Topoi
https://doi.org/10.1007/s11245-023-09922-5

Artificial Intelligence and Autonomy: On the Ethical Dimension of Recommender Systems

Sofia Bonicalzi[1,2] · **Mario De Caro**[1,3] · **Benedetta Giovanola**[3,4]

Accepted: 11 April 2023

Abstract
Feasting on a plethora of social media platforms, news aggregators, and online marketplaces, recommender systems (RSs) are spreading pervasively throughout our daily online activities. Over the years, a host of ethical issues have been associated with the diffusion of RSs and the tracking and monitoring of users' data. Here, we focus on the impact RSs may have on personal autonomy as the most elusive among the often-cited sources of grievance and public outcry. On the grounds of a philosophically nuanced notion of autonomy, we illustrate three specific reasons why RSs may limit or compromise it: the threat of manipulation and deception associated with RSs; the RSs' power to reshape users' personal identity; the impact of RSs on knowledge and critical thinking. In our view, however, notwithstanding these legitimate concerns, RSs may effectively help users to navigate an otherwise overwhelming landscape. Our perspective, therefore, is not to be intended as a bulwark to protect the *status quo* but as an invitation to carefully weigh these aspects in the design of ethically oriented RSs.

Keywords Artificial intelligence · Autonomy · Identity · Reasons-responsiveness · Recommender systems · Reflective-endorsement

1 Recommender Systems and Two Forms of Autonomy

Feasting on a plethora of social media platforms, news aggregators, and online marketplaces, recommender systems (henceforth RSs) are spreading pervasively throughout our daily online activities. While being a more and more central element of the revenue models of several technology-based private companies, RSs are also progressively expanding beyond the traditional commercial and entertaining sphere, up to politics and other crucial social and cultural sectors (Papakyriakopoulos et al. 2020; Botes 2023).

✉ Sofia Bonicalzi
sofia.bonicalzi@uniroma3.it

[1] Department of Philosophy, Communication, and Performing Arts, Roma Tre University, Rome, Italy

[2] CVBE Cognition, Values, Behaviour, Ludwig-Maximilians-Universität München, Munich, Germany

[3] Department of Philosophy, Tufts University, Medford, USA

[4] Department of Political Sciences, Communication, and International Relations, University of Macerata, Macerata, Italy

Simply put, RSs are algorithms based on artificial intelligence (AI)—mostly on machine learning techniques—that support user-tailored decision-making by providing suggestions out of a wider catalog, i.e., about news, videos, advertisements, or exercises, based on the users' or like-minded users' past choices or personal information (Aggarwal 2016; Jannach et al. 2010). On the one hand, by making effective recommendations, RSs may help users navigate the environment by alleviating choice overload and decision fatigue. Indeed, even without invoking the highly disputed *paradox of choice* (Schwartz 2016), users equipped with bounded rationality and limited cognitive resources may find it easier to make decisions aligned with their interests and goals when irrelevant options are filtered out (Bollen et al. 2010). On the other hand, in virtue of their reach of impact and influence, RSs may reshape, to a large extent, the media landscape and the social dimension in which users interact and, consequently, affect the way people think and choose. So, while RSs' short-term benefits are under everyone's eyes, the medium and long-term effects of their pervasive influence are less predictable and more ethically controversial.

Over the years, a host of ethical issues have been associated with the diffusion of AI-based algorithms and especially with RSs, and specific attention has been paid to the

Published online: 18 May 2023

tracking and monitoring of users' data (Milano et al. 2020). In particular, growing ethical concerns have been raised with regard to the RSs' role in the diffusion of inappropriate content (Yesilada and Lewandowsky 2022); breach of privacy and data protection laws, extending to the selling of personal information to third parties (Himeur et al. 2022); opacity in how recommendations are generated due to the complexity or even secrecy of the underlying mathematical models (Herlocker et al. 2000), with the connected problems of accountability (Diakopoulos and Koliska 2017); lack of fairness and biases in how data is sampled and used to shape recommendations (Carraro and Bridge 2022); democracy-damaging social effects, with pernicious phenomena the like of filter bubbles, polarization, and echo chambers (Cinelli et al. 2021; Helberger 2019; Pariser 2011); violation of personal autonomy and identity (Botes 2023; Burr et al. 2018; Klenk and Hancock 2019).

Here, we focus on the impact that RSs may have on personal autonomy as the most elusive among these often-cited sources of grievance and public outcry. In both everyday speech and philosophical theorizing, autonomy loss is linked with episodes of coercion or brainwashing, and with conditions of physical or mental impairment (Varkey 2021). Compared to these dreaded threats, however, it is not clear whether technology-mediated recommendations may actually work as autonomy-defeater mechanisms of soft power or are the mere online equivalent of the more innocuous word of mouth (Herlocker et al. 2000). To the extent that recommendations nudge people to make decisions that are allegedly more aligned with their true selves and interests (Thaler and Sunstein 2009), as in the case of health recommender systems (De Croon et al. 2021; Tiribelli 2023), they might even arguably enhance autonomy as self-governance and proneness to respond to evidence (Levy 2017). Or, by making different platforms more responsive to the users' interests and urging the media to produce news that tracks diverse information needs, new technologies may provide users with more effective and tailored tools to deal with the online and offline reality (Helberger 2019). At the same time, however, RSs are routinely viewed with suspicion as they are seemingly able to bypass users' cognitive defenses and steer their behaviors towards results that ultimately mirror third-party goals more than their benefit (Burr et al. 2018).

With the present paper, we aim to contribute to the discussion as to whether personal autonomy is threatened by the deployment of RSs. Primarily, we propose to distinguish between two overarching understandings of autonomy that should not be conflated and which we will term *descriptive autonomy* (henceforth DA), which is the focus of the present paper, and *normative autonomy* (henceforth NA). DA refers to a host of features that one must fulfill to be, or to act as, an autonomous agent. In the relevant literature, these features are variably identified in an *internalist* fashion (e.g., self-expression through one's actions) or an *externalist* one (e.g., reasons-responsiveness and reflective thinking) (for more details, see Sect. 2).

NA identifies a key principle of ethics i.e., the right, of Kantian and Millian flavor, that agents retain to be treated as autonomous individuals (Varkey 2021). This means that they ought to be seen as ends rather than as means, as Kant would put it, or, to put it in terms that also a utilitarian philosopher would accept, as moral subjects with a right to choose about their own life and whose interests and preferences ought not to be arbitrarily overridden or interfered with. Based on NA, we ought to treat users "as individuals capable of making informed decisions" (Keeling 2018) rather than as exploitable consumers. Respect for NA is a binding value enshrined in various legislative tools, including the General Data Protection Regulation (GDPR), as well as a key principle of biomedical ethics (Beauchamp and Childress 2001; Gillon 1994; Varkey 2021).

Fulfilling the conditions for DA, in terms of general capacities, is necessary to be entitled to NA. However, when it comes to individual actions, DA might be occasionally violated without individuals ceasing to be entitled to NA. For instance, when users are effectively prevented from making informed decisions by being targeted by fake news, their DA might be violated without them losing their NA. Or, conversely, NA might be violated (e.g., when a website is purposely designed to pick on users with a misleading advertising campaign) without DA being truly threatened (e.g., because experienced users are by now decently able not to fall for the tricks of a website). To offer another example, in discussions about nudging, it is doubtful whether helping people to align actions to their preferences, which may *prima facie* seem respectful of NA since no deception is involved, violates the conditions for DA as the ability to express one's own *self* through one's actions (Keeling 2018). Overall, failure to distinguish between whether RSs violate DA or NA, not clarifying the terms of this double-edged challenge, makes discussion inevitably unsatisfactory. In this paper, our focus is mainly on how RSs interfere with our fulfilling the requirements for DA, although in ways that may help explain why RSs may violate NA as a right to be treated as an end and not as an exploitable consumer.

We commence (in Sect. 2) by providing an overview of what the notion of DA consists of according to some of the most well-known philosophical views on the matter. Depending on the specific feature of autonomy that is deemed central within a theory, there are multiple ways in which RSs may raise challenging concerns. Then, we proceed to briefly illustrate how RSs effectively work in practice and highlight that they may hinder DA in various ways (Sect. 3). More specifically, we identify (in Sect. 4) three major sources of ethical worry and discuss their impact on DA: the threat of manipulation and deception associated

with RSs (Sect. 4.1); the RSs' power to reshape users' personal identity (Sect. 4.2); the effect of RSs on users' knowledge and critical thinking (Sect. 4.3).

In discussing these three challenges, however, one must pay attention to some pitfalls that may plague discussions of DA. In particular, one of the risks is that of taking refuge in an idealized notion of DA inspired by commonsense and, to some extent, philosophical theorizing, but that can be proven untenable, especially in consideration of the findings of contemporary cognitive sciences. In fact, decades of empirical literature have shown that our *autonomous self* is, by its nature, much more penetrable than one would ordinarily assume and that human decision-making is intrinsically context dependent. In this respect, evidence on cognitive manipulation and biases suggests that even apparently rational, *self*-defining behaviors may be driven by environmental inputs and automatisms bypassing conscious awareness (Bargh and Chartrand 1999; Doris 2015; Wegner 2002). On the one hand, this consideration warns about the worrisome easiness with which potentially *malevolent* RSs may take advantage of people's cognitive shortcomings (Matz et al. 2017). On the other hand, this implies that it would be naïve to conceive of the self as tendentially insulated from external influence: risk talk about the autonomy-defeating role of RSs ought to be grounded in a philosophically sound but also empirically sensitive notion of DA.

2 Philosophical Frameworks of Descriptive Autonomy

As mentioned in Sect. 1, DA refers to a host of conditions that agents must fulfill to be autonomous or to act autonomously. Although it can be pre-theoretically understood as self-governance or governance through one's own will, DA is a multi-dimensional philosophical concept. Different accounts have indeed considered different features as plausibly central to DA and, correspondingly, have identified as many DA-defeater conditions. It would be implausible to recount all the theories of DA in a single paper, but we can offer general indications drawing on some of the most discussed accounts of it (for analogous taxonomies, see Bonicalzi 2019; Burr et al. 2018; Buss and Westlund 2018). First, one should distinguish between views of DA of a compatibilist nature (i.e., DA is compatible with determinism being true) and of an incompatibilist nature (i.e., DA is incompatible with determinism being true). The incompatibilist view presupposes, as a necessary but not sufficient condition for DA, that agents are acting autonomously when their will is not fully determined by the combination of previous events and the laws of nature (Bonicalzi 2019; Mele 2001). Here, however, we put this view aside to the extent that it is unclear whether this condition can be met and because usually this debate is scarcely interested in the specific mechanisms that may work as DA-defeaters.

Conversely, various compatibilist views articulate a rich array of ways in which agents can act autonomously or not in a world that is possibly, but not necessarily, ruled by deterministic laws (Bonicalzi 2019; Mele 2001). The supporters of these alternative compatibilist views are usually arguing one against the other in individuating different conditions as allegedly necessary for DA. Here we take a more neutral perspective, screening off the details and seeing these views as individually capturing various key aspects of what DA must consist of. A most basic distinction in this respect is that between *internalist* and *externalist* views (Bonicalzi 2019). The heart of internalist or *coherentist* (Buss and Westlund 2018) views is that DA is related to the agent's acting out of her true and self-defining motives, as opposed to some external driving forces that the agent does not recognize as her own. Adopting this internalist view of DA, as a form of reflective endorsement, there is an overlap (discussed in sect. 4.2) between the notion of DA and that of personal identity to the extent that such self-defining motives are plausibly central also to the agent's identity.

Different internalist accounts vary in the identification of what the relevant motivating states, which should coherently fit within the agent's global psychological mindset, must be. Famously, Frankfurt insists that these must be desires (1971), independently of how rational or irrational they may look from a third-party point of view. Instead, others have discussed the prominent role of evaluative judgments (Watson 1975) or long-term plans (Bratman 2018) as defining elements of people's psychological makeup. Other strong defenses of internalist views have then been put together by authors the like of Arpaly and Schroeder (1999), Sripada (2017), and Talbert (2017).

Overall, internalist accounts have some notorious drawbacks, for instance concerning the problematic notion of *identification* with one's motives (Arpaly and Schroeder 1999). And, more crucially for the contrast with externalist views, internalist positions usually end up biting the bullet regarding cases of manipulation (consider Mele's "Ann & Beth case" (Mele 2013)) or, at least in the case of Frankfurt, addiction: to provide a known example, assuming that the content of the agent's motives is irrelevant to circumscribe DA as long as the agent wills to be guided by them, even the "unwilling addict" (Frankfurt 1971) may be deemed autonomous. Implausible as this last point may appear, these views deserve merit for capturing the idea that DA has to do with self-expression, as opposed to acting out of external motives about which one feels alienated. In this respect, internalist views of DA have some commonalities with more empirically oriented definitions of autonomy whereby autonomous actions are conceived of in terms of internally generated

changes of state, as opposed to habitual or environmentally triggered ones (Bonicalzi and Haggard 2019).

Conversely, externalist views of DA suggest that some intersubjective (this is the sense of the adjective "externalist" in this context) validation must be in place for agents to act autonomously. In particular, *reasons-responsiveness views*, such as those articulated by Fischer and Ravizza (1998), Wolf (1990), and Nelkin (2011) tend to define as non-autonomous cases of patent irrationality, manipulation, or lack of basic moral knowledge, especially when this is due to the agent's troubled causal history. Fischer and Ravizza, for instance, suggest that agents act autonomously when they have a grasp of the objectively relevant reasons for acting in one way or another so that they exert a form of control over the ensuing action (1998). This presumably requires the ability to monitor the state of the environment and calculate the likely consequences, as well as the costs and benefits, of a set of possible actions.

From this perspective, in picking an option, autonomous agents are those who can govern themselves in a reasonable manner, which is responsive to the agent's *true* interest—as opposed to the agent's true self discussed by internalist views. In this respect, an agent acting out of motives that she believes are conducive to rational behaviors but in fact are not—paradigmatic cases being Susan Wolf's evil *Jojo* (Wolf 1990) or Watson's *Robert Harris* (Watson 1987)—would likely count as a violation of DA. Externalist views can rule out borderline cases better than internalist views, and preserve the intuitive idea that autonomous agents can exert their self-control in a way that seems reasonable from the outside as well as from the inside. However, by introducing a distinction between autonomy-satisfying and autonomy-defeating causal histories, they might find it difficult to distinguish in non-arbitrary ways what conditions count as autonomy defeaters and what as regular causal factors affecting people's behavior (Pereboom 2014).

Starting from manipulation cases, one of the cruces of different views of DA consists in how to cogently define DA-defeater conditions in non-ad hoc ways (Bonicalzi 2019; Mele 2001; Pereboom 2014). While this is the mainstay of the debate on DA, in this context we have no stake in arguing in favor of one or the other view. Although at odds in how they approach borderline cases, both internalist and externalist views pick relevant elements or components of what DA as global self-governance necessarily implies in more ordinary cases, i.e., acting out of motives one can recognize as her own, behaving in ways that are responsive to commonly accepted reasons for actions, and exerting some reflective thinking, eventually turning away from motives about which one is decided to take distance. Intermediate or comprehensive positions, including those defended by Mele (2001) and Christman (1991), already propose a balance of internalist and externalist factors whereby DA depends both on reflective endorsement and critical thinking as the ability to put one's motives into perspective and eventually change one's mind.

A key question remains, however, whether all or some of these criteria are not just necessary but also jointly sufficient for DA. This is doubtful to the extent that, notwithstanding the differences between internalist and externalist criteria, all these views look almost exclusively at the sphere of individual agency, decision-making, and self-control *in the abstract*. This rather atomistic, competence-focused, approach overlooks the centrality, championed by more *socially situated* or *relational* views of DA, of concrete agential opportunities to act under various constraints. These opportunities are offered or hindered by the social and political environment in which one happens to live and act and ought not to be disregarded in discussing DA (Mackenzie and Stoljar 2000; Oshana 2006; Stoljar 2017). Although in the present article we will not explore this relational dimension in detail, in Sects. 3 and 4 we will point out that the concrete ways in which RSs are designed have a specific impact on DA, as far as they significantly contribute to shaping the environment in which we live and act.

Furthermore—as we will argue more extensively in Sect. 4—both internalist and externalist notions of DA risk suffering from other forms of idealization and, for this reason, require some amendments. Concerning internalist DA, research in cognitive sciences has shown that even allegedly self-defining features are influenced by the (social) context, so that they may express themselves or not depending on the inputs one receives from the environment. As for externalist DA, the well-known framework of bounded rationality, imported from behavioral economics (Simon 1957), provides a more realistic picture of the limitations affecting people's ability to model the environment, compute probabilities, remember different opportunities, predict the probabilistic outcome of each action, and finally make *rational* decisions.

3 A Description of the Fundamental Cogs of Recommender Systems

On the grounds of the philosophical framework of the DA-defining conditions outlined in Sect. 2, in sect. 4 we will discuss whether RSs may represent a source of threat to DA. To do so, a brief explanation of what RSs consist of and how they work is in order.

RSs are machine learning systems with a primary aim: filtering information and using them to make recommendations, so that people, individually or as a group (Pérez-Almaguer et al. 2021), are not bombarded by too many irrelevant options and can choose more effectively or have a more satisfactory experience with the system (Aggarwal

2016; Jannach et al. 2010). This recommending service promises to be particularly precious in the online landscape where more and more choices have to be made, and where users tend to make fast and rather automated decisions without considering the details of the various options (Mirsch et al. 2017).

By making recommendations (i.e., proposals regarding what users may like), conventional RSs, strictly speaking at least, are not to be cataloged as *persuasive technologies*—the aim of the latter being to actively bring people to like an item, guide them towards buying a given product, or convince them to engage in a certain activity, often independently of their original taste or preference (Fogg 2002; also, see below here Sect. 4.1 on the distinction between recommendation and persuasion). However, in the last few years in particular, the borders between RSs and persuasive technologies have become more and more blurred. Indeed, attempts have been made to implement persuasive messages within RSs and to leverage heuristic shortcuts to enhance the likelihood that users engage with the sponsored product or activity rather than opting out (Alslaity and Tran 2019; Yoo et al. 2012).

Fine-tuning personalized recommendations, iteratively trained with users' data to optimize outputs, is among the key strategies RSs deploy to increase engagement (Matz et al. 2017). To handle and personalize recommendations or feed "homophilic clusters" of users (Cinelli et al. 2021), RSs rely on a variety of methods the most widespread of which are *collaborative filtering* and *content-based* methods, which are more and more coupled into hybrid recommendation approaches (Aggarwal 2016).

Collaborative filtering methods are designed to formulate recommendations based exclusively on multiple users' past decisions and ratings. For instance, news RSs may promote stories that other users have spent time reading in the past. RSs using collaborative filtering may proceed in a kind of model-free way, as with memory-based recommendation systems that deploy the *nearest neighbor search* or more sophisticated techniques to find the closest item or user and suggest similar items or items that similar users have previously selected (Su et al. 2022). Or they may rely on more complex and faster model-based approaches that deploy the extracted information to build a model of the interaction and formulate predictions of what similar users may like.

Such intelligent systems are designed to approximate (an always revisable) model of the human mind, compute the probability that a given course of action is chosen, and propose options that can both meet the users' preferences and, in many cases, generate revenues for the company that runs the RS (Aggarwal 2016). Concretely, decisions about what advertisements to display—say, on a web page—depend on the weighing of multiple variables, such as the likelihood that the user will select that product and the expected income in case of purchase. The users' response, e.g., in terms of clicks or time spent on a platform, is a central cog of the feedback loop through which the system updates its model of the environment (Burr et al. 2018).

Both in the case of memory-based and model-based approaches, the underlying assumption is that if different users exhibited a certain decision-making pattern in the past, they are likely to have similar tastes or needs in the future. In this way, feedbacks and usage history are interpreted as signals to adjust the user's evolving psychological profile (Ilievski and Roy 2013), without the need to explicitly ask people about their preferences or even to have a clear grasp of the features of the recommended item. Simply by taking the action that the RS suggests, the user provides learning opportunities to the system.

Instead, content-based methods formulate predictions relying on extra available information, such as the anagraphic data users have provided or the actual features of a given item, based on which different users are matched with different product categories (Jannach et al. 2010). For instance, the user's anagraphic data or past ratings can be matched with the technical features of a movie, e.g., length, cast, and year of production, to map the user's preferences and provide adequate suggestions that are similar to what she has chosen in the past. Overall, the algorithmic model structuring otherwise different RSs tends to maximize predictive accuracy to increase the likelihood that users will choose a given content, while other aspects of user experience, such as general satisfaction, novelty, or availability of alternative options, are often neglected (Raza and Ding 2022).

Collaborative filtering has the advantage of gathering information simply by observing the users' habits, but notoriously suffers from the so-called *cold start* problem (Lika et al. 2014): when the system does not have access to enough information—for instance because too few interactions have been registered or because the user has just subscribed to a given platform—the resulting model is quite poor since the system needs large data sets to work properly. Conversely, content-based methods do not need to rely on repeated interactions but require a consistent flow of new items (e.g., songs, news, videos) to keep up with the users' preferences and provide appreciated solutions. Sophisticated RSs, including the algorithm used by technological giants such as Netflix or Amazon (Pawlicka et al. 2021), have thus implemented hybrid methods, which combine the functioning and overcome the respective weaknesses of collaborative filtering and content-based methods.

In situations where both collaborative filtering and content-based methods would fail because of the low number of interactions or items, RSs may also alternatively rely on *knowledge-based* approaches. Knowledge-based RSs depend on the information explicitly provided by the users and directly asked by the system (Aggarwal

2016). For instance, a rental website may provide suggestions based on the users' explicitly provided information about the maximum price or the minimum size of a property.

Depending on their purpose, different RSs have also more specific challenges to meet. For instance, news RSs, implemented in news aggregators and social media platforms, must deal with timeliness, quality control, limited lifespan of news items, growing abundance of data, and abrupt changes in the users' interests in ways that are uncommon to the RSs active in the e-commerce or the entertainment world (Raza and Ding 2022).

When they work properly, these ubiquitous digital information acquisition strategies are designed to allow smooth interactions between the user and the RS by capitalizing on our "digital footprints" (Matz et al. 2017). RSs hit their target by virtue of the system's ability to collect the data that users provide—implicitly (e.g., time spent on an item, sharing behaviors) or explicitly (e.g., ratings)—and organize inputs in ways that are particularly apt to raise their interest (Susser et al. 2019). The RS of a most enticing and addictive social network like TikTok, for instance, is built to capture attention and stimulate engagement by predicting the user's preferences and promoting videos that similar users have liked. Entering the platform, users are submerged by a flow of short videos without having made any direct choice to watch them and with no control over the contents they are exposed to (Qin et al. 2022). This by itself represents a departure from more traditional social media platforms, such as Facebook or Instagram, where the front page is occupied by the user's profile and the biographical information she has willingly shared.

In principle, the more a system is allowed to learn the more it should present the user options that are aligned with her actual preferences and interests. It goes without saying, however, that the system may fail to distinguish people's *real* interests from their tendency to addictively indulge in watching more content, potentially leading to disempowering feelings of regret or anger (Altuwairiq, Jiang, & Ali 2019). The system may even misinterpret *why* the user has made a certain choice, thus providing irrelevant feedback, as anyone who had looked for an item and is then presented with the same content over and over after a single purchase has likely experienced—a machine learning problem that is known as *reward hacking* and is a type of misalignment between the optimization of the algorithm's reward function and the user's goal (Cohen et al. 2022). Besides these practical concerns, however, the implementation of RSs on a large scale—which has been sometimes associated with a form of "technological paternalism" (Spahn 2012)—raises some specific issues for DA, as we will discuss in Sect. 4.

4 Three Reasons why Recommender Systems may Limit or Compromise Descriptive Autonomy

The philosophical accounts of DA presented in sect. 2 provide a useful framework for evaluating the threat posed by RSs while paying attention to the multiple nuances of this philosophical notion. Without being exhaustive, we discuss three main reasons why the impact of RSs on DA may be problematic: the threat of manipulation and deception associated with RSs (Sect. 4.1); the RSs' power to reshape users' personal identity (sect. 4.2); the effect of RSs on users' knowledge and critical thinking (Sect. 4.3). Here, we briefly consider these three reasons in turn, discussing how these sources of worry concern in different ways the internalist and the externalist components of DA.

4.1 The Threat of Manipulation and Deception Associated with RSs

One first source of worry is, roughly, that RSs do not just recommend items but actually persuade and manipulate users, or even deceive and coerce them, by circumventing their defenses in various ways, thus steering their ensuing behavior towards a predetermined outcome.

Although lately it has become a commonplace, it is no news that computers may work as persuasive technologies (as when they push users to appreciate or purchase an item) and, by extension, as manipulators (when the user is tricked into forming a belief or making an action), deceivers (when the reality is frankly altered), or even constrainers (when extra costs are openly imposed on alternative actions).

Under the heading *Captology*, a captivating acronym for *Computers As Persuasive Technologies* coined by behavioral scientist B. J. Fogg (1997), the study of computers as persuasive technologies started growing into a field of research of its own as early as in the late nineties. Drawing on the psychological literature, Fogg understood persuasion as the "attempt to shape, reinforce, or change behaviors, feelings, or thoughts about an issue, object, or action". As such, a persuasive computer or technology in general "is an interactive technology that changes a person's attitudes or behaviors" (Fogg 1998, p. 225), and does so *intentionally*, i.e., in view of a goal, although not necessarily *malevolently*, and generally avoiding coercion or deception.

RSs, in particular, may persuade by offering suggestions that are not necessarily based on the users' explicit queries (e.g., in the form of online search), but can be *spontaneously* provided by the system based on the information

the algorithm can extract and elaborate on. To clarify, the mere fact that RSs provide spontaneous suggestions, even persuasive ones, is not enough reason to conclude that DA is threatened or that users are manipulated: if suggestions are sufficiently aligned with the users' interests, they might enhance self-expression (favoring internalist DA) or help otherwise overwhelmed users to focus on a limited set of options and thus exert their reasons-responsiveness better (favoring externalist DA). In many cases, producing a reasons-based justification of the output, after all, advantages the system itself: it increases the likelihood that people trust the recommendation because they perceive the system as transparent (Sikka et al. 2012).

Even the fact that RSs may bring users to act in ways they would not have otherwise chosen, or that RSs reduce the likelihood that they will take some specific courses of action, is not necessarily problematic. Indeed, being autonomous in any philosophically and empirically plausible understanding of DA does not entail that users must be autonomous from *each and every* external influence, even when this implies that they will move on to novel options they would not otherwise think of or will give up on options that they would have otherwise considered. On the one hand, internalist views maintain that agents remain autonomous to the extent that they endorse the relevant motivational drives that are responsible for their conduct, whatever their initial traceable origin has been. Externalist views, on the other hand, accept that environmental and social inputs may work as relevant reasons for acting in one way or another.

More radically, a massive body of literature in theoretical and empirical moral and social psychology, as well as in behavioral economics, has shown that the self is not an island: although single results and methods are widely contested, evidence has been provided that situational factors, even trivial ones, may contribute to guiding people's choices and may even be more influential than apparently stable character traits (Bargh and Chartrand 1999; Doris 2015; Wegner 2002). The most influential of all contexts is actually the social dimension, with its apparatus of written and unwritten norms, expectations, and standards that widely influence human behavior (Bicchieri 2016).

This socially situated literature is often conceptualized as being in tension with DA both in its internalist (self-disclosure) and externalist (reflective thinking and reasons-responsiveness) components (Doris 2015). A plausible implication of the underlying evidence discussed in this literature, however, is not that there are no general action-driving traits or dispositions underlying DA but rather that individual behavior is the byproduct of multiple interacting causal vectors, including the situational (i.e., environmental, social, and political) variables that one progressively encounters and that may rule over traits that one would consider as central to one's self.

Esternalistically-conceived DA, as the propensity for reflective thinking and reasons-responsiveness, might then be rather progressively gained or even boosted by the implementation of RSs deploying algorithmic-based argumentation techniques (Benbasat and Wang 2005; Heras et al. 2017), not least because users may be forced to deal with their cognitive shortcomings. Recommendation, persuasion, and even some types of nudging (Levy 2017), as general communication practices, may thus appeal to the user's reflective thinking, without necessarily exploiting her cognitive weaknesses and idiosyncratic susceptibility to certain recommendation modes. However, it may also happen that pernicious *backfire effects* (Lewandowsky et al. 2012), whereby individuals who are explicitly presented with a correction or a suggestion may paradoxically become more likely to misbehave, partially compromise learning opportunities.

More generally, however, people's ability to take control over weaknesses and biases that encapsulate cognition, or to exert their reasons-responsiveness propensities, is notoriously limited and quite arbitrarily influenced by contingent factors. To name a few of such cognitive shortcomings, the likelihood of choosing an option may be deeply influenced by *framing effects* (i.e., how the option is presented), *availability heuristics* (i.e., the easiness with which some related pieces of information come to mind), *representativeness heuristics* (i.e., the closeness of the option to a presumptive prototype), or the *halo effect* (i.e., the ungrounded way in which a past impression influences a belief about an item) (Bargh and Chartrand 1999; Doris 2015; Kahneman 2013; Thaler and Sunstein 2009). Without discarding the philosophical notion of DA, the existing empirical evidence invites us to discuss the impact of RSs on the grounds of a more nuanced, less-demanding, socially situated, and empirically sound understanding of it.

With this caveat in mind, in our view discussions concerning the specific DA-defeating role of RSs ought to focus on situations where they do not merely help people to satisfy their pre-existing and evolving needs but may rather override them—and may do so utilizing questionable (i.e., manipulative or deceiving) methods. Misalignments of preferences may after all easily arise because most RSs tend to subserve third-party interests—such as increasing sales, generating new needs, stimulating engagement, or boosting political consensus—, which might be in contrast with the users' short or long-term goals. Also, other sources of misalignments may depend on the fact that different users, such as vendors and buyers on e-commerce websites, may compete whenever their utility functions differ.

On a traditional take, computers are not supposed to have intentions or an agenda on their own since they inherit those of their creators—be they manufacturers and programmers or, at a different level, policymakers and company owners

(Fogg 1998, 2002). On a second, more technical, understanding, modern intelligent computers can be thought of as autonomous, goal-directed, agents that possess a model of the environment and whose targets are reached whenever they effectively bring the user to act in accordance with the system's priorities, with the significant drawback of acting on motivations that the user may ultimately fail to recognize as her own. There is indeed discussion as to whether computers or algorithms, and not just human users, can be conceived of as autonomous agents themselves, or also as moral agents who are able to "change state without direct response to interaction" (Floridi and Sanders 2004)—and even as human-like agents with bounded rationality and limited resources who rely on heuristics/model-free computation (Burr et al. 2018). This problem of whether artificial agents are themselves autonomous is, however, beyond the scope of this article.

Even independently of the content that is vehiculated, the practice of providing recommendations may become methodologically worrisome when it trespasses into aggressive or subtle forms of persuasion, manipulation, deception, or even mild coercion (e.g., when users are forced to do something, such as watching a video or visualizing an ad, to access another content) (Botes 2023; Burr et al. 2018; Klenk and Hancock 2019; Susser et al. 2019).

The distinction between DA and NA becomes relevant when one tries to understand the situations in which the attempt to manipulate or deceive users fail so that RSs do not represent an effective threat to the user's DA. In such cases, the RSs' being programmed to operate via manipulative or deceiving strategies may nonetheless count as a violation of NA. In fact, it suffices that RSs are designed in such a way that, for hitting their behavioral target, manipulation or deception may be successful—a paradigmatic example being clickbait and phishing scams (Giachanou et al. 2022), even if experienced users have learned to avoid these traps.

As said, threats to DA occur—and can be experienced at a first-person level when users realize they have spent hours on TikTok, bought an extra set of garden tools, or endorsed an inadequate political candidate—when RSs deploy techniques of covert manipulation or deception by exploiting common, and sometimes even idiosyncratic, cognitive weaknesses and frailties, or when they leverage forms of hidden influence that, by design, prevents users from making informed choices (see Susser et al. 2019). In many cases the implementation of a RS within an online platform is thus part of a company's marketing strategies or subserve private financial or political interests.

However, RSs might be deployed even to promote prosocial, health-related, or sustainable practices. Relatedly, a growing and debated area of research is that of digital nudging, in which behavioral economics concepts and knowledge normally applied to the traditional forms of nudging have been transferred to the online sphere (Mirsch et al. 2017). In fact, some nudges perform particularly well in the online environment and could be implemented within the RS' structure (Jesse and Jannach 2021; Karlsen and Andersen 2019), because the underlying psychological effects are easily exploitable by the corresponding choice architecture. These include *anchoring effects* (providing a reference point to compare other choice options) or *priming effects* (using stimuli that facilitate an upcoming decision) (Mirsch et al. 2017). Nonetheless, even when the ultimate aim is not merely to marketize products and without rejecting nudging practices altogether, caution is required: RSs might be tailored to earn the users' trust in ways that sound per se (i.e., methodologically) questionable (e.g., because they take advantage of cognitive shortcomings), for instance by boosting the level of personalization so that users feel they are understood by the system (Alslaity and Tran 2019; Yoo et al. 2012), or by stimulating emotional engagement (Kramer et al. 2014).

In sum, assuming that the self is naturally penetrable by external influences, the problem that may arise with RSs may depend on the *content* that is vehiculated (as in the case of contents that are enticing but wrong or misleading (more on this in Sect. 4.3) or that override the user's preferences) but, more specifically, on the *methods* by which it is vehiculated. The underlying overall idea, roughly, is that RSs do not just make recommendations but may persuade users to act in ways the motivations of which they would not have endorsed (or they will regret endorsing) had they had sufficient time or cognitive resources to deal with them properly (so impacting negatively on the internalist component of DA), or that bypass their reflective thinking altogether (so impacting negatively on the externalist component of DA). Thus, since the intimate relationship with motivational states and the ability for reflective thinking are respectively seen as building blocks of DA in the internalist and the externalist perspective, the threat to DA becomes very tangible.

4.2 The RSs' Power to Reshape Users' Personal Identity

A second way in which RSs may threaten DA is by causing alteration to one's own identity as grounded in a coherent psychological and narrative bundle.

The internalist component of DA, relying on self-expression and reflective endorsement of stable agential and psychological traits, overlaps with the notion of personal identity or, rather, with the subjective experience of personal identity. To appreciate the relationship between DA and personal identity—and particularly between internalist and diachronic views of DA, grounded in the fulfillment of long-term plans, and personal identity through time—, it is useful to consider that the most popular accounts of identity refer

to criteria of psychological and narrative continuity that are central elements to DA as well (see MacIntyre 1984; Parfit 1984). Alterations of the perception of one's own identity as a person—and of the related bundle of mental states and self-narratives—may thus indirectly affect DA.

Different theories of personal identity provide alternative answers to what it means for an agent to be and to perceive oneself as a coherent unit through time. Psychological views of personal identity—inspired by the Lockean view of identity as rooted in memory and self-reflective consciousness—highlight the centrality of strong memory connections that, jointly with other mental elements (e.g., intentions), grant the agent cross-temporal continuity and coherence. More specifically, first-person narrative views of identity rely on a peculiar type of psychological continuity naturally associated with the notion of a *real self*: personal identity is rooted in the unified and coherent experience of the agent's self-told story, embracing the agential and mental states that belong to the agent's self. Personal identity through time thus depends on the agent's acting on motives that coherently fit one's self-told story (Shoemaker 2021).

On the grounds of a psychological/narrative criterion of identity, as a building block of the internalist component of DA, one ought thus to consider the role of RSs. Assuming that RS-based recommendations are aligned with the agent's true self-defining motives, personalization may in principle reinforce the agent's identity by strengthening the agent's preferences and needs in ways that she would recognize as respectful of her self-told narrative. There are, however, some reasons why this positive result might be often—although not necessarily—hampered even without considering cases of outright misalignment of preferences.

First, although fine-grained, personalized recommendations may improve over time due to multiple interactions, RSs proceed by assigning users to broad social categories associated with bundles of tastes or preferences. One problem is that the tags with which users are implicitly or explicitly identified may poorly reflect categories that are central to the agent's perceived, subjective self so that the corresponding suggestions may become annoying or irrelevant. At the same time, by telling people who they are, these tags contribute to mediating and reshaping the subjective experience of one's own identity as defined by the background of social categories to which one is iteratively compared (Milano et al. 2020). RSs in fact tend to boost the relevance of the selected preferences and categories: through iterative interactions and feedback loops, these potentially stereotyped associations may gain more and more centrality within the individual's mental life and, consequently, reshape her personal identity (Helberger 2019). For instance, a user may end up experiencing herself as an avid consumer of certain types of news, a fan of a music genre, or a member of a given social group, while these features or categories were not particularly central to her pre-existing, subjectively perceived identity.

These phenomena can be exacerbated by psychological biases such as the *salience effect*, which defines the tendency to pay more attention to certain news, items, or events that for some reason appear as noteworthy or emotionally charged, quite independently of their objective relevance or likelihood to happen. Furthermore, by iteratively feeding users with rewarding contents, RSs may selectively reinforce certain behavioral tendencies, especially reward-seeking behaviors (Burr et al. 2018), at the expense of others. Besides being worrisomely involved in addictive inclinations (Cooper et al. 2017), the over-stimulation of the reward system may significantly alter the user's feeling of her own self. Directly feeding the underlying self-told narratives, category-based recommendations may thus modify users' experience of their selves in ways that, in the long run, they might recognize as unsatisfactory. By doing so, RS-based recommendations indirectly affect the internalist component of DA as grounded in the reflectively endorsed and coherent experience of one's own mental life and narrative.

To some extent, the reshaping of personal identity due to the interaction with exogenous variables is a recurring experience. Cognitive sciences have shown that our psychological states and self-told narratives—our perception of ourselves as a certain type of individual—are continuously modulated by the external inputs we are exposed to when navigating the social environment, online and offline (Bargh and Chartrand 1999). Furthermore, our insight about preferences and motivations are often the byproduct of rationalizations aiming to justify or interpret our pre-reflective intuitions, often when we find ourselves in social contexts and must make sense of a much more fragmented mental life (Haidt 2001) — a process that may fall back into misattributions of agency and confabulation (Johansson et al. 2005; Wegner 2002). Given this general lack of introspective skills and the tendency to self-deception, agents themselves may often have an unrealistic or inflated view of their preferences and personality (Bargh and Chartrand 1999; Doris 2015). In such cases, RSs may even provide a more clear-sighted representation of the agent's psychological profile, thus supporting DA by providing reliable and transparent feedbacks on the agent's *objective* tastes and needs and helping people realign their choice behavior.[1]

So, the experience of one's identity is by default open-ended and in progress, as well as potentially confabulatory. Nonetheless, DA-related concerns about the role played by RSs reach the warning level especially when users' decision-making is systematically steered, taking advantage

[1] We are grateful to an anonymous reviewer for a useful comment along these lines.

of common cognitive shortcomings, towards options that mirror third-party interests more than tracking preferences and categories that users would see as central to their own identity and personality. And because our individual and social identity is built over time and through multiple learning experiences, problems may arise when users' time and cognitive resources, especially at a young age, are depleted by the provision of contents that are tailored-made to overwhelmingly capture their attention. Prolonged exposure to absorbing contents taking up much of people's mental energy might deprive them of the time and cognitive resources needed for a healthy mental and social life. Indeed, the experience of being carried away with online contents characterizes the addictive use and abuse of social media. This process is known to bring about cognitive and developmental problems (Giraldo-Luque et al. 2020), and to elicit feelings of regret and anger (Altuwairiqi et al. 2019; Burr et al. 2018).

Potential corrective measures include facilitating users' more active involvement in the design of meaningful choice opportunities, for instance by asking them to express their views about the categories they feel closer to their individual and social identity. Although, as mentioned, people's introspective power is limited (Bargh and Chartrand 1999; Doris 2015; Johansson et al. 2005), at least in some cases this can be done by explicitly asking users to judge the outputs of the RS or respond to questionnaires about their preferences and needs (Tiribelli 2023). However, in many other cases, both average and experienced users may find it difficult to evaluate the system's performance or specify their goals by responding to direct queries. In such cases, especially when unwanted misalignments of interests are the case, the problem must be tackled upstream by more sophisticated training strategies that improve the algorithm's adaptive performance and learning processes (Christiano et al. 2018).

To avoid depleting people's cognitive resources and favor their personal and social development, remedial strategies must include training algorithms to respond to the users' real needs and not just to their biases. However, this is not an easy task: observation and tracking of users' behavior do not directly lead to solving the theoretical and computational problem of distinguishing users' long-term goals from short-term gratifications (Burr et al. 2018; Hadfield-Menell et al. 2017). More generally, users' DA could be safeguarded by making the underlying algorithmic decision-making (e.g., the mechanisms by which the system assigns users certain tags) more transparent through forms of explainable A.I. It is worth noticing, however, that the attempt to produce user-friendly explanations may generate sound but ultimately fake narratives, which superficially meet the human desire for a comprehensible and meaningful story but are not truly representative of the ongoing processes (Perez 2018; Re and Solow-Niederman 2019).

In sum, the internalist criterion of DA overlaps importantly with the notion of personal identity as grounded in forms of psychological and narrative continuity. In the long run, by assigning users to social categories, RSs participate in the construction of users' identity in ways that may or may not be responsive to their view of themselves or that may be detrimental to their healthy and satisfactory development. As such, RS-based recommendations may affect the internalist component of DA in ways that deserve serious consideration when designing and training the underlying algorithms.

4.3 The Effect of RSs on Users' Knowledge and Critical Thinking

A third way in which RSs may threaten DA is by harming knowledge acquisition and the development of critical thinking.

As seen, externalist views of DA highlight that the ability to rationally govern oneself and navigate the social environment are central features of DA. The opportunity of developing and expressing these abilities—i.e., by taking informed decisions, learning about different ideas and opinions, and thus developing, to the extent possible, a clear and unbiased grasp of reality—is largely mediated by processes of knowledge acquisition. How the pieces of news we consume or the information we acquire are organized by the various media outlets we turn to plays a fundamental role in these knowledge-acquisition processes. Since RSs filter how users consume news and acquire relevant information about the outside world, as well as their preferences and social affiliations, they may contribute powerfully to the dynamics of knowledge acquisition and the development of critical thinking (Bakshy et al. 2015). This way, they may affect the externalist component of DA, not forgetting that the capacity of making meaningful and informed decisions is also a grounding factor of NA (Keeling 2018).

Although knowledge and information acquisition is nourished by multiple different sources, it would be wrong to underestimate the weight of RSs. From the perspective of individual users, people have increasingly consumed large quantities of news on social media, where RSs rule the roost, while migrating away from more traditional media (Bakshy et al. 2015; Shearer and Gottfried 2017; Raza and Ding 2022). And, on the news producer side, various reports indicate that algorithmic news recommendation is a central focus of current and future digital experimentation processes, with the aim to develop more and more personalized insight into consumers' preferences (Helberger 2019).

Moreover, if we consider not only online outlets but also social media in their function of broadcasting news (as it is necessary given their increasing role in shaping our informational ecosystem (Napoli 2019)), there seem to be at least two ways—one by bringing people to diverge from accepted

standards, the other by promoting excessively standardized behaviors—in which RSs may worrisomely affect knowledge and critical thinking, thus potentially menacing DA in its externalist fashion.

First, online outlets, and social media in particular, have been associated with the spread of misinformation, fake news, and half-baked theories because they may easily avoid the checks to which more traditional media are usually subject, and provide streams of content that are misleading or false altogether (Tommasel and Menczer 2022). The RSs implemented into online outlets have the potential to escalate this effect by promoting behaviors that are radicalized or worrisomely divergent from validated standards and at the same time particularly enticing and appealing to the target audience. This way, they may boost worrisome and already widespread social tendencies, such as the demise of trust in experts, often accompanied by processes of disintermediation, or support questionable political agendas (Beckett and Douze 2016; Whittaker et al. 2021).

Multiple studies have indeed pointed to the role social media have played in the spread of anti-scientific theories during the Covid pandemic (Bin Naeem et al. 2021), in fueling Euroscepticism during the UK referendum campaign leading to Brexit (Hänska and Bauchowitz 2017), or in being a crucial factor of the recent shocking growth of the Flat Earth movement (Landrum et al. 2021). Supporters of such anti-scientific or conspiracy theories often gesture directly at the value of DA—suggesting that people "ought to do their own research!" (Levy 2022)—in opposition to mainstream views shaped by expert opinions. The conclusion is often that experts ought to be silenced: individuals can then *autonomously* form their own opinion by getting to know alternative facts that are culpably hidden from public scrutiny.

The myth of the lonesome individual against the mainstream is grounded in a misconceived understanding of DA. Besides other concerns, it clashes with our physiological limitations in terms of cognitive and computational capacities, as well as attentional resources (Gigerenzer and Selten 2001). In reality, reasons-responsiveness and reflective thinking are indeed central aspects of DA, but they are expressed also, and actually quite often, by the ability to exert forms of epistemic deference when circumstances so require, and expert sources are available (Levy 2019). Reversely, in this respect, RSs may actually support DA by pre-selecting trustworthy content and excluding unreliable informational sources for an audience that is unavoidably short of time and attentional assets.

Second, however, iterative recommendations concerning analogous items may have the effect of closing off people's opportunity to appreciate diverse stimuli, thus promoting behaviors that progressively become standardized and homogeneous within sealed clusters of users. Exposure to similar contents across multiple interactions—independently of their intrinsic value or disvalue but due to excess of homogeneity—, tends to diminish users' ability to engage in critical thinking, question their certainties, or be responsive to multiple actionable reasons, which are all key aspects of externalist DA. In the long run, a passive lack of familiarity with challenging points of view in the media landscape may cause users to actively avoid confrontation more generally (Helberger 2019; Raza and Ding 2022), with important implications on the ability to navigate the social environment online and offline. On the one hand, it is imaginable that RSs can cultivate people's mindset by pointing to interesting or informative content that they would not have otherwise considered—once again, people's intellectual independence ought not to be overstated. On the other hand, research has shown that, in practice, people on social media tend to remain in their comfort zone. For instance, a 2015 large-scale analysis of widespread habits of Facebook users indicated that ideologically homophilic contents largely outnumber cross-cutting contents in terms of what people are exposed to and what they choose to share (Bakshy et al. 2015; Burr et al. 2018).

Beyond their effect on the intellectual growth of single users, standardization and homogenization are also detrimental to one of the media's key functions, i.e., to contribute to the creation and maintenance of a diverse and independent *public forum* where multiple views can compete on a par. This democracy-enhancing function easily clashes with the attempt to develop approaches to meet people's short-term preferences by producing quickly rewarding user-tailored contents (Helberger 2019).

Research has shown that standardization and homogenization are magnified by both psychological and algorithmic biases. On the psychological side, due to common communication dynamics and conversational heuristics, users have been shown to trust the news recommended by the algorithms as constructing a faithful model of the outside reality, or value the opinion expressed in an argumentative piece as representative of the consensus on a topic (Burr et al. 2018). Well-studied technology-mediated psychological phenomena, such as *automation bias* and *automation complacency*, may further contribute to explaining the trust dynamics in place in human-computer interactions. The term "automation bias" refers to the tendency to see automated solutions as more reliable than human-based outputs, especially in cases of mismatch. "Automation complacency" refers instead to the inclination to accept automated solutions uncritically and passively (Parasuraman and Manzey 2010).

On the algorithmic side, the term "popularity bias" refers to the widespread tendency for which RSs tend to increasingly promote items that are already popular, while less popular ones, despite being potentially interesting (especially for

minorities), remain uncovered (Dinnissen & Bauer 2022). Paradoxically, while increasing standardization and homogeneity within a cluster, popularity bias decreases personalization, with less-than-obvious matches being systematically neglected. In filtering the contents to display, priority is thus given to those that will likely increase the probability that users will select the proposed item, often discarding relevant values, such as the attempt to be comprehensive, unbiased, or exhaustive. Mitigation strategies, including re-ranking to avoid over-representation of certain items, are now routinely implemented to increase accuracy and diversify user experiences (Klimashevskaia et al. 2022), but have rarely the potential to produce game-changing results.

In sum, RS-based promotion of nonstandard contents on the one hand, and excess of standardization on the other hand have been routinely and rightly associated with the risk of being trapped in filter bubbles and echo chambers; moreover, they have a considerable impact on the social media's role in democratic societies. In addition, RS-mediated recommendations, also in virtue of specific psychological and algorithmic biases, affect DA as well, especially for what concerns the ability to develop reflective thinking and reasons-responsiveness, which are central elements of its externalist face.

5 Conclusions

Discussions of the ethical implications of the RSs' widespread implementation have recently gained momentum across different research fields. In this paper, we aimed at drawing attention to DA as a multifaceted concept whose different expressions can be distinctively hampered by the ubiquitous deployment of RSs. To organize the discussion, we have distinguished between internalist DA (in the form of reflective endorsement) and externalist DA (in the form of responsiveness to reasons), aiming also to mitigate the over-idealization that may permeate both ordinary speech and philosophical theorizing about DA, and promote an understanding of people's decision-making modes and capacities more aligned with the results of cognitive sciences. On these grounds, we have illustrated three reasons why RSs may undermine DA: the threat of manipulation and deception associated with RSs, which may affect both internalist and externalist DA; the RSs' power to reshape users' personal identity, which may affect internalist DA; the impact of RSs on users' knowledge and critical thinking, which may affect externalist DA. Notwithstanding these legitimate concerns, RSs are entrenched in our daily activities in ways that would be difficult, inconvenient, if not impossible, to dismantle; moreover, RSs may effectively help users to navigate otherwise troubled waters or even expand the agent's intellectual interests and promote pro-social behaviors. Our perspective, therefore, is not to be intended as a bulwark to protect the *status quo* but as an invitation to carefully weigh these aspects in the design of ethically oriented RSs.

Funding Open access funding provided by Università degli Studi Roma Tre within the CRUI-CARE Agreement. The three authors benefitted from the PRIN Grant 20175YZ855 from the Italian Ministry for Education, University and Research (Ministero dell'Istruzione, dell'Università e della Ricerca). Benedetta Giovanola also benefitted from the Jean Monnet Chair (Grant Agreement 101085372) EDIT – Ethics for Inclusive Digital Europe, co-funded by the European Union. Views and opinions expressed are however those of the author only and do not necessarily reflect those of the European Union or the European Education and Culture Executive Agency (EACEA). Neither the European Union nor EACEA can be held responsible for them.

Declarations

Conflict of interest The authors have no conflicts of interest to declare.

References

Aggarwal CC (2016) Recommender systems. Springer, Cham

Alslaity A, Tran T (2019) Towards persuasive recommender systems. 2019 IEEE 2nd international conference on information and computer technologies (ICICT), 143–148

Altuwairiqi M, Jiang N, Ali R (2019) Problematic attachment to social media: five behavioural archetypes. Int J Environ Res Public Health 16(12):2136

Arpaly N, Schroeder T (1999) Praise, blame, and the whole self. Philos Stud 93(2):161–188

Bakshy E, Messing S, Adamic LA (2015) Exposure to ideologically diverse news and opinion on Facebook. Science 348(6239):1130–1132

Bargh JA, Chartrand TL (1999) The unbearable automaticity of being. Am Psychol 54(7):462–479

Beauchamp TL, Childress JF (2001) Principles of biomedical ethics, 5th edn. Oxford University Press, New York

Beckett C, Douze M (2016) On the role of emotion in the future of journalism. Social Media + Society 2(3)

Benbasat I, Wang W (2005) Trust in and adoption of online recommendation agents. J Association for Inform Syst 6(3)

Bicchieri C (2016) Norms in the wild: how to diagnose, measure, and change social norms. Oxford University Press, New York

Bollen D, Knijnenburg BP, Willemsen MC, Graus M (2010) Understanding choice overload in recommender systems. RecSys '10: Proceedings of the fourth ACM conference on Recommender systems, 63–70

Bonicalzi S (2019) Rethinking moral responsibility, Mimesis, Milan-London
Bonicalzi S, Haggard P (2019) From freedom from to freedom to: new perspectives on intentional action. Front Psychol 10:1193
Botes M (2023) Autonomy and the social dilemma of online manipulative behavior. AI Ethics 3:315–323
Bratman M (2018) Planning, time, and self-governance: essays in practical rationality. Oxford University Press, Oxford
Burr C, Cristianini N, Ladyman J (2018) An analysis of the interaction between intelligent software agents and human users. Minds & Machines 28:735–774
Buss S, Westlund A (2018) Personal autonomy. Edward N. Zalta (ed.). The Stanford Encyclopedia of Philosophy (Spring 2018 Edition), URL = <https://plato.stanford.edu/archives/spr2018/entries/personal-autonomy/>. Accessed 27 January 2023
Carraro D, Bridge DA (2022) A sampling approach to debiasing the offline evaluation of recommender systems. J Intell Inf Syst 58:311–336
Christiano P, Shlegeris B, Amodei D (2018) Supervising strong learners by amplifying weak experts. arXiv:1810.08575
Christman J (1991) Autonomy and personal history. Can J Philos 21:1–24
Cinelli M, De Francisci Morales G, Galeazzi A, Starnini M (2021) The echo chamber effect on social media. PNAS 118(9):e2023301118
Cohen MK, Hutter M, Osborne MA (2022) Advanced artificial agents intervene in the provision of reward. AI Magazine 43(3):282–293
Cooper S, Robison AJ, Mazei-Robison MS (2017) Reward circuitry in addiction. Neurotherapeutics 14(3):687–697
De Croon R, Van Houdt L, Htun NN, Štiglic G, Vanden Abeele V, Verbert K (2021) Health recommender systems: systematic review. J Med Internet Res 23(6):e18035
Diakopoulos N, Koliska M (2017) Algorithmic transparency in the news media. Digit Journalism 5(7):809–828
Dinnissen K, Bauer C (2022) Fairness in music recommender systems: a stakeholder-centered mini review. Front Big Data 5:913608
Doris J (2015) Talking to our selves. Reflection, ignorance, and agency. Oxford University Press, Oxford
Fischer JM, Ravizza M (1998) Responsibility and control: a theory of moral responsibility. Cambridge University Press, New York
Floridi L, Sanders JW (2004) On the morality of artificial agents. Mind Mach 14(3):349–379
Fogg BJ (1997) Captology: the study of computers as persuasive technologies, CHI EA '97: CHI. '97 Extended Abstracts on Human Factors in Computing Systems
Fogg BJ (1998) Persuasive computers: perspectives and research directions. Proceedings of the SIGCHI conference on Human factors in computing systems—CHI '98—Persuasive computers. 225–232
Fogg BJ (2002) Persuasive technology. Using computers to change what we think and do. Morgan Kaufmann, Burlington (MA)
Frankfurt HG (1971) Freedom of the will and the concept of a person. In: Frankfurt HG (ed) (1988) The importance of what we care about: philosophical essays. Cambridge University Press, New York, pp 11–25
Giachanou A, Zhang X, Barrón-Cedeño A, Koltsova O, Rosso P (2022) Online information disorder: fake news, bots and trolls. Int J Data Sci Anal 13(4):265–269
Gigerenzer G, Selten R (2001) Rethinking rationality. In: Gigerenzer G, Selten R (eds) Bounded rationality: the adaptive toolbox. The MIT Press, Cambridge (MA), pp 1–12
Gillon R (1994) Medical ethics: four principles plus attention to scope. Brit Med J 309(5):184
Giraldo-Luque S, Aldana Afanador PN, Fernández-Rovira C (2020) The struggle for human attention: between the abuse of social media and digital wellbeing. Healthcare 8(4):497
Hadfield-Menell D, Milli S, Abbeel P, Russell SJ, Dragan A (2017) IInverse Reward Design. 31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA
Haidt J (2001) The emotional dog and its rational tail: a social intuitionist approach to moral judgment. Psychol Rev 108(4):814–834
Hänska M, Bauchowitz S (2017) Tweeting for Brexit: how social media influenced the referendum. In: Mair J, Clark T, Fowler N, Snoddy R, Tait R (eds) Brexit, Trump and the media. Abramis academic publishing, Bury St Edmunds, UK
Helberger N (2019) On the democratic role of news recommenders. Digit Journalism 7(8):993–1012
Herlocker JL, Konstan JA, Riedl J (2000) Explaining collaborative filtering recommendations. CSCW '00: Proceedings of the 2000 ACM conference on computer supported cooperative work. 241–250
Heras S, Rodríguez P, Palanca J, Duque N, Julián V (2017) Using argumentation to persuade students in an educational recommender system. In: de Vries P, Oinas-Kukkonen H, Siemons L, Beerlage-de Jong N, van Gemert-Pijnen L (eds) Persuasive technology: development and implementation of personalized technologies to change attitudes and behaviors. PERSUASIVE 2017. Lecture Notes in Computer Science, vol 10171. Springer, Cham
Himeur Y, Sohail SS, Bensaali F, Amira A, Alazab M (2022) Latest trends of security and privacy in recommender systems: a comprehensive review and future perspectives. Computers & Security 118:102746
Ilievski I, Roy S (2013) Personalized news recommendation based on implicit feedback. Proceedings of the 2013 International ACM RecSys news recommender systems workshop and challenge, 10–15
Jannach D, Zanker M, Felfernig A, Friedrich G (2010) Recommender Systems: an introduction. Cambridge University Press, Cambridge
Jesse M, Jannach D (2021) Digital nudging with recommender systems: survey and future directions. Comput Human Behav Rep 3
Johansson P, Hall L, Sikström S, Olsson A (2005) Failure to detect mismatches between intention and outcome in a simple decision task. Science 310(5745):116–119
Kahneman D (2013) Thinking fast and slow. Penguin Books Ltd, London
Karlsen R, Andersen A (2019) Recommendations with a nudge. Technologies 7(2):45
Keeling G (2018) Autonomy, nudging and post-truth politics. J Med Ethics 44(10):721–722
Klenk M, Hancock J (2019) Autonomy and online manipulation. Internet Policy Review https://policyreview.info/articles/news/autonomy-and-online-manipulation/1431
Klimashevskaia A, Elahi M, Jannach D, Trattner C, Skjærven L (2022) Mitigating popularity bias in recommendation: potential and limits of calibration approaches. In: Boratto L, Faralli S, Marras M, Stilo G (eds) Advances in bias and fairness in information retrieval. BIAS 2022. Communications in Computer and Information Science, vol 1610. Springer, Cham
Kramer ADI, Guillory JE, Hancock JT (2014) Experimental evidence of massive-scale emotional contagion through social networks. PNAS Proc Natl Acad Sci United States Am 111(24):8788–8790
Landrum AR, Olshansky A, Richards O (2021) Differential susceptibility to misleading flat Earth arguments on YouTube. Media Psychol 24(1):136–165
Levy N (2017) Nudges in a post-truth world. J Med Ethics 43(8):495–500
Levy N (2019) Due deference to denialism: explaining ordinary people's rejection of established scientific findings. Synthese 196(1):313–327
Levy N (2022) Do your own research. Synthese 200(5):356

Lewandowsky S, Ecker UKH, Seifert CM, Schwarz N, Cook J (2012) Misinformation and its correction: continued influence and successful debiasing. Psychol Sci Public Interest 13:106–131
Lika B, Kolomvatsos K, Hadjiefthymiades S (2014) Facing the cold start problem in recommender systems. Expert Syst Appl 41(4):2065–2073
MacIntyre A (1984) After virtue. University of Notre Dame Press, Notre Dame
Mackenzie C, Stoljar N (2000) Relational autonomy. Oxford University Press, Oxford
Matz SC, Kosinskim N, Nave G, Stillwell DJ (2017) Psychological targeting as an effective approach to digital mass persuasion. PNAS 114(48):12714–12719
Mele AR (2001) Autonomous agents: from self-control to autonomy. Oxford University Press, Oxford
Mele AR (2013) Moral responsibility, manipulation, and minutelings. J Ethics 17(3):153–166
Milano S, Taddeo M, Floridi L (2020) Recommender systems and their ethical challenges. AI & Soc 35:957–967
Mirsch T, Lehrer C, Jung R (2017) Digital nudging: Altering user behavior in digital environments. In: Leimeister, J M, Brenner, W (eds.) Proceedings of the 13th international conference on Wirtschaftsinformatik, pp 634–648
Naeem SB, Bhatti R, Khan A (2021) An exploration of how fake news is taking over social media and putting public health at risk. Health Info Libr J 38(2):143–149
Napoli P (2019) Social media and the public interest. Columbia University Press, New York
Nelkin D (2011) Making sense of freedom and responsibility. Oxford University Press, Oxford
Oshana M (2006) Personal autonomy in society. Ashgate Publishing, Aldershot
Papakyriakopoulos O, Medina Serrano JC, Hegelich S (2020) Political communication on social media: a tale of hyperactive users and bias in recommender systems. Online Social Networks and Media 15:100058
Parasuraman R, Manzey DH (2010) Complacency and bias in human use of automation: an attentional integration. Hum Factors 52(3):381–410
Parfit D (1984) Reasons and persons. Oxford University Press, Oxford
Pariser E (2011) The filter bubble: what the internet is hiding from you. Penguin, New York
Pawlicka A, Pawlicki M, Kozik R, Choraś RS (2021) A systematic review of recommender systems and their applications in cybersecurity. Sens (Basel) 21(15):5248
Perez CE (2018) Deep learning's uncertainty principle. Intuition Machine. https://medium.com/intuitionmachine/deep-learnings-uncertainty-principle-13f3ffdd15ce#:~:text=The%20uncertainty%20principle%20as%20applied,interpretable%20don%27t%20generalize%20well. Accessed 27 January 2023
Pereboom D (2014) Free will, agency and meaning in life. Oxford University Press, Oxford
Pérez-Almaguer Y, Yera R, Alzahrani AA, Martínez L (2021) Content-based group recommender systems: a general taxonomy and further improvements. Expert Syst Appl 184:115444
Qin Y, Omar B, Musetti A (2022) The addiction behavior of short-form video app TikTok: the information quality and system quality perspective. Front Psychol 13:932805
Raza S, Ding C (2022) News recommender system: a review of recent progress, challenges, and opportunities. Artif Intell Rev 55:749–800
Re RM, Solow-Niederman A (2019) Developing artificially intelligent justice. Stanford Technol Law Rev 22:242
Schwartz B (2016) The paradox of choice: why more is less. Harper-Collins Publishers Inc, New York
Shearer E, Gottfried J (2017) News use across social media platforms. PEW Research Center [Online], 7 September 2017. http://www.journalism.org/2017/09/07/news-use-across-social-media-platforms-2017/. Accessed 25 Jan 2023
Shoemaker D (2021) Personal identity and ethics. Zalta, E N (Ed.) The Stanford Encyclopedia of Philosophy (Fall 2021 Edition), URL = https://plato.stanford.edu/archives/fall2021/entries/identity-ethics/. Accessed 27 Jan 2023
Sikka R, Dhankhar A, Rana C (2012) A survey paper on e-learning recommender systems. Intl J Comput Appl 47(9):27–30
Simon H (1957) Models of man. John Wiley, New York
Spahn A (2012) And lead us (not) into persuasion? Persuasive technology and the ethics of communication. Sci Eng Ethics 18:633–650
Sripada C (2017) Frankfurt's unwilling and willing addicts. Mind 126(503):781–815
Stoljar N (2017) Relational autonomy and perfectionism. Moral Philos Politics 4(1):27–41
Su Z, Huang Z, Ai J, Zhang X, Shang L, Zhao F (2022) Enhancing the scalability of distance-based link prediction algorithms in recommender systems through similarity selection. PLoS ONE 17(7):e0271891
Susser D, Roessler B, Nissenbaum H (2019) Technology, autonomy, and manipulation.Internet Policy Review 8(2)
Talbert M (2017) Akrasia, awareness, and blameworthiness. In: Robichaud P, Wieland JW (eds) Responsibility: the epistemic condition. Oxford University Press, Oxford, pp 47–63
Thaler RH, Sunstein C (2009) Nudge: improving decisions about health, wealth, and happiness. Penguin Books, London
Tiribelli S (2023) The AI ethics principle of autonomy in health recommender systems. Argumenta 16:1–18
Tommasel A, Menczer F (2022) Do recommender systems make social media more susceptible to misinformation spreaders? RecSys '22: Proceedings of the 16th ACM Conference on Recommender Systems, pp 550–555
Varkey B (2021) Principles of clinical ethics and their application to practice. Med Princ Pract 30:17–28
Watson G (1975) Free agency. In: Watson G (ed) Agency and answerability: selected essays. Oxford University Press, New York, pp 13–32
Watson G (1987) Responsibility and the limits of evil: variations on a strawsonian theme. In: Schoeman F (ed) Responsibility, character, and the emotions. Cambridge University Press, New York, pp 256–286
Wegner DM (2002) The illusion of conscious will. The MIT Press, Cambridge (MA)
Whittaker J, Looney S, Reed A, Votta F (2021) Recommender systems and the amplification of extremist content. Internet Policy Rev 10(2):1–29
Wolf S (1990) Freedom within reason. Oxford University Press, New York
Yesilada M, Lewandowsky S (2022) Systematic review: YouTube recommendations and problematic content. Internet Policy Rev 11(1):1–22
Yoo K, Gretzel U, Zanker M (2012) Persuasive recommender systems. Conceptual background and implications. Springer, Cham