



# Social influence for societal interest: a pro-ethical framework for improving human decision making through multi-stakeholder recommender systems

Matteo Fabbri<sup>1</sup>

Received: 29 July 2021 / Accepted: 13 April 2022 / Published online: 28 May 2022  
© The Author(s) 2022

## Abstract

In the contemporary digital age, recommender systems (RSs) play a fundamental role in managing information on online platforms: from social media to e-commerce, from travels to cultural consumptions, automated recommendations influence the everyday choices of users at an unprecedented scale. RSs are trained on users' data to make targeted suggestions to individuals according to their expected preference, but their ultimate impact concerns all the multiple stakeholders involved in the recommendation process. Therefore, whilst RSs are useful to reduce information overload, their deployment comes with significant ethical challenges, which are still largely unaddressed because of proprietary constraints and regulatory gaps that limit the effects of standard approaches to explainability and transparency. In this context, I address the ethical and social implications of automated recommendations by proposing a pro-ethical design framework aimed at reorienting the influence of RSs towards societal interest. In particular, after highlighting the problem of explanation for RSs, I discuss the application of beneficent informational nudging to the case of conversational recommender systems (CRSs), which rely on user-system dialogic interactions. Subsequently, through a comparison with standard recommendations, I outline the incentives for platforms and providers in adopting this approach and its benefits for both individual users and society.

**Keywords** AI ethics · Recommender systems · Social influence · Explainability

## 1 Introduction

Nowadays recommender systems (RSs) are among the main technical tools for organizing information on online platforms. From social media to e-commerce, information flows are managed by algorithms aimed at predicting the choices of users, and therefore at optimizing the offer of the platforms. As Milano et al. (2021) argue, “RSs are a ubiquitous feature of digital environments, because they respond to a pressing need to reduce information overload, and facilitate interactions on multisided platforms with large numbers of participants”. For this reason, the design of RSs poses pressing ethical questions as regards their internal ontology and external influence over individuals and society at large. In fact, to account for the ethical and societal implications of RSs, it is necessary to analyze how their technical structure informs the everyday habits and choices of individuals in the

contemporary information age. As this is true for every digital technology, I argue that this approach is especially significant for RSs, because they represent a direct link between artificial intelligence (AI) and web platforms, where some of the most important economic, social and political transactions take place. Indeed, the large-scale application of AI to human interactions in online environments may originate socio-technical systems that are even more influencing and totalizing than the ones which we experience today.<sup>1</sup>





With this background in mind, I would like to address the ethical challenges and the societal opportunities brought about by RSs from the perspective of their possibly benevolent use. This essay is not aimed at providing an ethical taxonomy of the risks posed by RSs, which has already been proposed in a seminal work by Milano et al. (2020). However, building on this established ethical analysis, my aim is to elaborate on some aspects of the design of RSs to underline the impact of their application to online platforms and to suggest new paths for their development in the

✉ Matteo Fabbri  
matteo.fabbri@oii.ox.ac.uk

<sup>1</sup> University of Oxford, Oxford, UK

<sup>1</sup> For an account of platforms as all-encompassing socio-technical systems, see Bratton (2016).

**Table 1** Description and exemplification of each category of stakeholders within the recommendation process

Stakeholder category	Description	Example
 <b>Users</b>	Targets of the recommendation	Consumers looking for a product to buy on Amazon
 <b>Providers</b>	The agents who make the options available by selling their products or services through the platform	Musicians and singers on Spotify
 <b>System</b>	The interests of the platform that hosts providers' product/content and targets recommendations to users	The profits that Uber gets through a commission on each drive
 <b>Society</b>	The overall effects of automated recommendations on the social aggregate composed by all the stakeholders and the wider social environment	The impact of targeted fake news about vaccines on public health

direction of social good. To this end, I rely on the concept of multi-stakeholder recommender system (MRS) proposed by Milano et al. (2021) to account for the different layers of the ontological structure of RSs and the different levels of abstraction (LoA) from which RSs can be considered. In this way, I can analyze the social implications of RSs from the perspective of the various sets of actors involved in the recommendation process. The analytic model based on MRSs is also useful to inform policies addressing algorithmic design and auditing. Whilst a comprehensive policy analysis of RSs is outside the scope of this essay, I will propose an overview of the ontology and the social implications of recommender algorithms used by mainstream online platforms. Building on this outline, I will focus on how RSs influence the behaviour of platform users, and how their influencing potential could be applied to social good.

## 2 The distribution of interests within the recommendation process: a multistakeholder approach

Firstly, it should be recalled that, according to the multi-stakeholder approach proposed by Milano et al. (2021), there are four stakeholders in a recommendation: users, providers, the system and society (see Table 1). Users are “the parties to whom the recommendation is targeted”; providers are the subjects “who make the options available”, who sell their product or service through the platform; the system refers

to “the interests of the platform on which the recommendations are generated”; and finally “recommendations made by a system can have systemic effects on society, for example by altering or reinforcing existing social norms”, or by modifying a social environment. This ontological structure is not fixed nor represented in all RSs, as the four stakeholders could be subdivided into smaller categories or grouped together, according to the specific platform considered. An individual might belong to more than one category of stakeholders at a time. Moreover, in some contexts, stakeholders might coincide with one another: for instance, in dating apps the users are also the providers.

In particular, let us consider the example of Amazon as an e-commerce platform: in that case, users correspond to the consumers who are looking for a product to buy; providers are those who sell the products through the platform; Amazon itself is the system, which manages the RSs through which the products are suggested to users. In this process, automated recommendations create the social environment in which the interests of all the stakeholders are developed: users' interest to find a product at a convenient price, sellers' interest to reach a wider pool of potential consumers, the platform's interest in becoming an all-encompassing marketplace, where many sellers and buyers are attracted because they can meet in an optimized way. The case of Amazon is peculiar because the platform can sometimes coincide with the provider: indeed, some products are directly sold by the platform, which keeps them in a physical warehouse, whilst others are sold by external providers which are hosted on

the platform in exchange for a fee and/or a percentage on every sale.

The ontological structure of MRSs makes it clear that the interests of each stakeholder may differ within the same recommendation process. For example, let us focus on the concept of the utility of the user, which is central to the user-centered approach to RSs. If utility is conceived as the satisfaction of preferences, then the task of a RS would be to find and suggest good items, “interpreted as those ‘things’ that are most relevant or that match most accurately the preferences of the user to whom the recommendation is targeted” (Milano et al. 2021). In the case of a commercial platform, the utility of the user might not always coincide with that of the system, which could recommend the less popular products of its catalogue to sell them and empty an overloaded section of its physical warehouse. In this situation, a bias would be introduced in the recommendation algorithm so that an unpopular item could be suggested to the user regardless of their preference. As the user-centred approach is not able to account for the different stakeholders, that systemic bias cannot be explained as an expression of the interests of the platform within a framework limited only to users’ utility. Therefore, analysing the design of RSs from a multi-stakeholder perspective would enhance the understanding of the social implications of online recommendations.

Moreover, the concept of utility as the satisfaction of users’ preferences which is implemented in the accuracy metrics of many RSs could imply a focus on exploitation rather than on exploration of the space of choices. In other words, if a recommendation process is based mainly on tracking and learning users’ preferences to repropose them later, then the individual’s desires could be led towards standardization and homologation, resulting in bad consumer choices overall. For example, if a user tends to buy unhealthy food and receives recommendations proposing the same kind of food every time, their diet can be affected in a negative way. This implies that the influence of RSs on users based on the exploitation of their predicted utility may not always lead to good societal outcomes. Introducing social interest as a stakeholder in the design of a RS could entail a shift towards exploration in the recommendation policy. In fact, if we relied on a multi-stakeholder framework to address the example above, the consumer could be offered a healthier product that does not correspond to his/her previous choices, but might lead to a new stream of preferences which are better for him/her as an individual and for society as a whole (consider for instance the cost of obesity for the public health system, etc.).

This example indicates that choosing a particular accuracy metric over another one to be implemented in a RS can have a very different impact on the outcomes of the recommendation. Moreover, it is evident that the influence of automated recommendations can be extended beyond a single one-time

choice. Therefore, the consequences of the design of RSs can impact both individuals and society at large. Indeed, according to Milano et al. (2021), MRSs work as “social planners” in multisided platforms, in the sense that “they direct the flow of information between a multitude of participants on different sides of the platform”. Since in the contemporary digital age online and offline worlds overlap and merge, as Floridi (2014) puts in evidence, the potential impact of RSs as social planners is wider than the context of the platform and therefore needs to be addressed by policies centred on fairness.

### 3 Regulatory gaps and challenges

One might argue that it is unlikely that proprietary algorithms would be spontaneously designed in an ethical way by the platform which owns them. Moreover, since public policies and regulations cannot act on the design of a private intellectual property, but only on the societal consequences of its implementation, the technical structure of RSs may not be regulated through a centralized approach. This argument questions the effectiveness of impact analyses and policy proposals, as their enforcement is not easily viable. In fact, on the one side, the attempts of major digital companies in the direction of self-regulation have rarely been successful, as the experience of Google’s withdrawn AI ethics board can show<sup>2</sup> (Johnson and Lichfield 2019). On the other side, even when a supra-national regulation for AI is developed, as in the case of the recently published draft of the EU Artificial Intelligence Act,<sup>3</sup> the influencing effects of RSs may not be sufficiently taken into account among the risks posed by AI-enabled technologies. In particular, I would like to briefly discuss the potential impact of this new proposal for a regulation on the design and implementation of RSs, to understand whether policymaking could actually be effective in shaping the field of automated recommendations.

In April 2021, the European Commission released the first-of-its-kind regulatory framework for high-risk AI systems, which covers applications ranging “from self-driving cars to hiring decisions, bank lending, school enrolment selections and the scoring of exams” and involves crucial domains such as law enforcement and justice, as the New York Times reports<sup>4</sup> (Satariano 2021). However, it must be noted that, whilst “companies that violate the new regulations could face fines of up to 6 percent of global sales”

<sup>2</sup> <https://www.technologyreview.com/2019/04/06/65905/google-cancels-ateac-ai-ethics-council-what-next/>

<sup>3</sup> Proposal for a Regulation laying down harmonised rules on artificial intelligence (Artificial Intelligence Act) (2021) (<https://ec.europa.eu/newsroom/dae/items/709090>).

<sup>4</sup> <https://www.nytimes.com/2021/04/16/business/artificial-intelligence-regulation.html>

(*ibidem*), the boundaries for the lawful uses of AI are still not traced clearly. This inconsistency is shown by the case of facial recognition, which “shall be prohibited” even for the purpose of law enforcement, “unless and in as far as such use is strictly necessary” for objectives that include “the targeted search for specific potential victims of crime” and “the prevention of a specific, substantial and imminent threat to the life or physical safety of natural persons or of a terrorist attack” (art. 5, para. d)<sup>3</sup>. Both these exemptions to the general ban are not specified further and could lead to an unfair extension of the use of “‘real-time’ remote biometric identification systems in publicly accessible spaces” (*ibidem*).

For what concerns the specific subject analysed in this paper, the challenges posed by RSs are not addressed to the extent that their influence and diffusion would require: in particular, a direct reference to automated recommendations can be found only in two paragraphs of the regulatory proposal. In the first place, the definition of “AI system” refers to “software that can, for a given set of human-defined objectives, generate outputs such as content, predictions, *recommendations*,<sup>5</sup> or decisions influencing the environments they interact with” (art. 3, para. 1)<sup>3</sup>. Secondly, the regulation underlines that the implementation of high-risk AI systems should be overseen by humans to avoid the “automation bias”, which consists in “automatically relying or over-relying on the output produced by [...] AI systems used to provide information or *recommendations*,<sup>5</sup> for decisions to be taken by natural persons” (art. 14, para. 4b)<sup>3</sup>. However, in both the paragraphs, automated recommendations are considered from the perspective of the outcome and not of the process: therefore, they are merely regarded as outputs of an AI system that can have an impact on human decision-making, whilst a specific focus on the design principles of RSs and the risks posed by their biases is completely lacking. In this sense, the regulatory focus lies on the influence of recommendations on people’s decisions, whilst regulators do not suggest specific interventions on the design of RSs to shape the direction of their development.

Given the widespread influence of online targeted advertising based on profiling and automated recommendations (Zuboff 2019), this regulatory gap in the first supra-national proposal to regulate high-risk AI implies either that RSs are not regarded as technologies involving risks by policymakers, or that their ethical and social implications are difficult to address. Since the first option can be considered unlikely, at least from an ethical perspective, this means that policies and regulations are not often able to face the specific challenges posed by RSs, as it has been argued above. In fact, the design and deployment of recommendation technologies

mostly pertain to the commercial domain, which is led by market rules and is not generally considered a sensitive context in which public oversight is (or should be) required. For this reason, even within the advanced regulatory framework provided by the EU, the ethical risks and social challenges posed by RSs still need to be addressed properly: indeed, it might be the case that regulation alone is not sufficient to deal with the influence generated by RSs.

#### 4 Towards the pro-ethical design of recommender systems for social good

Building on this background, I argue that the ethical and societal implications of RSs could be (at least partly) addressed through the application of their influencing potential to good social and political aims. However, since an exhaustive definition of “good social and political aims” cannot be provided in this essay, my argument relies on a common understanding of the label “AI for social good”, which groups the initiatives aimed at using AI to reach sustainable development goals (Vinueza et al. 2020). In fact, although the specific notion of *RSs for social good* has not been proposed in the literature yet, and there are not many examples of the deployment of RSs for socially beneficial purposes,<sup>6</sup> I argue that this framework could be applied also to the domain of algorithmic recommendations. Therefore, I provide a theoretical example to clarify how the influence of online recommendations could be directed towards social good.

Within the framework of MRSs, let us consider a social media platform in which misinformation regarding the safety of vaccines reaches many individuals and affects the acquisition of immunity across a population. It is known that the online spread of fake news is often caused by an effective microtargeting of social media users addressed by advertisements that reach them through RSs. In this context, the architecture of the platform facilitates the activity of malicious agents who exploit its business model to reach their target. As the case of Cambridge Analytica showed, it is very difficult to repress the activity of these malicious agents on case-by-case basis, because they rely on the same system based on data and recommendations that structures the functioning of the platform. Therefore, to address this situation, the system could implement a new interest in the ontology of its RSs. This systemic interest (which is also societal) might consist in presenting public initiatives that contrast misinformation or support scientific divulgation to users who could

<sup>5</sup> Italicized by the author.

<sup>6</sup> To date, digital mental health is one of the few fields in which algorithmic recommendations unrelated to commercial aims have been considered beneficial to the individual.

be exposed to fake news. A similar design mechanism can be found in the banners that alert about potential misinformation on Twitter, or that inform Google users about Covid-19 when one searches keywords related to health and diseases on the web browser.

Although this approach might outline a direction to address the ethical and societal challenges posed by RSs, it still leaves open the question about which incentives a private firm would have in adopting a recommendation policy like the one explained above. In fact, as this policy would imply a change in the structure of the system without bringing any additional profit, a business-focused enterprise may not have compelling reasons to pursue it, apart from a commitment to social good. Moreover, from a user-centred perspective, this approach could be considered paternalistic, as it puts the designer's ethical evaluations above the user's interest. Indeed, a user may prefer to receive recommendations about items or contents which are closer to their expected preferences than to socially preferable outcomes. In this case, a recommendation policy aimed at fostering social good may not be tolerant towards users' attitudes and choices. Therefore, two main objections are raised here: on the one side, the lack of incentives for private companies to modify their RSs may undermine the feasibility of the policy; on the other side, the impact of the policy on users' range of choices may limit their freedom to an even greater extent, because they will be exposed to pre-determined contents that are not linked to their preferences.

To answer these objections, I frame the approach presented above according to an argument put forward by Floridi (2016) about the relationship between toleration and paternalism in the design of digital technologies. He argues that "one form of paternalism, based on pro-ethical design, can be compatible with toleration [...], by operating only at the informational and not at the structural level of a choice architecture". Within this framework, the designer does not aim at orienting the user to de facto pre-determined choices, but he/she forces the user to make a choice before the latter is able to enjoy the service provided by the technology. In the case of RSs, this kind of informational nudging would imply that the system might ask users questions about the contents that are going to be recommended or the categories through which the recommendation is informed.

From the perspective of social good, this approach to RSs has the advantage of making the users aware of the potential implications of their preferences without constraining or biasing the range of contents to which they are exposed. Moreover, as regards the policy proposed above, users might be asked whether they would like to be shown contents and initiatives related to social good alongside standard recommendations based on expected preferences. Therefore, this approach would balance paternalism with toleration through enhancing users' awareness without limiting their freedom.

Furthermore, pro-ethical design applied to MRSs can help us address the question about incentives for private firms. In fact, informational nudging allows companies to gather data about users' explicit preferences, thereby making recommendations more targeted and precise. A similar method has been implemented by Spotify, whose RSs are designed combining exploitation with exploration and explicability (McInerney et al. 2018). In fact, explicit feedback from users improves the preference elicitation process, thereby allowing RSs (and the firms owning them) to get more data about users' interests and behaviour.

## 5 Conversational recommender systems and the problem of explanation

This process is particularly evident in conversational recommender systems (CRSs), which "allow users to elaborate their requirements over the course of an extended dialogue [...] rather than each user interaction being treated independently of previous history" (Tintarev and Masthoff 2015). Jannach et al. (2020) define CRS as a "software system that supports its users in achieving recommendation-related goals through a multi-turn dialogue". Throughout this dialogue, "the system can [...] elicit the detailed and current preferences of the user, provide explanations for the item suggestions, or process feedback by users on the made suggestions" (*ibidem*). CRSs are therefore particularly useful to address the problem of explanation, which is closely related to the pro-ethical design of RSs: indeed, if a user wonders why they have been exposed to a particular recommendation, the answer may not be straightforward.





Early research on explicability for RSs dates back to 1990s, but one of the first comprehensive surveys about the different kinds of explanations is provided by Tintarev and Masthoff (2007), who outline a common problem faced by users of services based on automated recommendations: that is, understanding why they have been recommended items which are irrelevant to their interests and attitudes. In particular, the authors mention a case reported by the Wall Street Journal in 2002<sup>7</sup>: a digital video recorder, named TiVo, recorded automatically programmes it assumed its owner would like, based on shows he/she had chosen to record previously (Zaslow 2002). However, TiVo frequently mislabelled the users' characteristics extracted from their recording history and ended up recording videos whose irrelevance upset the owners.

Almost two decades separate that rudimental offline RS and the contemporary personalized advertisements on online platforms, but many recommendations are still irrelevant to

<sup>7</sup> <https://www.wsj.com/articles/SB1038261936872356908>



**Table 2** Comparison of conversational and standard RSs from the perspective of their impact on the different stakeholders

	Conversational recommendations	Standard recommendations
 <b>Users</b>	<ol style="list-style-type: none"> <li>1. Improved relevance derived from users' explicit input and feedback</li> <li>2. Increased interpretability and transparency as explanations are embedded by design in the system</li> </ol>	<ol style="list-style-type: none"> <li>1. Focus on accuracy metrics and exploitative feedback effects</li> <li>2. Problems with explicability due to black-box models, proprietary constraints, and lack of direct user-system interaction</li> </ol>
 <b>Providers</b>	<ol style="list-style-type: none"> <li>1. Incentive for micro-targeted ads as individual users provide explicit ready-made personal data on preferences and interests</li> <li>2. Potential higher profitability (as users are more likely to be interested in and eventually buy the products advertised)</li> </ol>	<ol style="list-style-type: none"> <li>1. Targeting often relying on low-quality implicit data</li> <li>2. Higher risk of irrelevant or repetitive ads that fail to increase product sales</li> </ol>
 <b>System</b>	<ol style="list-style-type: none"> <li>1. Profiling based on explicit data voluntarily provided by users (less potential privacy breaches)</li> <li>2. Increased likelihood of diversifying the recommendation policies thanks to more granular data</li> </ol>	<ol style="list-style-type: none"> <li>1. Profiling often based on implicit data such as digital traces, click-through rates and browsing history</li> <li>2. Increased likelihood of keeping the same recommendation policy (often exploitation)</li> </ol>
 <b>Society</b>	<ol style="list-style-type: none"> <li>1. Informational nudging both ex ante and ex post</li> <li>2. Increased individual and collective awareness of the socio-technical structure and ethical implications of the process thanks to dialogic explanations</li> </ol>	<ol style="list-style-type: none"> <li>1. Only ex post informational nudging</li> <li>2. Limited understanding of the distribution of the interests at stake and its connection with the structure of the system (unaccountable social influence)</li> </ol>

the users, or they may focus on a narrow subset of their interests. Explanations implemented within the recommendation process would allow users to be aware of the reasons why particular items are brought to their attention and would therefore increase their trust in the platform (O'Donovan and Smyth 2005). The different ways through which a recommendation can be explained depend on how it is generated: explanation styles, which are determined by the chosen explanatory goal, can range from content-based and knowledge/utility-based to nearest-neighbour (also called “collaborative”) approaches, as Tintarev and Masthoff (2015) outline.

However, I argue that the conversational approach to RSs is one of the most appropriate to explain how an ongoing recommendation process unfolds. In particular, CRSs are often implemented as AI-enabled chatbots that ask directly to the user which are their interests and, consequently, what the recommendation should focus on. Unlike other explanatory approaches, which may rely on a fixed underlying algorithm, CRSs can be considered “more of an interaction style than a specific algorithm” (*ibidem*). In fact, the structure of CRSs allows an interactive relationship between the user and the recommendation process, which is informed by the words produced throughout the chat. According to Wärnestål (2005), CRSs are useful “to make interaction efficient and natural, to acquire preferences from the user in a context when she is motivated to give them” and also “to facilitate exploration of the domain and the development of the user's preferences”. In particular, the conversational context can promote an exploratory approach to recommendations:

indeed, “users can learn about new items and concepts [...] in an incremental fashion throughout the [...] dialogue and explore the domain, and, as a result, evolve their own preferences within it” (*ibidem*).

## 6 Informational nudging through conversational recommender systems: applications, implications, advantages

For these reasons, CRSs can be conceived as models that integrate recommendations with the related explanations in a single dialogic stream. The advantages coming from their implementation are threefold: firstly, they are more likely to recommend items that the user is actively looking for; secondly, the explicit feedback given by users through natural language allows a more nuanced analysis of their interests, thereby representing an incentive for platforms; thirdly, the fact that the explanation is embedded within the recommendation process allows users to understand better the reason for which certain products are recommended to them. Therefore, from the perspective of the MRS model, the conversational recommendation process can account for the interests of different stakeholders: users' interest for both transparency and relevance; the systemic interest in acquiring explicit data directly from users; the providers' interest to find potential customers that are probably more interested in the products they advertise (see Table 2). These peculiarities make CRS a suitable technical infrastructure

for implementing pro-ethical design: in fact, within the framework of a human–machine conversation that is not pre-determined, the informational nudging can occur more naturally than in other recommendation processes. This is because CRSs are flexible and responsive to the interaction style of the single user, which may change on a case-by-case basis: therefore, in this case, the effects of pro-ethical design can be better targeted to the individual who is chatting with the system.

A theoretical example might be useful to underline how the implications of pro-ethically designed CRSs are different from those of standard RSs: indeed, this model can outline how informational nudging through conversational recommendations may be used to foster social good, thereby addressing the ethical challenges posed by the widespread implementation of RSs in online platforms. Therefore, let us consider the case of a news aggregator that uses automated recommendations to show specific content on each user's personalized feed. Assume that a user named Alice tends to view articles coming from news sources that are notoriously regarded as channels for misinformation. If the RS implemented on the platform relies on a collaborative-based approach, the algorithm will show articles following the criterion usually explained through the expression: “users who read this article also read...”. If Alice has already been exposed to fake news, this nearest-neighbour approach is likely to increase the number of fake news she reads, thereby engaging the user in a so-called filter-bubble.<sup>8</sup>

The application of informational nudging, in this case, would consist in asking the user whether a certain news category associated with misinformation which they have been exposed to really fits their interest. Alternatively, an alert or banner can pop up to warn the user that the content they are viewing is highly correlated with fake news. However, if Alice were deeply engaged in a filter bubble or very attracted to fake news, this one-time informational nudging would not probably increase her critical awareness about her news consumption behaviour. In fact, the platform has no incentives to show too many alerts or pose too many questions to Alice before she is allowed to enjoy the service, as she may decide to switch to another news provider otherwise. Moreover, the design of collaborative-based RSs implies that the interface can be personalized only after the user has viewed at least a few contents: therefore, in this situation, informational nudging could occur only *ex post*, i.e. after the user has already been exposed to misinformation. Indeed, there would be no sense in warning the user about fake news which they are not interested in or going to view. These implications put in

evidence that, in the context of collaborative-based recommendations, the effect of pro-ethical design on users' choices might be limited.

Consider instead the same case contextualized in a platform relying on CRSs: in this situation, Alice would need to engage with a chatbot before she could view personalized contents. In fact, the personalization can only take place if the users specify their requirements, as the feed would just show random articles otherwise. This is because CRSs are interactive and explanatory by design: therefore, if they are the only kind of RS implemented in the platform, the user needs to interact with them directly to be able to get a personalized feed. Hence, in this case, the initial setup of the interface requires an explicit feedback from Alice. Then, as Alice engages in the chat with the CRS, the questions she may have to answer will inform the categories of articles that may be recommended to her. This critical interaction takes place before the existence of personalization, rather than after that some contents have already been proposed based on what the user has already viewed (*ex ante* versus *ex post* approach): this chronological inversion is a crucial advantage of CRSs and can enhance the utility of pro-ethical design.

In fact, if Alice expresses explicitly her preliminary interest in articles that are typically associated with fake news, then informational nudging can take place in a more targeted and effective way, because it will anticipate the actual exposure to misinformation. Indeed, in this case, Alice would be informed that a large proportion of articles pertaining to the category which she is interested in are related to misinformation: consequently, she might be asked whether she really wants to view such content. In this way, the platform based on CRSs can expose the users to multiple layers of informational nudging, as the choices they are required to make in interacting with the chatbot are preliminary but compulsory passages that must happen before the feed can be personalized.

Moreover, pro-ethical design implemented in CRSs allows a combination of *ex ante* and *ex post* nudging approaches: in particular, whilst users' preliminary critical evaluation is incentivized by the dialogue with the system, informational nudging can happen also after the user has viewed some articles, as the in the case outlined above. For this reason, I argue that the form of tolerant paternalism embedded in the structure of pro-ethically designed CRSs leaves freedom of choice to the user whilst effectively enhancing their critical awareness, which can be fostered both *ex ante* and *ex post* through an interactive conversation between the user and the system. Therefore, from a multi-stakeholder perspective, the conversational recommendation process has a beneficial impact on society, as it can help prevent phenomena that affect the contemporary information age, such as the spread of misinformation.

<sup>8</sup> For an analysis of the relationship between filter bubbles and diversity in RSs, see Nguyen et al. (2014).

## 7 Conclusion

In conclusion, I think that a recommendation policy informed by pro-ethical design would improve the implementation of the MRSs framework in the direction of social interest. In particular, from an applicative perspective, I showed that CRSs are among the best multi-stakeholder recommendation technologies that can foster RSs potential for social good through targeted informational nudging. I analysed the incentives that private firms would have in adopting CRSs, to demonstrate that their currently limited diffusion cannot be ascribed to a lack of technical or economic feasibility. Therefore, I argue that the application of the MRS model to pro-ethically designed CRSs can address, at least partly, the ethical challenges and the societal opportunities brought about by the widespread implementation of automated recommendations in online platforms. In fact, within this framework, the systemic interests of platforms could be translated to the wider interests of users and society as a result of the ontological differentiation of and interaction among different stakeholders in the recommendation process. Therefore, the approach based on MRSs and pro-ethical design applied to CRSs has the potential of being integrated into many different domains in which people's decisions are influenced by algorithmic recommendations: from e-commerce to information filtering, from culture to health, human choices could be directed to ethical purposes besides profit.

**Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

## References

Bratton BH (2016) *The stack: on software and sovereignty*. MIT press  
 Floridi L (2014) *The fourth revolution: how the infosphere is reshaping human reality*. OUP Oxford

- Floridi L (2016) Tolerant paternalism: pro-ethical design as a resolution of the dilemma of toleration. *Sci Eng Ethics* 22(6):1669–1688
- Jannach D, Manzoor A, Cai W, Chen L (2020) A survey on conversational recommender systems. arXiv preprint [arXiv:2004.00646](https://arxiv.org/abs/2004.00646)
- Johnson B, Lichfield G (2019) Hey Google, sorry you lost your ethics council, so we made one for you. MIT Technology Review. Retrieved on 19 Jan 2021 from: <https://www.technologyreview.com/2019/04/06/65905/google-cancels-ateac-ai-ethics-council-what-next/>
- McInerney J, Lacker B, Hansen S, Higley K, Bouchard H, Gruson A, Mehrotra R (2018) Explore, exploit, and explain: personalizing explainable recommendations with bandits. In Proceedings of the 12th ACM conference on recommender systems. 31–39
- Milano S, Taddeo M, Floridi L (2020) Recommender systems and their ethical challenges. *AI & Soc* 35(4):957–967
- Milano S, Taddeo M, Floridi L (2021) Ethical aspects of multi-stakeholder recommendation systems. *Inf Soc* 37(1):35–45
- Nguyen TT, Hui PM, Harper FM, Terveen L, Konstan JA (2014) Exploring the filter bubble: the effect of using recommender systems on content diversity. In Proceedings of the 23rd international conference on World wide web. 677–686
- O'Donovan J, Smyth B (2005) Trust in recommender systems. In Proceedings of the 10th international conference on Intelligent user interfaces. 167–174
- Proposal for a Regulation laying down harmonised rules on artificial intelligence (Artificial Intelligence Act) (2021) Retrieved on 08 May 2021 from: <https://ec.europa.eu/newsroom/dae/items/709090>
- Satariano A (2021) Europe Proposes Strict Rules for Artificial Intelligence. *New York Times*. Retrieved on 08 May 2021 from: <https://www.nytimes.com/2021/04/16/business/artificial-intelligence-regulation.html>
- Tintarev N, Masthoff J (2007) A survey of explanations in recommender systems. In 2007 IEEE 23rd international conference on data engineering workshop. IEEE. pp. 801–810
- Tintarev N, Masthoff J (2015) Explaining recommendations: design and evaluation. *Recommender systems handbook*. Springer, Boston, pp 353–382
- Vinuesa R, Azizpour H, Leite I et al (2020) The role of artificial intelligence in achieving the sustainable development goals. *Nat Commun* 11(1):1–10
- Wärnestål P (2005) User evaluation of a conversational recommender system. In Proceedings of the 4th Workshop on Knowledge and Reasoning in Practical Dialogue Systems
- Zaslow J (2002) If TiVo thinks you are gay, here's how to set it straight. *Wall Street J*. Retrieved on 08 Jan 2021 from: <https://www.wsj.com/articles/SB1038261936872356908>
- Zuboff S (2019) *The age of surveillance capitalism*. PublicAffairs, New York

**Publisher's Note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.