decrease in utility for the user—or they may expose the relevant parties to *future risks*. The ethics of risk imposition is the subject of a growing philosophical literature, which highlights how most activities involve imposition of risks (Hansson, 2010; Hayenhjelm & Wolff, 2012). In the case of RS, for example, the risks may involve exposing users to undue privacy violations by external actors, or the exposure to potentially irrelevant or damaging content. Exposure to risks of these sorts can constitute a wrong, even if no adverse consequences actually materialise.[2]

Given the previous analysis, we may now categorise the ethical issues caused by recommender systems along two dimensions (see Table 1):

i) whether a (given feature of a) RS negatively impacts the utility of some of its stakeholders or, instead, constitutes a rights violation, which is not necessarily measured in terms of utility; and
ii) whether the negative impact constitutes an immediate harm or it exposes the relevant party to future risk of harm or rights violation.

2 The idea that exposing someone to risks can constitute a wrong to them, even if the adverse consequences fail to materialise, is familiar from other contexts, e.g. medical ethics: for example, negligence in treating a patient constitutes a wrong, even if the patient ultimately recovers and does not suffer as a result of the negligence.

Table 1 summarises our proposed taxonomy, including some examples of different types of ethical impacts of recommender systems, to be discussed in section 5.

Table 1

	Immediate Harm	**Exposure to Risk**
Utility	e.g. inaccurate recommendations	e.g. A/B testing (see section 4.1)
Rights	e.g. unfair treatment	e.g. leaking of sensitive information

With the help of this taxonomy we are now ready to review the contributions provided by the current literature. We shall offer a general discussion of our findings in the conclusion.

4. The Ethical Challenges of Recommender Systems

The literature addressing the ethical challenges posed by RS is sparse, with the discussion of specific issues often linked to a specific instance of a RS, and appearing to be fragmented across disciplinary divides. Through a multidisciplinary, comparative meta-analysis, we identified six main areas of ethical concerns (see appendix for our methodology). They often overlap but, for the sake of clarity, we shall analyse them separately in the rest of this section.

4.1. Ethical content

Only a handful of studies to date address explicitly the ethics of RS as a specific issue in itself. Earlier work on the question of ethical recommendations focuses more on the content of the recommendations, and proposes ways to filter the items recommended by the system on the basis of cultural and ethical preferences. Four studies are particularly relevant. (Souali, El Afia, & Faizi, 2011) consider the issue of RSs that are not culturally appropriate, and propose an "ethical database", constructed on the basis of what are taken to be a region's generally accepted cultural norms, which act as a filter for the recommendations. (Tang & Winoto, 2016) take a more dynamic approach to the issue, proposing a two-layer RS, comprising a user-adjustable "ethical filter" that screens the items that can be recommended based on the user's specified ethical preferences. (Rodriguez & Watkins, 2009) adopt a more abstract approach to the problem of ethical recommendations, proposing a vision for a *eudaimonic* RS, whose purpose is to "produce societies in which the individuals experience satisfaction through a deep engagement in the world". This, the authors predict, could be made achievable through the use of interlinked big data structures.

Finally, (Paraschakis, 2016, 2017, 2018) provides one of the most detailed accounts. Focusing on e-commerce applications, Paraschakis suggests that there are five ethically problematic areas:

- the practices of user profiling,
- data publishing,
- algorithm design,
- user interface design, and
- online experimentation or A/B testing, i.e. the practice of exposing selected groups of users to modifications of the algorithm, with the aim of gathering feedback on the effectiveness of each version from the user responses.

The risks he identifies relate to breaches of a user's privacy (e.g. via data leaks, or by data gathering in the absence of explicit consent), anonymity breaches, behaviour manipulation and bias in the recommendations given to the user, content censorship, exposure to side effects, and unequal treatment in A/B testing with a lack of user awareness, leading to a lack of trust. The solutions put forward in (Paraschakis, 2017) revolve around a *user-centred* design approach (more on this in the next paragraph), introducing adjustable tools for users to control explicitly the way in which RS use their personal data, in order to filter out marketing biases or content censorship, and to opt out of online experiments.

With the exception of (Souali et al., 2011), who adopt a recommendation filter based on geographically-located cultural norms, the solutions described in this section rely on a user-centred approach. Recalling our taxonomy, they try to minimise the negative impact on the user's utility—in particular, unwanted exposure to testing, and inaccurate recommendations—and on the user's rights, in particular, recommendations that do not agree with the user's values, or expose them to privacy violations. However, user-centred solutions have significant shortcomings: they may not transfer to other domains, they may be insufficient to protect the user's privacy, and they may result in inefficiency, for example impairing the system's effectiveness in generating new recommendations, if enough users choose to opt out of profile tracking or online testing. Moreover, users' choice of parameters can reveal sensitive information about the users themselves. For example, adding a filter to exclude some kind of content gives away the information that the user may find this content distressing, irrelevant, or in other ways unacceptable. But above all, the main problem is that, although user-centred solutions may foster the transparency of recommender systems, they also shift the responsibility and accountability for the protection of rights and utility to the users. These points highlight how user-centred solutions in general are challenged by their demanding nature, as they may constitute a mere shift in responsibility when

the users are only nominally empowered but actually unable to manage all the procedures needed to protect their interests. This may therefore be an unfair shift, since it places undue burdens on the users, and is in any case problematic because the effectiveness of these solutions varies with the level of awareness and expertise of the users themselves, which may lead to users experiencing different levels of protection depending on their ability to control the technology.[3]

Implementing an "ethical filter" for a recommender system, as suggested by (Rodriguez & Watkins, 2009), would also be controversial in some applications, for example if it were used by a government to limit citizens' ability to access some politically sensitive contents. As for the eudaimonic approach, this goes in the direction of designing a recommender system that is an optimal decision support, yet it seems practically unfeasible, and at least much more research would be needed. Figuring out what is a "good human life" is something that millennia of reflection have not yet solved.

4.2. Privacy

User privacy is one of the primary challenges for recommendation systems (Friedman et al., 2015; Koene et al., 2015; Paraschakis, 2018). This may be seen as inevitable, given that a majority of the most commercially successful recommender systems are based on hybrid or collaborative filtering techniques, and work by constructing models of their users in order to generate personalised recommendations. Privacy risks occur in at least four stages. First, they can arise when data are collected or shared without the user's explicit consent. Second, once data sets are stored, there is the further risk that they may be leaked to external agents, or become subject to de-anonymization attempts (Narayanan, 2008). At both stages, privacy breaches expose users to risks, which may result in loss of utility (for example, if individual users are targeted by malicious agents as a result), or in rights violations (for example, if users' private information is utilised in ways that threaten their individual autonomy, see section 4.3 below). Third, and independently of how securely data are collected and stored, privacy concerns also arise at the stage of inferences that the system can (enable one to) draw from the data. Users may not be aware of the nature of these inferences, and they may object to this use of their personal data if they were better informed. Privacy risks do not only concern data collection because, for example, an external agent observing the recommendation that the system generates for a given user may be able to infer some sensitive information about the user (Friedman et al., 2015). Extending the notion of informed consent to

[3] For a critical analysis of empowerment see Jessica Morley and Luciano Floridi (forthcoming), "Against Empowerment: How to Reframe the Role of mHealth Tools in the Healthcare Ecosystem".

the indirect inferences from user recommendations appears difficult.[4] Finally, there is also another subtle, but important, systemic issue regarding privacy, which arises at the stage of collaborative filtering: the system can construct a model of the user based on the data it has gathered on other users' interactions. In other words, as long as enough users interact and share their data with the system, the system may be able to construct a fairly accurate profile even for those users about whom it has fewer data. This indicates that it may not be feasible for individual users to be shielded completely from the kinds of inferences that the system may be able to draw about them. It could be a positive feature in some domains, like medical research, but it may also turn out to be problematic in other domains, like recruitment or finance.

Current solutions to the privacy challenges intrinsic to recommender systems (especially those based on collaborative filtering techniques) fall into three broad categories, covering *architectures*, *algorithmic*, and *policy* approaches (Friedman et al., 2015). Privacy-enhancing architectures aim to mitigate privacy risks by storing user data in separate and decentralised databases, to minimise the risk of leaks. Algorithmic solutions focus on using encryption to minimise the risk that user data could be exploited by external agents for unwarranted purposes. Policy approaches, including GDPR legislation, introduce explicit guidelines and sanctions to regulate data collection, use, and storage.

The user-centred recommendation framework proposed by (Paraschakis, 2017), which we already encountered in the previous section, also introduces explicit privacy controls, letting the users decide whether their data can be shared, and with whom. However, as we have already remarked, user-centred approaches have limits, as they may constitute a mere shift in responsibility, placing an undue burden on the users. A possible issue that may arise specifically with user-enabled privacy controls is that the user's privacy preferences would, in themselves, constitute informative metadata, which the system (or external observers) could use to make sensitive inferences about the user, for example, to infer that a user who has strong privacy settings may have certain psychological traits, or that they may have "something to hide". When considering systemic inferences, due to the nature of collaborative filtering methods, even if user-centred adjustments could be implemented across the board in effective ways, they would arguably still not solve the problem.

Crucially, due to the nature of recommender systems – which, as we have seen, rely on user models in order to generate personalised recommendations – any approach to the issue of user privacy will need to take into account the likely trade-off between privacy and accuracy, but

[4] The recent ProPublica/Facebook exchange about auditing targeted ads may configure as a privacy breach of this kind (Merrill & Tobin, 2019).

also fairness and explainability of algorithms (Friedman et al., 2015; Koene et al., 2015). For this reason, ethical analyses of recommender systems are better developed by embracing a macro-ethical approach. This is an approach that is able to consider specifically ethical problems related to data, algorithms, and practices, but also how the problems relate, depend on, and impact each other (Floridi & Taddeo, 2016).

4.3. Autonomy and Personal Identity

Recommender systems can encroach on individual users' autonomy, by providing recommendations that nudge users in a particular direction, by attempting to "addict" them to some types of contents, or by limiting the range of options to which they are exposed (Burr et al., 2018; de Vries, 2010; Koene et al., 2015; Taddeo & Floridi, 2018). These interventions can range from being benign (enabling individual agency and supporting better decision making by filtering out irrelevant options), to potentially questionable (persuasion, nudging), to possibly malign (being manipulative and coercive (Burr et al., 2018)).

Algorithmic classification used to construct user models on the basis of aggregate user data can reproduce social categories. This may introduce bias in the recommendations. We shall discuss this risk in detail in the next section (4.4). Here, the focus is on a distinctive set of issues arising when the algorithmic categorization of users does not follow recognisable social categories. (de Vries, 2010) powerfully articulates the idea that our experience of personal identity is mediated by the categories to which we are assigned. Algorithmic profiling, performed by recommender systems, can disrupt this individual experience of personal identity, for at least two main reasons. First, the recommender system's model of each user is continuously reconfigured on the basis of the feedback provided by other users' interactions with the system. In this sense, the system should not be conceptualised as tracking a pre-established user identity and tailoring its recommendations to it, but rather as contributing to the construction of the user identity dynamically (Floridi, 2011). Second, the labelling that the system uses to categorise users may not correspond to recognisable attributes or social categories with which the user would self-identify (for example, because machine-generated categories may not correspond to any known social representation), so even if users could access the content of the model, they would not be able to interpret it and connect it with their lived experiences in a meaningful way. These features of recommender systems create an environment where personalization comes at the cost of removing the user from the social categories that help mediate their experiences of identity.

In this context, an interesting take on the issue of personal autonomy in relation to recommender systems comes from the "captology" of recommender systems. (Seaver, 2018a) develops this concept from an anthropological perspective:

> [a]s recommender[s] spread across online cultural infrastructures and become practically inescapable, thinking with traps offers an alternative to common ethical framings that oppose tropes of freedom and coercion (Seaver, 2018a).

Recommender systems appear to function as "sticky traps" (our terminology) insofar as they are trying to "glue" their users to some specific solutions. This is reflected in what Seaver calls "captivation metrics" (i.e. that measure user retention), which are commonly used by popular recommender systems. A prominent example is YouTube's recommendation algorithm, which received much attention recently for its tendency to promote biased content and "fake news", in a bid to keep users engaged with its platform (Chaslot, 2018). Regarding recommender systems as traps requires engaging with the minds of the users: traps can only be effective if their creators understand and work with the target's world view and motivations, so the autonomous agency of the target is not negated, but effectively exploited. Given this captological approach, and given the effectiveness and ubiquity of the traps of recommender systems, the question to ask is not how users can escape *from them*, but rather how users can make the traps work *for them*.

4.4. Opacity

In theory, explaining how personalised recommendations are generated for individual users could help to mitigate the risk of encroaching on their autonomy, giving them access to the reasons why the system "thinks" that some options are relevant to them. It would also help increase the transparency of the algorithmic decisions concerning how to class and model users, thus helping to guard against bias.

Designing and evaluating explanations for recommender systems can take different forms, depending on the specific applications. As reported by (Tintarev & Masthoff, 2011), several studies have pursued a user-centred approach to evaluation metrics, including metrics to evaluate explanations of recommendations. What counts as a good explanation depends on several criteria: the purpose of the recommendation for the user; whether the explanation accurately matches the mechanism by which the recommendation is generated; whether it improves the system's transparency and scrutability; and whether it helps the user to make decisions more efficiently (e.g. more quickly), and more effectively, e.g. in terms of increased satisfaction.

These criteria are satisfied by factual explanations.[5] However, factual explanations are notoriously difficult to achieve. As noted by (Herlocker, Konstan, & Riedl, 2000), recommendations generated by collaborative filtering techniques can, on a simple level, be conceptualised as analogous to "word of mouth" recommendations among users. However, offline word of mouth recommendations can work on the basis of trust and shared personal experience, whereas in the case of recommender systems users do not have access to the identity of the other users, nor do they have access to the models that the system uses in order to generate the recommendations. As we mentioned, this is an issue in so far as it diminishes the user's autonomy. It may be difficult to provide good factual explanations in practice also for computational reasons (the required computation to generate a good explanation may be too complex), and because they may have distorting effects on the accuracy of the recommendations (Tintarev & Masthoff, 2011). For example, explaining to a user that a certain item is recommended because it is the most popular with other users may increase the item's desirability, thus generating a self-reinforcing pattern where the item will be recommended more often because it is popular. This, in turn, reinforces its popularity, ending in a winner-takes-all scenario that, depending on the intended domain of application, can have negative effects on the variety of options, plurality of choices, and the emergence of competition (Germano, Gómez, & Mens, 2019). Arguably, this may be one of the reasons why Amazon does not automatically privilege products with less than perfect scoring but that have been rated by a large number of reviewers.

4.5. Fairness

Fairness in algorithmic decision making is a wide-ranging issue, made more complicated by the existence of multiple notions of fairness, which are not all mutually compatible (Friedler, Scheidegger, & Venkatasubramanian, 2016). In the context of recommender systems, several articles identified in this review address the issue of recommendations that may reproduce *social biases*. They may be synthesised around two approaches.

On the one hand, (Yao & Huang, 2017) consider several possible sources for unfairness in collaborative filtering, and introduce four new metrics to address them by measuring the distance between recommendations made by the system to different groups of users. Focusing on

[5] Factual explanations are usually contrasted to *counterfactual* ones, that describe what would have had to be the case, in order for a certain state or outcome (different from the actual one) to occur. For example, suppose that while browsing an e-commerce website, Alice is recommended a brand of dog food. A counterfactual explanation of why Alice received this recommendation would specify what would have had to be the case, for Alice *not* to be recommended this specific product (for example, had she not browsed dog collars, she would not have been recommended dog food). A factual explanation, on the other hand, would specify *why* this specific item was recommended, for example why this specific brand of dog food was deemed good for Alice.

collaborative filtering techniques, they note that these methods assume that the missing ratings (i.e., the ones that the system needs to infer from the statistical data to predict a user's preferences) are randomly distributed. However, this assumption of randomness introduces a potential source of bias in the system's predictions, because it is well documented that users' underlying preferences often differ from the sampled ratings, since the latter are affected by social factors, which may be biased (Marlin, Zemel, Roweis, & Slaney, 2007). Following (Yao & Huang, 2017), (Farnadi, Kouki, Thompson, Srinivasan, & Getoor, 2018) also identify the two primary sources of bias in recommender systems with two problematic patterns of data collection, namely *observation bias*, which results from feedback loops generated by the system's recommendations to specific groups of users, and *population imbalance*, where the data available to the system reflect existing social patterns expressing bias towards some groups. They propose a probabilistic programming approach to mitigate the system's bias against protected social groups.

On the other hand, (Burke, 2017) suggests to consider fairness in recommendation systems as a *multi-sided concept*. Based on this approach, he focuses on three notions of fair recommendations, taking the perspective of either the user/*consumer* (C-fairness*)*; or the *provider* (P-fairness); or a combination of the two (CP-Fairness). This taxonomy enables the developer of a recommendation system to identify how the competing interests of different parties are affected by the system's recommendations, and hence design system architectures that can mediate effectively between these interests.

In both approaches, the issue of fairness is tied up with choosing the right LoA for a specific application of a recommender system. Given that the concept of fairness is strongly tied to the social context within which the system gathers its data and makes recommendations, extending the same approach to any application of recommender systems may not be viable.

4.6. Polarization and social manipulability

A much-discussed effect of some recommender systems is their transformative impact on society. In particular, news recommender systems and social media filters, by nature of their design, run the risk of insulating users from exposure to different viewpoints, creating self-reinforcing biases and "filter bubbles" that are damaging to the normal functioning of public debate, group deliberation, and democratic institutions more generally (Bozdag, 2013; Bozdag & van den Hoven, 2015; Harambam, Helberger, & van Hoboken, 2018; Helberger, Karppinen, & D'acunto, 2016; Koene et al., 2015; Reviglio, 2017; Zook et al., 2017). A closely related issue is protecting these systems from manipulation by (sometimes even small but) especially active groups of users, whose interactions with the system can generate intense positive feedback, driving up the system's rate of

recommendations for specific items (Chakraborty, Patro, Ganguly, Gummadi, & Loiseau, 2019). News recommendation systems, streaming platforms, and social networks can become an arena for targeted political propaganda, as demonstrated by the recent Cambridge Analytical scandal in 2018, and the documented external interference in US political elections in recent years (Howard, Ganesh, Liotsiou, Kelly, & François, 2019).

The literature on the topic proposes a range of approaches to increase the diversity of recommendations. A point noted by several authors is that news recommendation systems, in particular, must reach a trade-off between the expected relevance to the user and diversity when generating personalised recommendations based on pre-specified user preferences or behavioural data (Helberger et al., 2016; Reviglio, 2017). In this respect, (Bozdag & van den Hoven, 2015) argue that the design of algorithmic tools to combat informational segregation should be more sensitive to the democratic norms that are implicitly built into these tools.

In general, the approaches to the issue of polarization and social manipulability appear to be split between bottom-up and top-down strategies, prioritizing either the preferences of users (and their autonomy in deciding how to configure the personalised recommendations) or the social preference for a balanced public arena. Once again, some authors take a decidedly user-centred perspective. For example, (Harambam et al., 2018) propose the use of different "recommendation personae", or "pre-configured and anthropomorphised types of recommendation algorithms" expressing different user preferences with respect to novelty, diversity, relevance, and other attributes of a recommendation algorithm. In the same vein, (Reviglio, 2017) stresses the importance of promoting serendipity even at the cost of sacrificing aspects of the user experience, such as diminished relevance of the recommendations.

5. Conclusion

Based on the review of the literature presented in the previous section, we can now revisit the taxonomy that we proposed in Section 3, and place the concerns that we have identified within the conceptual space that it provides. Table 2 summarises our results.

Table 2

	Immediate Harm	**Exposure to Risk**
Utility	Biased recommendations (4.1)	Opacity (4.4) Questionable content (4.1)
Rights	Unfair recommendations (4.5) Encroachment on individual autonomy and identity (4.3)	Privacy (4.2) Social manipulability and Polarisation (4.6)

Starting with privacy, the main challenge that is linked with privacy violations is the possibility of unfair or otherwise malicious uses of personal data to target individual users. Thus, from our review, it emerges that privacy concerns may be best conceptualised as *exposure to risk*. Moreover, the types of risk to which privacy violations expose users fall mainly under the category of *rights violations*, such as unfair targeting and use of manipulative techniques.

Issues of personal autonomy and identity also fall under the category of *rights violations*, and constitute cases of *immediate* violations. Unfair recommendations can be associated with a negative impact on utility but, as also noted by (Yao & Huang, 2017), fairness and utility are mutually independent, and unfairness may be best classified as a type of immediate right violation.

A notable insight that emerges from the review is that most of the ethical impacts of recommender systems identified in the literature are analysed from the perspective of the receivers of the recommendations. This is evident not only in the reliance on accuracy metrics measuring the distance between user preferences and recommendations, but also when considering that privacy, unfairness, opacity, and the appropriateness of content are judged from the perspective of the individual receiving the recommendations. However, individual users are not the only stakeholders of recommender systems (Burke, 2017). The utility, rights, and risks carried by *providers* of recommender systems, and by *society* at large, should also be addressed explicitly in the design and operation of recommender systems. And there are also more complex, nested cases in which recommendations concern third-parties (e.g., what to buy for a friend's birthday). Currently, this is (partially) evident only in the case of discussion on social polarization and its effects on democratic institutions (reviewed in section 4.6). Failure to address explicitly these additional perspectives of the ethical impact of recommender systems may lead to masking seriously problematic practices. A case in point may be that of introducing a "bias" in favour of recommending unpopular items to maximise catalogue coverage in e-commerce applications (Jameson et al., 2015). This practice meets a specific need of the provider of a recommendation

system, helping to minimise the number of unsold items, which in this specific instance may be considered a legitimate interest to be traded off against the utility that a user may receive from a more accurate recommendation. However, modelling the provider's interests as a bias added to the system is unhelpful if the aim is to identify what would be the right level of trade-off between the provider's and users' interests.

Any recommendation is a nudging, and any nudging embeds values. The opacity about which and whose values are at stake in recommender systems hinders the possibility of designing better systems that can also promote socially preferable outcomes and improve the balance between individual and non-individual utilities.

The distribution of the topics by discipline also reveals some interesting insights. Among the reviewed articles, the ones addressing privacy, fairness and opacity come predominantly from computer science. This is in line with the general trends in the field of algorithmic approaches to decision making, and the presence of established metrics and technical approaches to address these challenges.

In contrast, the challenges posed by socially transformative effects, manipulability, and personal autonomy are more difficult to address using purely technical approaches, largely because their definitions are qualitative, more contentious, and require viewing recommender systems in the light of the social context in which they operate. Thus, the articles identified in this review that relate to these issues are much more likely to come from philosophy, anthropology, and science and technology studies. The methodologies that they adopt are more varied, ranging from ethnographic study (Seaver, 2018b), to hermeneutics (de Vries, 2010), decision theory (Burr et al., 2018), and economics (Abdollahpouri et al., 2017).

This article offers a map and an analysis of the main ethical challenges posed by recommender systems, as identified in the current literature. It also highlights a gap in the relevant literature, insofar as it stresses the need to consider the interests of providers of recommender systems, and of society at large (including third-party, nested cases of recommendations), and not only of the receivers of the recommendation, when assessing the ethical impact of recommender systems. The next steps are, therefore, filling the gap, and articulating a comprehensive framework for addressing the ethical challenges posed by recommender systems, based on the taxonomy and the findings of this review.

6. Appendix: Methodology

We performed a keyword search on five widely used reference repositories (Google Scholar, IEEE Xplore, SCOPUS, PhilPapers and ArXiv), using a sting of the general form:

((moral* OR ethic*) AND (recommend* AND (system* OR algorithm*)))

The keyword search produced a total of 533 results, including 417 results on Google Scholar, 54 results on Scopus, 48 results on IEEE Xplore, seven results on PhilPapers, and seven results on ArXiv. After eliminating duplicate entries, and screening out the irrelevant entries based on the title and abstract, 50 relevant entries were left. These were reviewed in more detail. Finally, additional entries were added following the citations in the reviewed articles. The result was a corpus of 37 relevant works, discussed in this review and listed in the References.

7. References

Abdollahpouri, H., Burke, R., & Mobasher, B. (2017). *Recommender Systems as Multistakeholder Environments*. https://doi.org/10.1145/3079628.3079657

Adomavicius, G., & Tuzhilin, A. (2005). Toward the next generation of recommender systems: a survey of the state-of-the-art and possible extensions. *IEEE Transactions on Knowledge and Data Engineering*, *17*(6), 734–749. https://doi.org/10.1109/TKDE.2005.99

Bozdag, E. (2013). Bias in algorithmic filtering and personalization. *Ethics and Information Technology*, *15*, 209–227. https://doi.org/10.1007/s10676-013-9321-6

Bozdag, E., & van den Hoven, J. (2015). Breaking the filter bubble: democracy and design. *Ethics and Information Technology*, *17*(4), 249–265. https://doi.org/10.1007/s10676-015-9380-y

Burke, R. (2017). *Multisided Fairness for Recommendation*.

Burr, C., Cristianini, N., & Ladyman, J. (2018). An Analysis of the Interaction Between Intelligent Software Agents and Human Users. *Minds and Machines*, *28*(4), 735–774. https://doi.org/10.1007/s11023-018-9479-0

Chakraborty, A., Patro, G. K., Ganguly, N., Gummadi, K. P., & Loiseau, P. (2019). Equality of Voice: Towards Fair Representation in Crowdsourced Top-K Recommendations. *FATREC*. https://doi.org/10.1145/3287560.3287570

Chaslot, G. (2018, February 1). How Algorithms Can Learn to Discredit the Media – Guillaume Chaslot – Medium. *Medium*.

de Vries, K. (2010). Identity, profiling algorithms and a world of ambient intelligence. *Ethics and Information Technology*, *12*(1), 71–85. https://doi.org/10.1007/s10676-009-9215-9

Farnadi, G., Kouki, P., Thompson, S. K., Srinivasan, S., & Getoor, L. (2018). A Fairness-aware Hybrid Recommender System. *2nd FATREC Workshop: Responsible Recommendation*.

Floridi, L. (2008). Understanding Epistemic Relevance. *Erkenntnis*, *69*(1), 69–92. https://www.jstor.org/stable/40267374

Floridi, L. (2011). The Construction of Personal Identities Online. *Minds and Machines*, *21*(4), 477–479. https://doi.org/10.1007/s11023-011-9254-y

Floridi, L. (2016). The Method of Levels of Abstraction. In L. Floridi (Ed.), *The Routledge Handbook of Philosophy of Information* (pp. 67–72). Routledge.

Floridi, L., & Taddeo, M. (2016). What is data ethics? *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, *374*(2083), 20160360. https://doi.org/10.1098/rsta.2016.0360

Friedler, S. A., Scheidegger, C., & Venkatasubramanian, S. (2016). *On the (im)possibility of fairness *.*

Friedman, A., Knijnenburg, B., Vanhecke, K., Martens, L., Berkovsky, S., & Berkovsky CSIRO, S. (2015). Privacy Aspects of Recommender Systems. In F. Ricci, L. Rokach, & B. Shapira (Eds.), *Recommender Systems Handbook* (2nd ed., pp. 649–688). New York: Springer Science + Business Media.

Germano, F., Gómez, V., & Mens, G. L. (2019). The few-get-richer: a surprising consequence of popularity-based rankings. *ArXiv:1902.02580 [Cs]*. Retrieved from http://arxiv.org/abs/1902.02580

Hansson, S. O. (2010). The Harmful Influence of Decision Theory on Ethics. *Ethical Theory and Moral Practice*, *13*(5), 585–593. https://doi.org/10.1007/s10677-010-9232-0

Harambam, J., Helberger, N., & van Hoboken, J. (2018). Democratizing algorithmic news recommenders: how to materialize voice in a technologically saturated media ecosystem. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, *376*(2133), 20180088. https://doi.org/10.1098/rsta.2018.0088

Hayenhjelm, M., & Wolff, J. (2012). The Moral Problem of Risk Impositions: A Survey of the Literature. *European Journal of Philosophy*, *20*, E26–E51. https://doi.org/10.1111/j.1468-0378.2011.00482.x

Helberger, N., Karppinen, K., & D'acunto, L. (2016). *Exposure diversity as a design principle for recommender systems*. https://doi.org/10.1080/1369118X.2016.1271900

Herlocker, J. L., Konstan, J. A., & Riedl, J. (2000). *Explaining Collaborative Filtering Recommendations*.

Howard, P. N., Ganesh, B., Liotsiou, D., Kelly, J., & François, C. (2019). *The IRA, Social Media and Political Polarization in the United States, 2012-2018*.

Jameson, A., Mrtijn c>, W., Felfernig, A., de Gemmis, M., Lops, P., Semeraro, G., & Chen, L. (2015). Human Decision Making and Recommender Systems. In Francesco Ricci, L. Rokach, & B. Shapira (Eds.), *Recommender Systems Handbook*. Springer.

Jannach, D., & Adomavicius, G. (2016). Recommendations with a Purpose. *RecSys'16*. https://doi.org/10.1145/2959100.2959186

Jannach, D., Zanker, M., Ge, M., & Gröning, M. (2012). *Recommender Systems in Computer Science and Information Systems – A Landscape of Research*. https://doi.org/10.1007/978-3-642-32273-0_7

Karimi, M., Jannach, D., & Jugovac, M. (2018). *News Recommender Systems - Survey and Roads Ahead*. 1–49.

Koene, A., Perez, E., Carter, C. J., Statache, R., Adolphs, S., O'Malley, C., … McAuley, D. (2015). *Ethics of Personalized Information Filtering*. https://doi.org/10.1007/978-3-319-18609-2_10

Marlin, B. M., Zemel, R. S., Roweis, S., & Slaney, M. (2007). Collaborative Filtering and the Missing at Random Assumption. *UAI*.

Merrill, J. B., & Tobin, A. (2019). Facebook Moves to Block Ad Transparency Tools — Including Ours. *ProPublica*.

Narayanan, A. (2008). IEEE Xplore - Robust De-anonymization of Large Sparse Datasets. *SP '08 Proceedings of the 2008 IEEE Symposium on Security and Privacy*. https://doi.org/10.1109/SP.2008.33

Paraschakis, D. (2016). Recommender Systems from an Industrial and Ethical Perspective. *Proceedings of the 10th ACM Conference on Recommender Systems - RecSys '16*, 463–466. https://doi.org/10.1145/2959100.2959101

Paraschakis, D. (2017). Towards an ethical recommendation framework. *2017 11th International Conference on Research Challenges in Information Science (RCIS)*, 211–220. https://doi.org/10.1109/RCIS.2017.7956539

Paraschakis, D. (2018). *Algorithmic and Ethical Aspects of Recommender Systems in E-Commerce*. Malmö.

Pennock, D. M., Horvitz, E., & Giles, C. L. (2000). Social Choice Theory and Recommender Systems: Analysis of the Axiomatic Foundations of Collaborative Filtering. *AAAI-00*.

Reviglio, U. (2017). *Serendipity by Design? How to Turn from Diversity Exposure to Diversity Experience to Face Filter Bubbles in Social Media*. https://doi.org/10.1007/978-3-319-70284-1_22

Ricci, Francesco, Rokach, L., & Shapira, B. (Eds.). (2015). *Recommender Systems Handbook* (2nd ed.). Retrieved from https://www.springer.com/gb/book/9781489976369

Rodriguez, M. A., & Watkins, J. H. (2009). *Faith in the Algorithm, Part 2: Computational Eudaemonics*.

Seaver, N. (2018a). Captivating algorithms: Recommender systems as traps. *Journal of Material Culture*, 135918351882036. https://doi.org/10.1177/1359183518820366

Seaver, N. (2018b). Captivating algorithms: Recommender systems as traps. *Journal of Material Culture*, 135918351882036. https://doi.org/10.1177/1359183518820366

Souali, K., El Afia, A., & Faizi, R. (2011). An automatic ethical-based recommender system for e-commerce. *2011 International Conference on Multimedia Computing and Systems*, 1–4. https://doi.org/10.1109/ICMCS.2011.5945631

Taddeo, M., & Floridi, L. (2018). How AI can be a Force for Good. *Science*, *361*(6404), 751–752. https://doi.org/10.1126/science.aat5991

Tang, T. Y., & Winoto, P. (2016). I should not recommend it to you even if you will like it: the ethics of recommender systems. *New Review of Hypermedia and Multimedia*, *22*(1–2), 111–138. https://doi.org/10.1080/13614568.2015.1052099

Tintarev, N., & Masthoff, J. (2011). Designing and Evaluating Explanations for Recommender Systems. In *Recommender Systems Handbook* (pp. 479–510). https://doi.org/10.1007/978-0-387-85820-3_15

Yao, S., & Huang, B. (2017). Beyond Parity: Fairness Objectives for Collaborative Filtering. *NIPS*. https://doi.org/10.1177/0143831X03024002003

Zook, M., Barocas, S., boyd, danah, Crawford, K., Keller, E., Gangadharan, S. P., … Pasquale, F. (2017). Ten simple rules for responsible big data research. *PLOS Computational Biology*, *13*(3), e1005399. https://doi.org/10.1371/journal.pcbi.1005399