AI & SOCIETY
https://doi.org/10.1007/s00146-023-01720-2

OPEN FORUM

AI-powered recommender systems and the preservation of personal autonomy

Juan Ignacio del Valle[1] · Francisco Lara[1]

Received: 27 February 2023 / Accepted: 27 June 2023

Abstract
Recommender Systems (RecSys) have been around since the early days of the Internet, helping users navigate the vast ocean of information and the increasingly available options that have been available for us ever since. The range of tasks for which one could use a RecSys is expanding as the technical capabilities grow, with the disruption of Machine Learning representing a tipping point in this domain, as in many others. However, the increase of the technical capabilities of AI-powered RecSys did not come with a thorough consideration of their ethical implications and, despite being a well-established technical domain, the potential impacts of RecSys on their users are still under-assessed. This paper aims at filling this gap in regards to one of the main impacts of RecSys: personal autonomy. We first describe how technology can affect human values and a suitable methodology to identify these effects and mitigate potential harms: Value Sensitive Design (VSD). We use VSD to carry out a conceptual investigation of personal autonomy in the context of a generic RecSys and draw on a nuanced account of procedural autonomy to focus on two components: competence and authenticity. We provide the results of our inquiry as a value hierarchy and apply it to the design of a speculative RecSys as an example.

Keywords Artificial intelligence · Recommender systems · Autonomy · Identity · Regulation · Value sensitive design

1 Introduction

“They steal your data. They hack your brain. They rule the world” (Wylie 2019). “If they get to know you better than you know yourself, they can then sell you anything they want—be it a product or a politician” (Harari 2018). Those are just a couple of examples of the headlines, books, interviews… that have been published, particularly in the last decade, warning us about the danger that Artificial Intelligence (AI) represents to our lives. Personal autonomy is at the centre of these threats, and it is identified implicitly or explicitly almost by default as one of the core values to be preserved in the many ethical guidelines that have been issued over the last few years.[1]

A paradigmatic case of a potential impact on human autonomy is the use of Recommender Systems (RSs, RecSys) that have been around over the last two decades to assist us in the way we work, move, shop, exercise, enjoy digital content, get education, find friends, find love… From philosophy of technology, we know that technology is not neutral: it entails a mediating effect, enabling certain actions and disclosing certain parts of the world in specific directions, leaving others in the shadow (Verbeek 2011), and Recommender Systems are a notable example of this.

In addition, AI-powered applications, particularly those based on Machine Learning (AI/ML), add an element of complexity inherent to the very process of technology development. Indeed, data driven approaches encompass specific problems related to bias, explainability or unpredictability, to name a few, that have been identified and assessed carefully in recent years (see e.g., (Coeckelbergh 2020b, 2022a)).

✉ Francisco Lara
flara@ugr.es

Juan Ignacio del Valle
jidelvalle@correo.ugr.es

[1] Departamento de Filosofía I, Universidad de Granada, Granada, Spain

[1] For an overview of the main AI ethics guidelines see (Jobin et al. 2019), which identifies autonomy as one of the top ethical principles elicited thereof.

Published online: 21 July 2023

Despite these issues being well identified, the impact of Recommender Systems on personal autonomy is seldom analysed in depth and the very concept of autonomy is never described clearly, as it is normally assumed that we all share what it means. But what is this autonomy that we should preserve? Are we not already subject to a plethora of influences (e.g., our family, our friends, our *zeitgeist*) that make the common understanding of autonomy and other related concepts such as identity, the self, or authenticity misleading? And what exactly is the threat? What is so particular about AI that entails a risk and how can we evaluate the impact of this specific technology on user autonomy?

This paper aims at assessing the impact of the use of Recommender Systems, particularly when powered by AI/ML, on user autonomy. It aims to bring these rather abstract philosophical reflections closer to a useful conceptual framework that could be used to support the development and evaluation of this technology.

Section 2 will provide a brief introduction to Recommender Systems in general and highlight some of their main design characteristics. It will also provide an example of a specific, yet speculative, RecSys in a particularly interesting context —parenting—which we believe will be useful for framing the remainder of this paper and its results. Section 3 will introduce some issues related to the evaluation of technology and some key aspects of the methodology that will drive our assessment: Value Sensitive Design (VSD). The VSD methodology attempts to account for human values in a systematic manner throughout the system design process and provides an adequate interface for making philosophical concepts and discourse available to the engineering domain.

The main part of the paper is Sect. 4, which introduces the conceptual analysis of personal autonomy, assesses the impact of Recommender Systems thereon and proposes some mitigations. Section 5 will summarise the results of this assessment, and Sect. 6 will provide some concluding remarks.

2 Recommender systems: a brief overview and one example

Recommender Systems (RecSys) have been around since the early days of the Internet helping users to navigate within the vast ocean of information and the increasingly available options that have been made available for us ever since. We are all aware —and most of us are users— of RecSys, such as Google, Amazon, YouTube, Netflix, TikTok, Twitter, Facebook, LinkedIn, or Tinder. They support us to find the information we need, right products to buy, good films to watch, interesting content (videos, news…) to see, like-minded persons to connect… The range of tasks for which one could use a RecSys is expanding as the technical capabilities grow, with the disruption of Machine Learning representing a tipping point in this domain, as in many others. RecSys are now evolving into multi-purpose, conversational, ubiquitous, intelligent assistants, and examples such as Amazon's Astro will even aim to predict our needs.[2]

From a technical point of view, RecSys are "software tools and techniques that provide suggestions for available items that are most likely of interest to a particular user" (Ricci et al. 2015). Amongst the different characteristics commonly used to define these systems (Ricci et al. 2015), there are four elements we believe to be particularly relevant for describing a generic system:

- Function: The function or intended functionality of a system is a key element in any design description thereof. Normally, in AI-powered systems, particularly in those based on Machine Learning, this is translated into a Utility Function, a quantifiable element closely related to this intended functionality, used to drive the training process.
- Items: Items are the entities being recommended. In current Recommender Systems they are news, movies, people, travels, jobs…
- User model: A key feature in modern Recommender Systems is the user model, which is used to adapt (personalise) recommendations and the interactions of systems in general. Obviously, when evaluating the impact of an AI-powered Recommender System in human autonomy, a detailed description of how it builds up and employs the user model is of utmost importance.
- Recommendation technique: This is the way the RecSys is going to fulfil its function using the item features and user model. The technical details of this extensive area in continuous development (Ricci et al. 2015) are well beyond the scope of this paper; nevertheless, having a high-level understanding of the system's recommendation technique will be important to understand its impact on user autonomy. For example, it may rest on knowledge-based techniques, in which the recommendation depends on specific rules, or content-based techniques, in which the system aims at recommending items according to their features and how they match user preferences. However, the most interesting recommendation technique and —we believe—the one with the most potential to disrupt human autonomy is collaborative filtering, in which the recommendations are based on matching different user profiles in a rather opaque fashion.

These characteristics can define a wide range of systems, and our approach to RecSys encompasses not only typical applications such as the Amazon or YouTube recommender systems, but also other systems normally referred to as Virtual Assistants or Expert Systems. The ethical impact —including but

[2] https://www.wired.com/story/amazon-wants-its-home-robot-astro-to-anticipate-your-every-need/

not limited to user autonomy— of these systems will depend on their use case. Intuitively, one can foresee that this impact will be different when using the Netflix RecSys to when using e.g., the Artificial Moral Advisor, a speculative moral assistant proposed by Alberto Giubilini and Julian Savulescu to improve human moral decision-making (Giubilini and Savulescu 2018). Below, we provide our own speculative RecSys, which will hopefully lay bare some of the ethical issues we want to assess.

GePeTo: the parental RecSys.

John is a 40-year-old man and father of a 5-year-old girl. A close relative has recently passed away and John must tell his daughter. This is the first time he has dealt with the subject of death with his daughter, and he is unsure about what line he should take. John is using GePeTo a (speculative) app that can assist parents in daily discussions with their children in a wide variety of subjects including education, sex, love, friendship, fashion, emotions…

GePeTo has been trained using the entire Internet to fulfil its function and, in addition, it is based on generative AI (GPT3), which can create new content if so required. Finally, GePeTo requires John to identify himself, ideally with his Google or Facebook account, which provides the app with direct access to John’s digital profile built upon the things he searches for, watches and likes, e-mails he exchanges, friends he has and their own digital profiles…

GePeTo provides John with precise speaking points —as usual— but this time it always draws on Catholic ideas. Although John is Catholic, he is not a practising one, and he would not initially have used these arguments. However, upon reflection, John decides to use GePeTo recommendations as he concludes that drawing on religion does make discussing death with a child easier. In doing so, however, he is also moved to reflect on his own beliefs, and in this particular case they end up being reinforced.

Is there something wrong in this use case? Some of us would feel uneasy with this RecSys, while others would argue it is perfectly fine, and even claim that GePeTo’s assistance is qualitatively similar to discussing the issue with a (human) friend. We believe answering this question would be easier if we were provided with some more conceptual clarity about what is at stake, and philosophy is particularly well equipped for this endeavour. What is clear at this stage is that: first, by using GePeTo, John did not think about different options himself, i.e., he did not use his own imagination to foresee different courses of actions and their effects; second, he chooses a line of action that he would have unlikely chosen otherwise; and third, this choice, in turn, has an influence on him regarding a particular domain of his own belief system. This touches upon aspects such as user competence or the beliefs and preferences construction that seem to be central to user autonomy. From our point of view this makes GePeTo an example of a system that at least should be evaluated from an ethical point of view before deployment. The next section introduces an approach for performing this kind of evaluation which differs from traditional practice in engineering and involves a philosophically grounded conceptual assessment.

3 Evaluation of technology: soft impacts and value sensitive design

The deployment of new technology in society should be carefully assessed and, in some cases, formally regulated. This is clear in areas like safety and security, which have specific policies and methodologies for dealing with it, such as certification specifications or safety standards, and well-identified authorities that enforce them. Safety and security issues are examples of what are known as “hard impacts” in the area of ethics of New and Emergent Science and Technology (NEST) or NESTethics (Swierstra 2015). Hard impacts are technological risks with a quantifiable probability of directly causing clear and noncontroversial harm.

Soft impacts in contrast describe the way technology affects norms and values. The impacts of technology in domains such as privacy, fairness, or—particularly important for us—human autonomy are soft impacts. The norms and values at play are normally tacit, imperceptible to us, embedded in our culture and, like a Heideggerian hammer, soft impacts become present once the referred norms and values have already been disrupted.[3] Soft impacts are more ambiguous and defined in a qualitative fashion, making them less suitable for a systematic identification and evaluation, which normally relies on the use of our imagination to envisage future techno-moral scenarios (Swierstra 2015).

[3] For example, most of us become aware of privacy issues after we start receiving ads related to information from the e-mails we send and receive or even the conversations we have.

A more practical approach to anticipating soft impacts might be found in the Value Sensitive Design (VSD) methodology, which has become popular in recent decades, particularly in the Information Technology (IT) domain. VSD strives to systematically include human values through the technological design process by deriving measurable requirements from more abstract analyses, seldom considered in the practice of engineering.

VSD[4] attempts to bridge the mostly normative areas of value analysis with the rather descriptive ones of requirements elicitation. To that end, it proposes an integrative and iterative tripartite methodology composed of conceptual, empirical and technical investigations. Conceptual investigations "comprise analytic, theoretical, or philosophically informed explorations of the central issues and constructs under investigation" (Friedman and Hendry 2019). Empirical investigations will force us to leave the "philosopher's armchair" and examine the "understanding, context and experiences" of stakeholders (Friedman and Hendry 2019) in relation to the technology under assessment, and the implicated values. Technical investigations are concerned with the specific features of the technology, aiming to prescribe the correct design to support specific values or analyse how particular features of existing technologies foster or hinder certain values in a context of use.

Value hierarchy diagrams are a useful way to structure a VSD analysis and the result thereof, making the translation of values into design more "systematic, […] explicit, debatable and transparent" (van de Poel 2013). Figure 1 depicts an often-cited approach for value hierarchies. It is addressed as a taxonomy with different levels in which the upper part consists of intrinsic, normally abstract and qualitative values; an intermediate layer consists of norms, in which the term refers to "all kinds of prescription for, and restriction on, action" (van de Poel 2013), which stems from the assessment of the impacts on the values at play, and there is a bottom layer concerned with context-dependent and preferably quantitative requirements, which are means of compliance of the norms previously elicited.

This way of structuring the results of a VSD analysis enables opening the design process to external stakeholders, normally not involved in these technical activities and, in turn, it makes the philosophical reflection about values at play accessible to design teams, usually uninvolved in these rather abstract discussions.

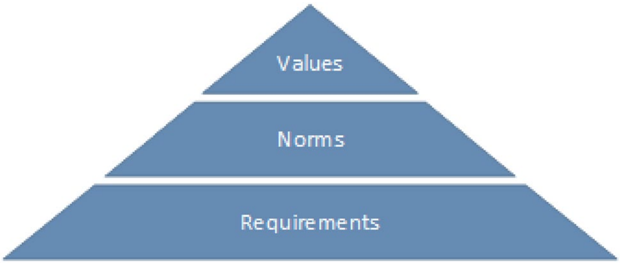

Fig. 1 Basic layers of a value hierarchy (van de Poel 2013)

VSD and value hierarchies seem a suitable approach for providing an evaluation framework as regards the impact of Recommender Systems on human autonomy, and in the rest of this paper we will endeavour to begin this assessment by focusing on the conceptual investigation.

4 Conceptual investigation: human autonomy and RecSys

4.1 Which autonomy? Procedural and relational accounts

Autonomy is generally understood as the "capacity to be one's own person, to live one's life according to reasons and motives that are taken as one's own and not the product of manipulative or distorting external forces, to be in this way independent" (Christman 2020). Contemporary discussions about human autonomy in political and moral philosophy usually pivot between its two main approaches: procedural and substantive autonomy, the latter normally taking the form of relational autonomy. Procedural autonomy is content free, it does not provide means to evaluate the desires, preferences, and so on, in virtue of which one is considered autonomous; it only "stresses internal self-reflection and procedural independence" (Christman 2020). It requires two conditions for a person to be considered autonomous (Christman 2004): competence, which is the capability required for governing oneself and includes "various capacities for rational thought, self-control and freedom from debilitating pathologies, self-deception and so on"; and authenticity, which includes "the capacity to reflect upon and endorse (or identify with) one's desires, values and so on" (Christman 2020). From a procedural point of view, valuing and evaluating autonomy is quite straightforward: autonomy is a value that should be preserved as it is directly linked with individual well-being and the focus should be made on preserving individual competence and authenticity. This conception of the value of autonomy has been further elaborated as the source of individual normativity (Korsgaard 1996), something that grounds our respect and

[4] There are many references that address VSD, including research papers aimed at its promotion (Friedman, Kahn, & Borning, Value sensitive design: Theory and methods., 2002) or criticising it (Manders-Huits, 2011), book chapters (Friedman, Kahn, & Borning, Value sensitive design and information systems., 2006) and even entire volumes devoted to this methodology (Friedman and Hendry 2019).

obligations towards others and vice versa. Finally, in yet a further development of this view, the value of autonomy is grounded in our reflexive capabilities: “we confer status and value on decisions simply because we reflectively made them” (Christman 2020).

Critics of procedural autonomy commonly see this focus on the individual, his self-reflection and procedural independence rooted in a modern conception of the self, which, in both essentialist and existentialist accounts,[5] is taken as fully rational and unaffected by the social or technical environment. Opposed to this, the constitutive model of the self posits that the person builds her identity embedded in—and in relation with—a particular socio-economic environment, including family, society, and institutions (MacKenzie 2008). This view is enriched by recent developments in the philosophy of technology that show how available technologies also have a key role in the subjectivation of the self (Verbeek 2011; Dorrestijn 2012) with new power relations embedded in their design and new actors, namely big technology companies that develop and deploy these technologies.

This is something that the basic procedural account of autonomy did not consider, as it allegedly approaches the individual as a person capable of transcending “her socialization, defining and reflecting on her values and commitments free of social influence” (MacKenzie 2008). Relational accounts of autonomy seek to prevent these issues, but this comes at a cost: autonomy is no longer considered an unconditional value, and its evaluation is shifted to the assessment of the social, economic and technological environment in which individuals are constituted. This path leads to a more abstract inquiry, less focused on individual capabilities and their evaluation, and more on the need for social action to transform society (Coeckelbergh 2022b), and—inevitably—to a perfectionist approach thereof.

4.2 The pragmatic turn

In recent decades there has been a trend in philosophy of technology that is especially useful for overcoming procedural-relational dualism: the pragmatic turn (Coeckelbergh 2020a; Keulartz et al. 2004), a revitalisation of pragmatism and, particularly, the application of John Dewey’s’ ideas to philosophy of technology (Hickman 1990). In regards to conceptual analyses, it consists of three characteristics: first, an instrumental account of concepts, which takes them as tools whose worth is based only on their explanatory power to a particular problem. As such, concept definitions should never be taken as the final foundation of any knowledge; on the contrary, they are always fallible and subject to change when our understanding of the world evolves. Second, pragmatism rejects “armchair philosophy”: pragmatic inquiry is theoretically based but had no value unless empirically confirmed. Third, pragmatism rejects well established dualisms in philosophy e.g., subject/object, means/ends, fact/value or, as we will argue, procedural/relational autonomy. According to pragmatism, these divisions are only useful as long as they help clarifying a problem and should be the outcome of empirical research, rather than precede it.

Therefore, in line with the pragmatic approach for a conceptual analysis of personal autonomy, we propose that the line separating the procedural and the relational accounts is blurred, and that we rather speak of a “relational load” of the components identified in the procedural account. We will also give weight to the context in which we are analysing personal autonomy: the use of Recommender Systems. In this way, our analysis will benefit from the strength of both models of autonomy: first, it will be focused on competence and authenticity and, therefore, on individual capabilities as the grounds for valuing and evaluating autonomy. Secondly, it will be sensitive to the social, economic, and—of the most relevance in our case—the technical influences and how these affect the above-mentioned components, using a more accurate model of the self than the one inherited from modernity.

4.3 A pragmatic account of autonomy

In the previous sections we have concluded that, to provide a concept of human autonomy in the context of RecSys that could be used as a conceptual investigation within a VSD assessment, we are going to use the procedural view of autonomy enriched with some relevant elements from the relational view. This provides two elements, competence and authenticity, that will form the basis of the conceptual analysis, but need to be sensitive to the technical environment in

[5] The essentialist model of the self posits a fixed or pre-existing inner-self, “which we must first uncover and then express or enact. On this picture, we are passive in the face of the authentic self, limited to activities of discovering and expressing it” (Walker & Mackenzie, 2015). This view also encompasses elements such as introspection, self-discovery, acknowledgement, self-expression and the like. In the existential model “the self is dynamic, created by our activities and decisions over time” (Walker & Mackenzie, 2015) and is characterised as reflective endorsement, self-creation and self-definition, amongst others.

Both models are questionable when they are held at their extreme. On the one hand, the essentialist one relies on the “empirically implausible notion that the self has a fixed essence” (Levy 2011), which can be fully accessed through introspection. On the other hand, the extreme existential model, best represented by Sartre, relies on an equally implausible “radical freedom” in the constitution of the self, recognising that “literally nothing –not their genes, not their past history, not their social relationships or their talents and skills, not morality and not God– stands in the way of their self-creation” (Levy, 2011).

which the self is constituted. This extended sensitivity to technical influences will rely on the Foucauldian-inspired constitutive model of the self that has been posited by leading scholars in the frame of the Technology Mediation Theory (Verbeek 2011). Here we find the balance between what is fixed—including both individual constraints and the relationships of power within a specific cultural and technological environment—and what is created, and how this balance constitutes the self: "[t]he latter Foucault, […] addressed the ways in which human beings can establish a relation towards structures of power. […] Humans are not only the objects of power here but also subjects that create their own existence against the background of and in confrontation with these powers" (Verbeek 2011). This is important as it lays bare that being influenced is unavoidable and provides some insight into the kind of relation that people should establish with the technology they use: a critical relation that requires them to acknowledge that technology is not neutral and that they should be aware of the sources of non-neutrality to regain, to a certain extent, the components of autonomy they give away when they use technological artifacts for some tasks.

4.3.1 Competence and its components

The competence condition in the procedural model of autonomy focuses on what is required to enable one's governance in compliance with one's preferences, desires and values. Competence can be analysed in a systematic way using basic decision theory[6] as the capability to accurately perceive a given situation and imagine different courses of action and their potential consequences so that one can rationally choose the option with the largest expected subjective utility. Therefore, in the context of decision making supported by Recommender Systems, basic competence requires:

- Information management, involving accurately perceiving the current situation —including one's self state—and possible actions, seeking and having access to relevant and manageable information, which should not be intentionally deceiving or based on misinterpretations or false assumptions.
- Deliberation, which implies self-control, as this kind of decision making should normally avoid impulsive actions, always enabling one to give reasons for decisions. Deliberation also requires minimal rationality and imaginative skills for understanding the implications of one's decisions, and the capability of envisaging alternative possible courses of action under at least some imaginable conditions (Christman 2004; Mackenzie and Walker 2015).

Relational accounts of autonomy normally focus on the socio-economic conditions that could restrict this capability. For example, it is commonly claimed that those raised in an oppressive society or having been denied minimal education will likely be unable to imagine alternatives just replicating "the oppressive social conditions that autonomous living is meant to stand against" (Christman 2004). However, we will instead focus on the fact that technology also has this enhancing or limiting role in user competence and thus autonomy. This technological non-neutrality is one of the central themes in philosophy of technology, especially in its post-phenomenology branch (Ihde 2009), and has been clearly described by Peter Paul Verbeek in his Technology Mediation Theory (Verbeek 2011). This theory posits that our experience and actions are mediated by the technology we use, which can no longer be seen as pure instruments, as good or bad as the person using it. Technology itself gives shape to individual moral subjectivity, turning itself into what has been referred to as moralising technology (Verbeek 2011). Recommender Systems mainly entail what is referred to as an "alterity relation" in Technology Mediation Theory. In this kind of mediation, the user interacts with the RecSys and the world stays in the background, and in doing so they delegate a significant part of their competence. First, information management can no longer be seen as individual capability limited by the user's capacity. The information is selected and represented in a certain way, which might now be driven by hidden interest from the system developer or another third party. In the same way, deliberation capabilities are somehow bypassed, particularly when the system fails to give meaningful reasons for their recommendations, or those reasons do not reveal the user's preferences.

Our assessment of the competence component of autonomy provides us with a finer granularity for dealing with two issues key to personal autonomy previously identified in the literature in the context of RecSys and virtual assistants in general, namely manipulation and cognitive degeneration.

Manipulation is a complex phenomenon, worsened with the use of autonomous and self-learning technology. Fortunately, in recent years several contributions in academia[7] have analysed the different types of influences that software agents might exert on their human users (Jongepier and Klenk 2022; Klenk 2020, 2019; Susser et al. 2019). A full review of this topic is beyond the scope of this paper, and we

[6] For an overview of basic decision theory, see e.g., (Russell & Norvig, 2010), which introduces the main concepts (actions, outcomes as states of the world, subjective utility…) in the context of AI development.

[7] We are grateful to an anonymous reviewer who suggested reviewing the latest references on this topic.

will focus on the types of manipulation that clearly lead to a diminishment of personal autonomy, i.e., those that intentionally and covertly influence the decision-making capability of people "by targeting and exploiting their decision-making vulnerabilities" (Susser et al. 2019).

Interestingly, some authors have argued that manipulation does not imply covertness and even that it does not always threaten autonomy. For example, Michael Klenk has pointed out that we can be aware of the potential influences and yet still use the system, and this should not be considered a loss of autonomy (Klenk 2019). For him, the key issue is not whether the manipulation method is hidden or not, but the fact that manipulation does not take the user's reason into account, which does not necessarily entail a threat to autonomy (Klenk 2022). Although we would agree with this view—and considering third-party goals is also important in our opinion—we would insist that covertness is central to the type of manipulation we intend to mitigate, the type that does lead to a certain loss of autonomy. In this way, we agree with (Susser et al. 2019) and argue that when someone is aware of being influenced, this knowledge becomes part of his decision-making process and should not be counted as manipulation—at least not the kind of manipulation (manipulation as loss of autonomy) that concerns us.

In a seminal paper on the interaction of Intelligent Software Agents (ISA)—a term that encompasses Recommender Systems—and human users, Christopher Burr and his colleagues provide a taxonomy that includes different forms of first order manipulations: deception, nudges and coercion (Burr et al. 2018).

Deception occurs when a decision is rationally made but based on false or misrepresented premises. Clickbait is a paradigmatic example of this, but we can think of less explicit and more elaborate cases of deception. For instance, in the example introduced in Sect. 2, GePeTo designers—maybe a religious organisation—might want to promote religious beliefs amongst children, which leads the system to provide John with the abovementioned suggestions based on Catholic ideas, supported by exaggerated views about their suitability against other compelling options. What is common in both cases, and arguably in all cases of deception, is that the information is intentionally[8]misrepresented. This is an impact on the user's competence, and particularly, on his information management skills, which should be avoided in all cases.

Deception could be mitigated if the designers of the system were required to publish its complete intended function and, in cases where a formal evaluation is needed, to provide the competent authorities with relevant design information about how this function is achieved. Due to its link to the system's intended functionality and technical design, deception is a hard impact, in accordance with Sect. 3, and can be largely solved by traditional means, for example, compliance with certification specifications, which in turn might require the development of industry standards to provide guidelines on the publication of this information for the general public and designated authorities.

Regarding nudges, these attempt to exploit known biases below the level of individual awareness, therefore effectively bypassing deliberation capability (Thaler and Sunstein 2008). While both nudges and deception are intentional, they differ substantially because in the former the information presented is correct, or at least not deliberately erroneous, unlike in the latter. In addition, nudges exploit not only the limited information management skills of users, but also their deliberation abilities and everything these involve (self-control, rationality and so on). Finally, nudges are commonly seen as a way of helping people to make better decisions, allegedly without preventing them from choosing otherwise, thus considerably limiting their effect on user autonomy.

Typical examples of nudging include using different sizes and colours for different options, arranging different suggestions in a particular way, or default opt-ins for some allegedly better options for the user. Returning to the RecSys described in Sect. 2, the designers of GePeTo—maybe another religious organisation—could legitimately believe that the spread of Catholicism will result in a better society and happier individuals. However, their strategy is based on a default opt-in for prioritising such ideas, always allowing the user to opt-out. This could be an example of a nudge that does not harm user autonomy as the advocates of libertarian paternalism claim (Thaler and Sunstein 2008). Let us now imagine that the system uses emotion recognition techniques to better adapt its recommendations, using a particular ordering, based on the predicted real-time preferences depending on the user's mood, without preventing him or her from choosing otherwise. Has any line been crossed here?

According to the account of autonomy we are using, any RecSys using nudges does indeed have an impact on user autonomy as regards competence capability in both information management and deliberation skills. Notwithstanding the potentially good objectives with respect to helping users to achieve better decision making, the system should clearly provide information about its use of nudges upon demand. Once again, this is a rather hard impact: it is directly linked to the intended functionality and technical design of the system, in this case, the design of its choice architecture, i.e. the representation of the space of options (Thaler and Sunstein

[8] The designer's intent is important for moral and practical reasons. A design that translates bad intentions from the designer is obviously morally questionable. From a practical point of view, there is a direct link between the system's designed intent, and whether its impacts should be considered hard or soft, and the kind of mitigations that should be provided —hard impacts will be mitigated mainly with technical means (see Sect. 3).

2008). A potential mitigation would entail complementing the information required in the previous case with detailed information about the design of the choice architecture and the ways the system allows the user to opt-out thereof.

Finally, with regard to coercion, it does not seem to be a problem in the RecSys context, as we can assume that no one is forcing the user to utilise the system. However, we can also see coercion as a self-inflicting effect of a person being so accustomed to utilising the RecSys that he or she becomes dependent on it, being unable to perform properly in its absence.

Dependency has been very well articulated by John Danaher as a potential harm to user well-being in the event the system becomes temporary or permanently unavailable, which is a consequence of user cognitive degeneration due to not exercising the cognitive capabilities that we are outsourcing to technology (Danaher 2018). As mitigation, Danaher suggests that as users we differentiate between instrumentally and intrinsically valued tasks to reflect on—and limit—the dependency issue. According to this author, the latter ones are those tasks directly linked with user well-being and characterised by a sense of reward when they are accomplished, whereas the former ones are those that enable us to obtain other good things, their outsourcing being less problematic.

While we agree with Danaher's approach and the mitigations he proposes, we consider that this would only cover the degeneration of user competence. We will argue that user cognitive degeneration goes beyond this and affects another important element of human autonomy: authenticity, whose assessment will require some more philosophical depth, as we shall see.

4.3.2 Authenticity as a reflection of one's personal identity

In the procedural account of autonomy, authenticity is defined as the capability to reflect upon and endorse the preferences, desires, and values that guide the agent's rational decision-making. We have already mentioned that the main weakness of this model is its apparent blindness to external influences in the development of these elements. Following our pragmatic approach, we will complement procedural authenticity with the most relevant elements from the relational account.

Modern authenticity, a very influential concept particularly for mid-XX century existentialists, has since then declined in popularity, being forcefully questioned from various angles. Especially in recent decades, different authors coming from schools of thought as diverse as liberalism, communitarianism, post-modernism, feminism,[9] etc., have stressed the relational load in the conception of human authenticity and —strongly related to this— in the concept of personal identity, which will be particularly useful to assess the impact of Recommender Systems.[10]

They all point out the fact that one's preferences, values, desires, and everything that confers meaning to one's existential experience and actions are not developed *ex-nihilo*. On the contrary, they are mediated by our language, moral rules, relatives, habits, social roles, institutions, and so on, which provide a "background of intelligibility" (Taylor 1991), against which things take on importance and are integrated—or rejected—into our personal identity upon reflection.

This is clearly described by Christine Korsgaard when she introduces the influential concept of practical identity as "a complex matter and for the average person there will be a jumble of such conceptions. You are a human being, a woman or a man, an adherent of a certain religion, a member of an ethnic group, a member of a certain profession, someone's lover or friend, and so on. And all of these identities give rise to reasons and obligations." (Korsgaard 1996).

This reflection on one's personal identity—"reflective endorsement" in Korsgaard's terms—is not only important for well-being in general, but it is also what confers normative power to one's decisions. And all this is lost when the decision-making process is delegated to a Recommender System, particularly those powered by AI/ML. First, because this very reflective endorsement is missing; second, because as a consequence, user identity is artificially developed and maybe —although not necessary intentionally— in misalignment with well-being; third, because the user model i.e., the user's personal identity modelled by AI/ML-powered systems, will be developed upon predefined categories, which in the worst case will be automatically generated and, as such, meaningless for the user. Let's unpack these aspects.

The first issue has already been identified and partially covered in Sect. 4.3.1., when discussing cognitive

[9] E.g., Christine Korsgaard (Korsgaard 1996) and Charles Taylor (Taylor 1991), who have been our main sources. Katja de Vries and her colleagues (de Vries 2010) draw on Paul Ricoeur's *ipse* and *idem* identities to explain how profiling technologies mediate the development of one's identity. Michel Foucault is unavoidable in current discussions on subjectivity and external influences and feminist philosophers such as Judith Butler are normally referenced when introducing anti-essentialist accounts of identity, particularly in the case of gender.

[10] Personal identity in philosophy refers to two distinct yet intertwined concepts: numerical identity and narrative identity (DeGrazia 2005). Numerical identity is related to one's psychological persistence and the criteria for one's continuous existence. Narrative identity focuses on the story an individual gives to himself, which is key to make sense of who he is and his actions. Personal identity in its narrative sense is strongly linked to the authentic creation of one's preferences, values and desires, which is what we will argue might be at risk with the use of Recommender Systems and will be the focus of our assessment.

degeneration relative to user competence. The problem is that RecSys are being used in an increasing number or areas, particularly since the disruption of AI, which also leads to multiplying the potential effects. We argue that some tasks and decisions are related to reflecting on one's identity and that the abovementioned cognitive degeneration would have more profound implications if these tasks were delegated, and such reflection bypassed. Indeed, the reflective endorsement introduced before is part and parcel of the development of one's personal identity i.e., there is a feedback effect on individual identity when one makes sense of an experience or an action, and this is bypassed when this reflection is delegated to a Recommender System. We complement the dualism of instrumentally vs. intrinsically valued tasks presented in Sect. 4.3.1. and suggest that those that should not be delegated (at least uncritically) are tasks that help to shape the important traits related to individual identity, and this mainly depends on the user on a case-by-case basis.

The second and third issues can be seen as a specific account of two well-known problems in the AI domain and inherent to the very development of this technology: alignment and explainability, which are analysed in the RecSys context. A key feature of AI in its Machine Learning variant is that these systems are not programmed but trained with a specific goal, their inner working being largely opaque for their users and even their designers. The training goal mainly depends on the RecSys purpose e.g., an entertainment RecSys such as YouTube's might aim at user engagement, whereas a search engine such as Google or a decision-making assistant such as our speculative example GePeTo could instead target single interactions, the opposite being an indicator of failing recommendations.

Our account of the alignment problem emphasises that these training goals have a partial—and sometimes questionable—link with user well-being, which is an essential element in decision-making, particularly when this leads to reflecting a specific trait of one's identity. Yet, these systems do have an impact on the development of user identity. A clear example is the case of entertainment RecSys trained to maximise user attention[11]: a posteriori, we have realised that the best way to succeed in this goal is to provide users only with what they like—sometimes in more extreme positions—resulting in an artificial polarisation and radicalisation of their preferences and values. Interestingly, in most cases of polarisation, system designers do not try to manipulate their users in a specific direction; polarisation is a non-intended effect of a system trained with a bad—although not necessarily malicious—goal.

We argue that this kind of system that has the power to influence the development of user identity should be trained with the well-being of users as a priority, clearly informing them whether or not other third-party goals are considered. The alignment problem could be mitigated with the publication of the system's goals —its utility function using technical terms— which would further complement the description of the system's intended functionality proposed in Sect. 4.3.1.: on the one hand, we would have a description of the intended design, and on the other, a description of how this design has been translated to a utility function in the context of an AI/ML based system development. Ideally, this would be based on internationally recognised standards —not existing today to our knowledge—that would guide the publication of this information.[12]

Even if the designers of the system made public its intended function and how this was implemented in the training process, with this all being somehow aligned with user well-being, we would still be facing the explainability issue. This is especially relevant when the system's recommendation technique is based on collaborative filtering, which as we mentioned in Sect. 2 is particularly important for our assessment. Indeed, this technique is based on a user model, which represents the user's preferences, values and desires. The problem is that this model is automatically generated in the training process, grouping user data, both explicitly provided (e.g., sex, age…) and implicitly gathered (e.g., previous interactions with the system and with other applications), into clusters.[13] These clusters are not related to any socially endorsed categories so, even if they were disclosed, they would be meaningless for the user. This has been very well articulated by Silvia Milano and her colleagues: "the labelling that the system uses to categorise users may not correspond to recognisable attributes or social categories with which the user would self-identify (for example, because machine-generated categories may not correspond to any known social representation), so even if users could access the content of the model, they would not

[11] This has been described in detail by James Williams in "Stand Out of Our Light" where he explains that the goal commonly guiding the design of certain Recommender Systems is maintaining the attention of users for as much time as possible i.e., "maximizing the amount of time you spend with their product, keeping you clicking or tapping or scrolling as much as possible, or showing you as many pages or ads as they can" (J. Williams, 2018).

[12] This cannot replicate the current approach of informing users about the system's terms of use i.e., endless pages of text that very few people read. Here we argue that there is a need for standardising the way the intended function and utility functions —in case the system in based on AI/ML— are published and that compliance with such a standard would be a suitable mitigation for this impact.

[13] We focus on unsupervised learning technique, which is more complex and —we believe— interesting. In supervised learning the system is trained to classify the user within a set of predefined classes selected by the designers e.g., race, age, religion, nationality and so on.

be able to interpret it and connect it with their lived experiences in a meaningful way." (Milano et al. 2020).

Our account of the transparency problem highlights the difference between using an opaque RecSys and the way individuals make sense of—and reflectively endorses—their own identity using given social categories and the reasons and obligations encompassed by these. We argue that this problem could be partially mitigated if the system was able to provide the working user model upon demand in a meaningful way. This would involve the use of Explainable AI techniques to translate such an opaque user model to e.g., a standard social category list in order for users to be able to access—and ideally modify—the model that the system has generated of them at any time. An avenue of multidisciplinary research could assess the standardisation of a set of social categories to build these predefined classes, along with a methodology for updating the list on a timely basis. Even if standardisation is a reductive approach and might be only applicable to specific cultures (e.g., Western social categories and their content might differ from the Asian ones), this standard would ease both the development of these systems and user education for them to make better sense of the information provided.[14]

Unlike the impacts assessed in Sect. 4.3.1., we claim that the impacts affecting user identity are soft ones. As such, their relationship with the system's intended function is not straightforward and, although we tentatively propose some technical mitigations, we do not believe that their solution relies solely on a better technical design. On the contrary, we would rather propose that in this case the burden should be mainly on users, being the first mitigation discussed above the most important one: users should always carefully consider whether they are using a RecSys for tasks that are intrinsically valued (in particular, if these shape important traits of their identity) and be aware of the potential consequences if they decide to delegate them, using all available information provided by the system.

5 A first draft of the evaluation framework

This section takes stock of the previous assessment and populates the value hierarchy introduced in Sect. 3, summarised in a table (Table 1) rather than a taxonomy for legibility. Since at this stage we are considering a generic Recommender System, the lower layers of the hierarchy can only be tentative and should be refined when a proper description of the system and its intended use is available. We also provide an initial allocation of the elicited technical requirements to the system's design elements identified in Sect. 2.

Let us illustrate the use of this framework with our speculative parenting RecSys introduced in Sect. 2. This framework operationalises the previous reflections about hard and soft impacts and prioritises the technical design and the adherence to agreed standards and regulations for the former while giving weight to the user's responsibility for the latter.

First, deception is a hard impact, directly linked with the intentions of the designers to deceive users, and how this has been translated to the system's design. This would be prevented upfront if designers were required to publish the intended functionality and the relevant information about how it is achieved. In addition, if this information were required by regulation and followed user-friendly standards, it would be hard for designers to keep using this practice, which is morally problematic in all cases.

The use of nudges would be more challenging as they also represent a hard impact caused by specific design, although not all nudges are morally questionable. In fact, many of them reflect good intentions on the part of designers. However, all nudges potentially diminish user autonomy. Let us assume that GePeTo has been developed by an institution that aims at spreading Catholic ideas. It is legitimate, as long as this aim is known, for the system to output a particular representation of the proposed solutions, prioritising those aligned with the system's intended function. GePeTo could also perform better if it used emotion recognition techniques (e.g., detecting John's anger or frustration in a given task or decision). According to our conceptual analysis, both cases are equally problematic regarding user autonomy. Thus, our framework requests that the system informs about all kinds of nudges used and ideally provides the user with the tools to opt-out at any time, making the manipulation mechanisms overt and thus defusing one of the main problems of this kind of autonomy-threatening influence, according to our analysis. In this way, John could agree, e.g., that the system filters out some recommendations but choose his emotional state not to be monitored despite GePeTo's likely loss of performance. Even if GePeTo's recommendations filtering were to still influence John in a certain way, as we have argued, once the manipulation mechanism becomes overt, it could no longer be considered a threat to his autonomy.

Our framework also considers soft impacts and provides some mitigations. Soft impacts are less directly linked to technical design, and much more difficult to grasp and anticipate and will unlikely be fixed solely with better technical design. In our assessment, the mitigations of these impacts are mainly allocated to the user. First, it highlights the

[14] We believe this represents a promising avenue for further multidisciplinary research. We consider that the standardisation proposals identified above i.e., standardisation of the way that developers publish the system's utility function, and standardisation a generic user models, that would be accessible to the users, are especially relevant and could be included in the standardisation roadmap of current organisations working in this domain (e.g., CEN/CENELEC, ETSI, IEEE).

Table 1 RecSys—Human autonomy value hierarchy

Values	Norms	Requirements (tech and non-tech)	
Competence • Information management • Deliberation	Deception in the use of a RecSys shall be avoided in all cases	Designers should publish -ideally in compliance with widely accepted standards- the RecSys' intended function, and the relevant design information about how this function is achieved	Tech (Function)
	The RecSys shall clearly provide information about its use of nudges upon demand	Designers should complement the information about the RecSys intended function case with detailed information about the choice architecture design and the ways the system allows the user to opt-out thereof	Tech (Function)
	Dependency due to competence degeneration shall be prevented for intrinsically valued tasks	Users should differentiate between instrumentally and intrinsically valued tasks: Intrinsically valued are those tasks directly linked to user well-being and characterised by a sense of reward when they are accomplished; instrumentally valued are those tasks that enable them to get other good things	Non-tech
Authenticity as Personal Identity	Reflective endorsement and reflexion on one's personal identity should be promoted	Users should avoid uncritically relying on RecSys for tasks and decisions that help shape important traits of individual identity. Awareness of the use of technology with a potential to affect one's personal identity construction in ways that are difficult to understand, anticipate and control	Non-tech
	Alignment: The system's training goal shall be aligned with the essential elements of reflective endorsement, particularly user well-being, and refrain from unduly considering third-party goals	Designers should publish the system's Utility Function, ideally in compliance with a standard to provide this information	Tech (Function)
	Transparency in the working user model shall be ensured	The system should provide information on the working user model in a meaningful way. Standardisation of social categories to provide this information—and fulfilment of the standard—may be a means of compliance	Tech (User model; Rec Technique)

importance of differentiating between intrinsically valuable tasks and instrumental ones and, extending an initial characterisation thereof, provides some tools to reflect on the intrinsic value of the tasks being delegated, giving weight in the context of personal identity to individual reflective endorsement.

In general, John should reflect on whether the tasks he is outsourcing to GePeTo are valuable in the sense that their completion and the effort invested in them lead to a subjective feeling of well-being and accomplishment. Particularly in the domain of parenting, and focusing now on John's personal identity, he could reflect on the roles that are related to the task being delegated e.g., "father", "son", "Catholic"… and whether the obligations they entail are important for him.

If following this reflection John decides to use the system anyway, our framework provides two proposals to mitigate the alignment and transparency problems. For the alignment issue, GePeTo's designers shall provide the details of the training goals. For example, we could imagine that GePeTo, a decision-making assistant, is trained to maximise single interactions as this is an indication that first proposals are considered satisfactory. GePeTo might also request its users provide an indication of a measure of their satisfaction in each transaction so that a measure of subjective well-being is somehow gathered. Finally, GePeTo is required to provide a description of John's working user profile at any time and in a meaningful way. For example, GePeTo could generate a narrative about John's profile; the proposed social categories standard list would ease providing and understanding this information. Ideally GePeTo would also allow John to make corrections to his user profile if needed. Although understanding these details would require a significant effort on the part of users, they would enable a better and more informed use of the system.

We have recently witnessed the roll-out of ChatGPT-3, an AI-powered multipurpose chatbot developed by OpenAI, which makes our GePeTo a less speculative example. Unlike GePeTo, ChatGPT-3 is not designed for a specific purpose, say parenting advice, but it can certainly be used to that end, as well as for many others. This makes ChatGPT-3 fall into our wide conception of recommender systems. Although the initial concerns about ChatGPT-3 deal with problems relating to privacy[15] and how it uses collected data, it might also represent a threat to user autonomy, and the framework presented in this paper is well-suited to make this assessment. This case is particularly interesting because ChatGPT-3 entails an extremely broad multistability, i.e., the capacity of the technology to be taken up for different uses and to be meaningful in different ways, namely dominant and alternative stabilities (Ihde 2009). In this context, hard impacts —linked to specific design characteristics— will have a lower weighting in the evaluation of loss of autonomy, while the analysis of soft impacts will be particularly important in this context. This includes our assessment of the risk of dependency and the issues regarding the constitution of the user's practical identity and their mitigations, which would be allocated, once again, to the user.

6 Conclusions

This paper aims at providing a conceptual assessment of human autonomy when using a generic Recommender System. We have drawn on recent work in the ethics of New and Emergent Science and Technology to highlight the importance of technological soft impacts (i.e., impacts on values such as autonomy) and the difficulty of their anticipation. Drawing on the VSD methodology, we have carried out a conceptual investigation of human autonomy in the context of a generic Recommender System use. We have provided a philosophically grounded, richer account of human autonomy, which has enabled us to proceed with a more nuanced assessment of the impacts of these systems thereof and the potential mitigations that could be proposed. Although some of these mitigations are ascribed to the technical design, we have argued that in other cases the burden is on users, particularly focusing on their awareness of the potential effects and mitigations. Although this would entail a significant effort from them, this would not be the first example of technology requiring a minimum education and even training from users.

We believe Recommender Systems are an extremely useful tool in many cases and they will continue to extend their influence on other aspects of our lives in the years to come. Hopefully we have convincingly shown that not everything should be recommended, and that people should be aware of the potential impact of this technology. Obviously, everything depends on context: the risk to autonomy is not the same when using Google maps for directions as when using GePeTo, our speculative parental RecSys. This nuance is already reflected in the European Commission's AI Act, which put stronger regulatory requirements in the so called "high risk" applications[16] and we are confident that our research could support this regulatory framework.

We believe preserving human autonomy should be afforded more importance than that given in current ethical guidelines. This should be especially salient when one is confronted with statements from top executives from high tech companies like Amazon: "Today you have to ask for things, but a lot of this asking is starting to fade into the background, because the AI is getting good enough that it's beginning to predict what I might want."[17] It is our opinion that we should have a feeling of unease with this view of the future. This is at least the case for us, as we consider that, in many aspects, it is very important, in the often-quoted words of Harry Frankfurt, "to want what we want to want".

Funding Funding for open access publishing: Universidad de Granada/CBUA. This work was supported by Agencia Estatal de Investigación (PID2019-104943RB-I00) FEDER/ Junta de Andalucía (B-HUM-64-UGR20). The authors have no competing interests to declare that are relevant to the content of this article.

Data availability Data sharing not applicable to this article as no datasets were generated or analysed during the current study.

Declarations

Conflict of interest The authors have no competing interests to declare that are relevant to the content of this article.

References

Burr C, Cristianini N, Ladyman J (2018) An analysis of the interaction between intelligent software agents and human users. Minds Mach 28(4):735–774. https://doi.org/10.1007/s11023-018-9479-0

[15] E.g., the recent banning of ChatGPT-3 in Italy due to privacy-related problemshttps://www.wired.com/story/italy-ban-chatgpt-privacy-gdpr/

[16] Regulatory framework proposal on AI: https://digital-strategy.ec.europa.eu/en/policies/regulatory-framework-ai

[17] See footnote 2.

Christman J (2004) Relational autonomy, liberal individualism, and the social constitution of selves. Philos Studies 117(1–2):143–164. https://doi.org/10.1023/b:phil.0000014532.56866.5c

Christman J (2020) Autonomy in Moral and Political Philosophy. The Stanford Encyclopedia of Philosophy. https://plato.stanford.edu/archives/fall2020/entries/autonomy-moral/. Accessed 29 Oct 2022

Coeckelbergh M (2020a) Introduction to philosophy of technology. Oxford University Press, Oxford

Coeckelbergh M (2020b) AI Ethics. MIT Press, Cambridge

Coeckelbergh M (2022a) The political philosophy of AI: an introduction. John Wiley & Sons, Hoboken

Coeckelbergh M (2022b) Self-improvement: technologies of the soul in the age of artificial intelligence. Columbia University Press, Columbia, p 152

Danaher J (2018) Toward an ethics of AI assistants: an initial framework. Philos Technol 31(4):629–653. https://doi.org/10.1007/s13347-018-0317-3

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

DeGrazia D (2005) Human identity and bioethics. Cambridge University Press, Cambridge. https://doi.org/10.1017/CBO9780511614484

Dorrestijn S (2012) Technical mediation and subjectivation: tracing and extending foucault's philosophy of technology. Philos Technol 25(2):221–241. https://doi.org/10.1007/s13347-011-0057-0

Friedman B, Hendry DG (2019) Value Sensitive Design: Shaping Technology with Moral Imagination. MIT Press, Cambridge

Friedman B, Kahn PH, Borning A (2002) Value Sensitive Design: Theory and Methods

Friedman B, Kahn PH, Borning A (2006) Value sensitive design and information systems. In: Zhang P, Galletta D (eds) Human-computer interaction and management information systems: Foundations. M. E. Sharpe, New York

Giubilini A, Savulescu J (2018) The artificial moral advisor. The "ideal observer" meets artificial intelligence. Philos Technol 31(2):169–188. https://doi.org/10.1007/s13347-017-0285-z

Harari YN (2018, September 14) Yuval Noah Harari: The myth of freedom. The Guardian. https://www.theguardian.com/books/2018/sep/14/yuval-noah-harari-the-new-threat-to-liberal-democracy. Accessed 29 Oct 2022

Hickman LA (1990) John Dewey?s pragmatic technology. Indiana University Press, Bloomington

Ihde D (2009) Postphenomenology and technoscience: the peking university lectures. State University of New York Press, Albany

Jobin A, Ienca M, Vayena E (2019) Artificial Intelligence: the global landscape of ethics guidelines. Nat Mach Intell 1(9):389–399. https://doi.org/10.1038/s42256-019-0088-2

Jongepier F, Klenk M (eds) (2022) The philosophy of online manipulation. Routledge, New York

Keulartz J, Schermer M, Korthals M, Swierstra T (2004) Ethics in technological culture: a programmatic proposal for a pragmatist approach. Sci Technol Human Values 29(1):3–29. https://doi.org/10.1177/0162243903259188

Klenk M (2019) *Autonomy and online manipulation*. Internet Policy Review. https://policyreview.info/articles/news/autonomy-and-online-manipulation/1431. Accessed 26 May 2023

Klenk M (2020) Digital well-being and manipulation online. Ethics of digital well-being. A multidisciplinary approach. Springer, New York, pp 81–100

Klenk M (2022) (Online) manipulation: sometimes hidden, always careless. Rev Soc Econ 80(1):85–105. https://doi.org/10.1080/00346764.2021.1894350

Korsgaard CM (1996) The sources of normativity. Cambridge University Press, Cambridge

MacKenzie C (2008) Relational autonomy, normative authority and perfectionism. J Social Philos 39(4):512–533. https://doi.org/10.1111/j.1467-9833.2008.00440.x

Mackenzie C, Walker M (2015) Neurotechnologies, personal identity, and the ethics of authenticity. Handbook of Neuroethics. Springer Netherlands, Dordrecht, pp 373–392

Milano S, Taddeo M, Floridi L (2020) Recommender systems and their ethical challenges. AI Soc 35(4):957–967. https://doi.org/10.1007/s00146-020-00950-y

Manders-Huits N (2011) What Values in Design? The Challenge of Incorporating Moral Values into Design. Sci Eng Ethics 17:271–287. https://doi.org/10.1007/s11948-010-9198-2

Ricci F, Rokach L, Shapira B (2015) Recommender systems: introduction and challenges. In: Ricci F, Rokach L, Shapira B (eds) Recommender systems handbook. Springer US, Boston, pp 1–34

Susser D, Roessler B, Nissenbaum H (2019) Technology, autonomy, and manipulation. Internet Policy Review. 8(2). https://policyreview.info/articles/analysis/technology-autonomy-and-manipulation

Swierstra T (2015) Identifying the normative challenges posed by technology's 'soft' impacts1. Etikk i Praksis 9(1):5–20

Taylor C (1991) The Ethics of Authenticity. Harvard University Press, Cambridge

Thaler RH, Sunstein CR (2008) Nudge. Penguin, UK

van de Poel I (2013) Translating values into design requirements. Philos Eng Technol 15:253–266. https://doi.org/10.1007/978-94-007-7762-0_20

Verbeek PP (2011) Moralizing Technology: Understanding and Designing the Morality of Things. University of Chicago Press, Chicago

Wylie C (2019) *Mindf*ck*. Profile Books, London