Self-determination through explanation: an ethical perspective on the implementation of the transparency requirements for recommender systems set by the Digital Services Act of the European Union

Matteo Fabbri
matteo.fabbri@imtlucca.it
IMT School for Advanced Studies
Lucca, Italy

ABSTRACT

In the contemporary information age, recommender systems (RSs) play a critical role in influencing online behaviour: from social media to e-commerce, from music streaming to news aggregators, individuals are constantly targeted by personalized recommendations suggesting contents that may interest them. Despite such diffusion, the extent to which recommendations influence users' decisions is still underexplored, given that independent audits on the structure and functioning of RSs deployed on online platforms are usually prevented by proprietary constraints. The nudging potential of RSs can represent a risk for vulnerable people: indeed, judicial cases involving platforms' responsibility for displaying recommendations that may lead to political radicalization or endangerment of minors have recently caught public attention. The Digital Services Act of the European Union (DSA) is the first supranational regulation that sets specific transparency and auditing requirements for RSs implemented by online platforms with the aim of enhancing users' self-determination: in particular, it allows users to modify the parameters on which recommendations rely so to let them choose autonomously which kind of content they want to see. This research focuses on whether and how the enforcement of this regulation can mitigate the unfair consequences of the power imbalance between online platforms and users. To this aim, I discuss the harms arising from digital nudging based on RSs and propose explanations as a tool that can reduce the impact of those harms by increasing users' awareness. Through a comparative analysis of relevant articles of the DSA, the General Data Protection Regulation (GDPR) and the AI Act, I outline how the provisions of the DSA fill some of the gaps left by other relevant European regulations, while leaving the so-called right to explanation substantially unaddressed. As a result of this analysis, I argue that, in order for the implementation of the DSA provisions on recommender systems to be effective, policy-makers should: 1) enhance users' awareness through clear and easily accessible explanations on how the recommendation process works and how they can be influenced by it; 2) grant users the possibility of intervening directly on the strategies through which RSs target them on the platform's interface.

AIES '23, August 08–10, 2023, Montréal, QC, Canada

ACM ISBN 979-8-4007-0231-0/23/08…$15.00
https://doi.org/10.1145/3600211.3604717

CCS CONCEPTS

• **Social and professional topics** → *Computing / technology policy*.

KEYWORDS

Regulation of AI, Recommender Systems, Digital Services Act, Transparency, Digital Nudging

ACM Reference Format:
Matteo Fabbri. 2023. Self-determination through explanation: an ethical perspective on the implementation of the transparency requirements for recommender systems set by the Digital Services Act of the European Union. In *AAAI/ACM Conference on AI, Ethics, and Society (AIES '23), August 08–10, 2023, Montréal, QC, Canada.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3600211.3604717

1 INTRODUCTION

In the contemporary information age, recommender systems (RSs) play a crucial role in determining the way in which people interact and obtain information online: from social media feeds to news aggregators and e-commerce websites, users are constantly targeted by personalized recommendations about contents or products they may like. From a technical perspective, RSs can be defined as algorithms aimed at estimating predictive ratings for some items which a user has not seen yet (Adomavicius and Tuzhilin, 2005) [6] in order to generate recommendations about content which may interest them. The Digital Services Act of the European Union (DSA) [1] [5], which is the first supranational regulation addressing automated recommendations specifically, defines RS as "a fully or partially automated system used by an online platform to suggest in its online interface specific information to recipients of the service or prioritize that information, including as a result of a search initiated by the recipient of the service or otherwise determining the relative order or prominence of information displayed" (DSA, art. 3 (s)). This definition highlights the method ("fully or partially automated"), aim ("to suggest"), content ("specific information"), target ("recipients of the service"), input ("as a result of a search initiated by the recipient") and output ("determining the relative order or prominence of information displayed") of a recommendation process. As it can be observed, RSs concern the main aspects

[1] REGULATION (EU) 2022/2065 OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL of 19 October 2022 on a Single Market For Digital Services and amending Directive 2000/31/EC (Digital Services Act).

of the user's experience: this is why their influencing potential should not be underestimated. In fact, whilst RSs should be aimed at improving user experience, they can give rise to a variety of ethical concerns related to privacy, autonomy and fairness [21], to name but a few. Indeed, the political economy of platforms based on profiling and recommendations has been notably addressed by [34] with the concept of "surveillance capitalism". However, independent research and ethical auditing on the design and functioning of the RSs implemented on online platforms is usually prevented by proprietary constraints.

For these reasons, there is a normative discrepancy between the widespread use of RSs in various domains and the methods through which their ethical and societal impact can be evaluated. Issues related to transparency and explainability have become increasingly pressing, as the implementation of opaque models may have problematic consequences on the users' ability to retrieve relevant information and define their online identity. As algorithmic recommendations often rely on implicit personal data, such as browsing and click-through history, and their functioning is not explained to users, their influence is not accountable. Although explanations for RSs have been addressed by research in Explainable AI [31], their effects on the design of algorithms and on the different stakeholders within the recommendation process have not been assessed extensively. Moreover, even when explanations are provided in real-world platforms, users are not able to interact explicitly with them, apart from providing limited feedback. The limitations regarding the transparency and accountability of automated recommendations are supposed to be addressed by the provisions of the DSA, which would require very large online platforms, including marketplaces and social media, to let users shape the design of the RSs managing their online experience. However, the effectiveness of the application of the regulatory provisions will depend on the extent to which people understand how RSs work and how they can shape their functioning: therefore, explanations should have a prominent role in this context.

In this paper, I focus on whether and how the new European regulatory context around RSs can address the risks and opportunities stemming from this pervasive digital technology, especially from the perspective of mitigating the unfair consequences of the power imbalance between platforms and users. Firstly, I discuss the possible harms arising from RSs as instances of digital nudging and introduce explanations as a tool that can reduce the impact of those harms by increasing users' awareness. Secondly, I consider the impact of the DSA provisions about RSs and online targeted advertising within the regulatory context set by relevant articles of the AI Act (AIA) and the General Data Protection Regulation (GDPR) of the European Union. This comparative analysis outlines how the provisions of the DSA fill some of the gaps left by other European regulations, while substantially lacking measures to effectively enhance users' autonomy. As a result of this analysis, I argue that, in order for the aims of the DSA provisions about RSs to be fulfilled, the principle of users' self-determination needs to be substantiated by: 1) easy accessibility of explanations on how the recommendation process works and how users can be influenced by it; 2) an extended possibility for users to intervene directly on the strategies through which RSs target them on the platform's interface.

2 CONTEXT

2.1 From personalization to epistemic fragmentation

[21] propose an initial taxonomy of the ethical challenges posed by automated recommendations: among the social effects of RSs, they identify a "lack of exposure to contrastive views", giving rise to the so-called filter-bubbles, which can be exploited by manipulative agents in order to increase the frequency with which a content is recommended within specific online communities. [8] put in evidence, phenomena such as polarization on social media arise because of a subtle manipulation of the contents delivered individually but spread collectively by RSs: through strategic content tagging and by exploiting the networked structure of platforms, political campaigners may be able to redirect public attention on controversial contents which appear on the social media feeds of users. In this regard, [29] has famously pointed out the widespread political implications of digital technologies, including RSs, which have allowed people to "filter what they want to read, see, and hear", not coming "across topics and views that you have not sought out".

In fact, the concept of recommendation is inherently related to that of personalization, although the corresponding phenomena are distinct. In fact, the latter represents the pre-condition for the former. On the one hand, recommendations make sense only if they can be personalized, because, if they were not personalized, they would not be able to reduce the information overload on platforms, which is their main utility for users and providers [14]. On the other hand, personalization can be applied mainly through algorithmic recommendations (in the form of targeted advertisements, suggested contents, etc.): therefore, even if personalization as a design concept makes sense independently of recommendations, its application within the infosphere often relies on them. Therefore, automated recommendations depend on personalization, whilst personalization is embedded within recommendations from the perspective of its application.

This distinction is required in order to understand how the socio-technical structure of RSs is related to the epistemic fragmentation of users [20], a prominent problem in online platforms. Epistemic fragmentation can be defined as the phenomenon by which individual users lose contact with their peers through online targeted advertising. In particular, as each user is targeted individually by automated recommendations, one cannot know which content another person sees: in this sense, users' knowledge about their common experience on the platform is fragmented, because what they see is the result of personalization and cannot be shared among different individuals. This aspect is even more relevant considering that the effects of personalization do not necessarily imply that each user sees a different array of contents. In fact, an analysis of news recommendations on Google News by [23] found that "users with different political leanings from different states were recommended very similar news".

Epistemic fragmentation is not only a result of the individualization of recommended contents, but it also derives from the opaqueness of the recommendation process, which prevents users from becoming aware of the platform dynamics. This situation can give rise to ethical concerns especially when personalisation is based on implicit user profiling, through which "the system determines what the user is interested in" thanks to implicit data, which include "web usage mining [. . .], IP address, cookies" and other metadata [7]. Indeed, if a user is profiled through implicit data, the recommendations will be less transparent and explainable compared to a situation in which "the user customizes the information source himself" (ibidem) by providing explicitly data such as personal interests, demographic information and ratings. In the context of an epistemically fragmented user experience, the influence of RSs relying on implicit profiling may hold negative aggregate social implications. In fact, when users do not have control over which kind of data is used for their profiling, the recommendations are more likely to bring unwanted contents to their attention.

As a result, users may suffer, on a first dimension, from absolute harms of inclusion or exclusion, which "originate in the nature of the content that is either included or excluded from what is shown to an individual consumer" [20]: the former occur when genuinely bad and offensive contents (i.e. false claims or racist stereotypes used for promotional purposes) are displayed on the users' profile, whilst the latter occur when essential contents (i.e. important public heath announcements) are omitted, without the user's consent or control on the process. On a second dimension, users can be affected by contextual harms of inclusion or exclusion, which "do not stem from the nature of the content per se, but depend on the context in which the content is delivered" (ibidem): for example, a contextual harm of inclusion may occur when unhealthy food is suggested to obese people or children, who may be more likely to buy them; conversely, a contextual harm of exclusion can be recognised when a job-seeker does not encounter advertisements for positions in their area. The categories of harms produced by RSs do not arise only from implicit profiling but may also be a consequence of the data that users choose to provide explicitly. For example, a user may want to provide explicit data about personal unhealthy habits, such as gambling, because they are interested in finding products or offers in the related domain, regardless of their impact on wellbeing. In the same way, some users may give a high rating to recommendations about contents featuring stereotypes that other people may find offensive or unethical: if the latter share interests with the former, they may see such unwanted recommendations due to collaborative filtering algorithms. These cases show that even personalization based on explicit profiling may originate unexpected harms, which cannot be evaluated just from the point of view of the single user but need to be interpreted within the context of both the platform environment and the socio-technical structure of RSs. Therefore, the harms generated by personalized recommendations do not depend only on the individual case of application, but also on the policy informing the system.

2.2 Digital nudging and recommendation policies

The origin of harms caused by RSs lies in their potential to influence users' choices. In particular, since algorithmic recommendations "influence which information is easily accessible to us and thus affect our decision-making processes though the automated selection and ranking of the presented content", they can be interpreted "as digital nudges, because they determine different aspects of the choice architecture for users" [17]. According to the original definition in behavioural economics proposed by [30], nudges are the features of a choice architecture "that have an influence on which decisions people make" [17]. Nudging "should be aimed at helping people make better decisions than they probably would if the nudge would not be there" (ibidem) without forcing them to adopt a specific choice. The nudging potential of RSs depends on the effectiveness of the recommendation policies implemented in the algorithmic design, which usually rely on the exploitation or exploration of the space of choices.

An exploitative policy aims "to recommend an item that has the highest expected probability of satisfying the user's preferences" [22], whilst an explorative policy is focused on recommending "content with uncertain predicted user engagement for the purpose of gathering more information" about users' interests [19]. When RSs rely exclusively on exploitative policies, users can be led into feedback loops that may reinforce their current preferences, resulting in bad consumer choices in the long run. For example, a user that usually buys unhealthy food through a delivery app based on exploitative RSs may receive recommendations about the same kind of food every time they want to make an order and therefore their health could be impacted negatively. In this case, an explorative policy could instead propose different kinds of products that do not correspond to the preferences previously expressed by the user, eventually inducing them to find healthier food they like.

Since the aim of RSs is to recommend items which users may purchase or consume, it is relevant to know whether and how explanations, which stem mainly from explicit profiling, can impact on the users' perception of the recommendation and their subsequent behaviour. This issue relates to the harms of inclusion and exclusion described above: indeed, if the system manages to change users' interests through explanations, they will end up seeing different contents from the ones they were originally aiming for. Nonetheless, this may make them perceive to have been assigned to categories which they think they have willingly chosen to belong to, given that the recommendation is seemingly transparent because of explanations. The risks coming from the manipulation of users' preferences are intrinsic to RS-powered digital nudging, but [17] report finding no paper about whether "users felt manipulated or coerced by the proposed nudge". In this context, understanding the extent to which users are influenced by recommendations, on the one side, and their explanations, on the other, is crucial for the assessment of the impact that the current and upcoming regulations will have as regards transparency and self-determination.

The default integration of information about the content within the recommendation could be beneficial for users' awareness of their own preferences. Following the same example as above, a food recommendation might be designed so that the nutritional values of a product that a user has (exploitative policy) or has not (explorative policy) bought before are displayed to them before they can proceed to the order: in this way, the user could be informed about the characteristics of their dietary choices. Moreover, providing explanations would make users aware of the extent to which their preferences have been taken into account by the policy informing the recommendation. Although their impact on users' decision-making is still underexplored, explanations for RSs can be considered a kind of pro-ethical informational nudging [11], as they improve user-system interaction in direction of transparency and trustworthiness just through the provision of information. In fact, [17] classify explanations as nudging mechanisms within the Decision Information category based on making information visible.

3 REGULATORY FRAMEWORKS

3.1 From platform to court

The classification of harms presented above covers different cases in which automated recommendations would have a negative impact on users' wellbeing. The wide-ranging implications of harms caused by RSs go beyond the individual, acquiring a societal relevance. A case of absolute harm of inclusion covered by the international press concerns the "blackout challenge" on TikTok, which encourages users to film themselves as they choke themselves to the point of fainting and then regain consciousness on camera: various cases emerged in which minors died while trying the challenge. After the most recent cases, which happened in the USA [10] and UK [28], some American families decided to sue the platform as it let the challenge spread and target children through its recommendation algorithm [18]. While this may at first seem a problem of content moderation, it is, at a deeper level, a consequence of the use of RSs in social media platforms, where their main aim is to increase users' engagement. As RSs are often based on uninterpretable machine learning models, it might be difficult to attribute the liability for the harm to the platform. In fact, the platform could argue that contents are displayed to users according to recommendation policies that take their preferences into account, so, if the user liked or kept consuming a harmful content which is later reproposed to them, the system should not be blamed. Moreover, as access to the platform by individuals under a certain age should be supervised by parents, it is the parents' duty to control the online activity of their children. To challenge this argument, one should prove that it is the recommendation policy itself to be biased towards contents aimed at maximizing engagement regardless of the vulnerability of the user: according to this perspective, the platform would be liable for designing RSs that influence users' behaviour to fulfil the interests of the system (DSA, art. 35).

A related argument about platforms' responsibility for the content suggested by their RSs is embraced by petitioners in the Gonzalez vs Google case, which deals with "whether Section 230 [of the US Communication Decency Act] shields Google from liability for allegedly recommending ISIS content posted to YouTube to other YouTube users" [9]. This lawsuit emerged as a result of the deaths caused by the 2015 terrorist attacks in Paris, France, which were carried out by people recruited by ISIS after being exposed to social media content disseminated by the organization through YouTube RSs. In particular, the question posed to the US Supreme Court concerns whether "section 230(c)(l) immunize[s] interactive computer services when they make targeted recommendations of information provided by another information content provider, or only limit the liability of interactive computer services when they engage in traditional editorial functions (such as deciding whether to display or withdraw) with regard to such information" [24]. Petitioners argue that "Section 230(c)(1), which shields intermediaries from liability for "publishing" third-party content, applies only to claims based on the "display" of content, not the "recommendation" of content" (ibidem). In May 2023, the Supreme Court dismissed the case on the ground that it could not by addressed by antiterrorism law, as the "plaintiffs' complaint seems to fail under [. . .] our decision in Twitter" vs Taamneh, which concerned the same issue of Gonzalez vs Google [25]. If the Supreme Court's ruling had excluded targeted recommendations from the protection provided by Section 230, implying that "the "recommendation" of content is different from the display of content", platforms would have been forced to change their moderation and recommendation processes and users might have lost their "rights to like and promote content in forums where they act as community moderators and effectively boost some content over other content" [27]. As the old debate between freedom of expression and (online) safety eventually focuses on the impact of the influence of RSs, it is crucial for users to understand how algorithmic recommendations function and to shape their design. In fact, the prerequisite for users' self-determination is the knowledge of the sociotechnical systems with which they interact.

3.2 Digital Services Act (DSA): filling the gap left by the AI Act

The DSA addresses this issue with a specific article, according to which "Providers of online platforms that use recommender systems shall set out in their terms and conditions, in plain and intelligible language, the main parameters used in their recommender systems, as well as any options for the recipients of the service to modify or influence those main parameters" (DSA, art.27 (1)). The aim of this provision is to "explain why certain information is suggested to the recipient of the service": therefore, the parameters need to include, at least, "the criteria which are most significant in determining the information suggested to the recipient of the service" (i.e., content) and the reasons for its "relative importance" (i.e., ranking) (DSA, art. 27 (2)). Additionally, when options to modify or influence the main parameters are stated in the terms and conditions, "providers of online platforms shall also make available a functionality that allows the recipient of the service to select and to modify at any time their preferred option" (DSA, art. 27 (3)). In order to make this requirement work in practice, "That functionality shall be directly and easily accessible from the specific section of the online platform's online interface where the information is being prioritised" (ibidem). Moreover, "providers of very large online platforms [VLOPs] and of very large online search engines [VLOSEs] that use recommender systems shall provide at least one option for each of their recommender systems which is not based on profiling"[2] (DSA, art. 38). It is worth noticing that, while the provisions of Article 27 apply to all online platforms, the application of Article 38 is limited to VLOPs and VLOSEs, which therefore represent the only environments in which users will always have the option to choose between at least two types of recommendations[3].

The provisions of Article 27 aim to address four of the aspects of the definition of RS provided by Article 3(s): method, target, input and output. In particular, as a result of the enforcement of the DSA,

[2] Profiling is defined here according to Article 4 (4) of the GDPR.

[3] It is plausible to state that all the VLOPs and VLOSEs identified by the European Commission (https://digital-strategy.ec.europa.eu/en/policies/dsa-vlops) use profiling for automated recommendations, so the provision of Article 38 applies to all of them.

the traditionally passive role of the target might be reversed, as the recipient could determine the method (through the choice of parameters) and, indirectly, also the input (the type of data to be processed through the parameters) that the RS would use to produce its output. This opportunity to enhance transparency and users' self-determination has not been welcomed by a prominent digital company like Meta, which has stated that "the breadth of some of the auditing obligations under the DSA should be clarified/improved as these could become a barrier for growth in the sector" [2]. However, online platforms that are not VLOPs or VLOSEs using RSs based on profiling will not be obliged to provide options for users to modify or influence the parameters if this possibility is not specified in the terms and conditions, and platforms arguably have no interest in providing this possibility voluntarily. Therefore, Article 27 formally grants users the right to influence the recommendation process but only in some limited cases which may not be likely to happen, as [15] point out. Moreover, the practical impact of these provisions will probably depend on users' ability to understand the type and the policy of recommendations.

The rationale of the norms on RSs transparency, introduced in Recital 70, outlines a wider regulatory scope than the one of Article 27: indeed, the statement that "online platforms should consistently ensure that recipients of their service are appropriately informed about how recommender systems impact the way information is displayed, and can influence how information is presented to them" (DSA, recital 70) does not seem to be reflected in the actual provisions of Article 27, at least to the extent that the adverb "consistently" would entail[4]. Nonetheless, online platforms "should clearly present the main parameters for such recommender systems in an easily comprehensible manner to ensure that the recipients understand how information is prioritised for them" (ibidem). A right to explanation for RSs could be identified in this formulation: in fact, the "easily comprehensible manner" of presenting the parameters of RSs so that "the recipients understand how information is prioritised for them" can come to effect only if RSs are explainable.

Relatedly, the DSA will also require VLOPs that display advertisements to "compile and make publicly available in a specific section of their online interface, through a searchable and reliable tool that allows multicriteria queries and through application programming interfaces, a repository" (DSA, art. 39 (1)) featuring the following information: "(a) the content of the advertisement, including the name of the product, service or brand and the subject matter of the advertisement; (b) the natural or legal person on whose behalf the advertisement is presented; (c) the natural or legal person who paid for the advertisement, if that person is different from the person referred to in point (b); (d) the period during which the advertisement was presented; (e) whether the advertisement was intended to be presented specifically to one or more particular groups of recipients of the service and if so, the main parameters used for that purpose including where applicable the main parameters used to exclude one or more of such particular groups; (f) the commercial communications published on the very large online platforms [. . .]; (g) the total number of recipients of the service reached and, where applicable, aggregate numbers broken down by Member State for the group or groups of recipients that the advertisement specifically targeted." (DSA, art. 39 (2)). The first four points of the cited paragraph concern the metadata of the advertisement: its content, who paid for it, the duration of its permanence on the platform. According to point (e), the platform is required to indicate whether the advertisement was targeted and, if so, the main parameters used for including or excluding categories of users from the targeted. Point (g) would allow to understand indirectly the correspondence between specific clusters of users and the advertisement by which they have been targeted in each EU country. The enforcement of this article has the potential to address the epistemic fragmentation of users due to online targeted advertising considered by [20]. Indeed, if users can access a public repository with information about the parameters used by platforms to segment them into groups for targeting purposes, they can have an idea of how many other people see a particular advertisement and why they see it. The access to this information can reduce the individualization and fragmentation of online experience, as users could eventually become aware of collective platform dynamics, although probably not at a very granular level.

The provisions outlined above are part of a wider regulatory scope. In particular, the DSA aims to address the systemic risks and harms that may emerge from the implementation of RSs in VLOPs and VLOSEs so to avoid violation of fundamental rights and the endangerment of vulnerable people like minors. According to Article 34, "Providers of very large online platforms and of very large online search engines shall diligently identify, analyse and assess any systemic risks in the Union stemming from the design or functioning of their service and its related systems, including algorithmic systems, or from the use made of their services", including: "(a) the dissemination of illegal content through their services"; "(b) any actual or foreseeable negative effects for the exercise of the fundamental rights [. . .] to human dignity", "to respect for private and family life", "to the protection of personal data", "to freedom of expression and information", "to non-discrimination", "to respect for the rights of the child" and "to a high level of consumer protection"; (c) "any actual or foreseeable negative effects on civic discourse and electoral processes, and public security"; (d) "any actual or foreseeable negative effects in relation to gender-based violence, the protection of public health and minors and serious negative consequences to the person's physical and mental well-being". The risks assessments operated by very large online platforms should take into account, among other aspect, "the design of their recommender systems and any other relevant algorithmic system" (DSA, art. 34(2), which will need to be adapted following risk mitigation measures (DSA, art. 35(1)).

Following the unprecedented regulatory scope of the DSA, the European Commission has founded the European Centre for Algorithmic Transparency (ECAT), whose mission is to contribute with "scientific and technical expertise to the Commission's exclusive supervisory and enforcement role of the systemic obligations on Very Large Online Platforms (VLOPs) and Very Large Online Search Engines (VLOSEs) provided for under the DSA" [1]. The area of competence of the ECAT features "recommender systems, information retrieval and search engines", which will be the subject of research aimed at uncovering their "ethical, economic, legal and social impact" and at developing risk assessment and mitigation

[4] The right to information outlined here is mirrored by Article 13-15 of the GDPR, which will be considered later.

measures for the protection of fundamental rights (ibidem). Such research effort would provide an evidence base for the implementation of the DSA, whose high-level provisions regarding RSs are not currently backed by standards that can bridge the gap between regulatory principles and market practices. The ECAT will also include an inspections team which “will actively help assessing whether very large online platforms and search engines comply with their obligations under the Digital Services Act” by "analysing the design, functioning and impact of advanced algorithms, like recommender systems, in their production environments" through "formal investigations" including "on-site inspections at platforms' premises" (ibidem).

The provisions of the DSA fill the gaps of the EU Artificial Intelligence Act (AIA) [3] concerning RSs. Before the latest amendments approved by the European Parliament in June 2023, references to automated recommendations could be found in only two paragraphs of the AIA proposal: the first occurrence is the definition of AI system as “software that can, for a given set of human-defined objectives, generate outputs such as content, predictions, recommendations, or decisions influencing the environments they interact with” (AIA, art. 3(1)); the second occurrence is the explanation of “automation bias” as the tendency of “automatically relying or over-relying on the output produced by [. . .] AI systems used to provide information or recommendations for decisions to be taken by natural persons” (AIA. art. 14(4b)). In both the occurrences, “automated recommendations are considered from the perspective of the outcome and not of the process: therefore, they are merely regarded as outputs of an AI system that can have an impact on human decision-making, whilst a specific focus on the design principles of RSs and the risks posed by their biases is completely lacking” [11]. This choice may appear inconsistent with the widespread impact that algorithmic recommendations have on users, which can also include serious harms, as the case mentioned above underlines.

In the compromise text that includes the amendments voted in June 2023[5] [26], RSs are mentioned in two instances. Firstly, Recital 40b outlines how and to which extent the AIA addresses RSs, by specifying that “recommender systems are subject to this Regulation so as to ensure that” they “comply with the requirements laid down under this Regulation, including the technical requirements on data governance, technical documentation and traceability, transparency, human oversight, accuracy and robustness”. Only RSs implemented by VLOPs, and especially social media, are considered by the AIA, which complements the DSA by enabling “such very large online platforms to comply with their broader risk assessment and risk-mitigation obligations in Article 34 and 35” of that regulation. Secondly, and most importantly, the AI component of RSs becomes part of the high-risk AI applications listed in Annex III (1(8(ab))) as “AI systems intended to be used by social media platforms that have been designated as very large online platforms [. . .] in their recommender systems to recommend to the recipient of the service user-generated content available on the platform”.

While the AIA refers to the DSA for the identification of VLOPs and the enforcement of the norms concerning RSs, the fact that the AI technologies enabling automated recommendations are eventually included in this regulation testifies a welcomed change of paradigm from the previous versions. The reasons for which RSs have not been considered a high-risk AI technology in the early drafts of the AIA maybe concern the fact that recommendations impact indirectly rather than directly on individuals. A comparative example might be helpful: automated credit risk assessment, which has been included in Annex III since the beginning, is supposed to output a score that helps human decision-makers determine whether a client is suitable to receive a loan. In this case, the system is devoted to performing a content-specific task that supports human decision making (although human decisions often tend to be determined rather than supported by it). Algorithmic recommendations, instead, are not content- but context-specific: the content of their output can vary widely depending on the user, but they are directed by a defined aim within a particular context, i.e. maximizing user engagement in a social media platform.

For this reason, the recommendation does not raise ethical concerns per se, but as regards its domain of application: this may be the reason for which RSs have been initially excluded from the scope of the AIA, which regulates the risks of AI technologies per se, but included in the DSA, which instead addresses specific algorithmic systems as enablers of the services provided by online platforms. The inclusion of the AI systems enabling RSs implemented by VLOPs in Annex III underlines regulators' awareness of the risks stemming from the influence of automated recommendations. Given that the AIA has not been enforced yet, I would like to switch this analysis to another relevant regulation currently in force, i.e. the GDPR, to evaluate its potential impact on RSs transparency.

3.3 General Data Protection Regulation (GDPR) and the right to explanation

Article 22 of the GDPR [4] addresses “automated individual decision making, including profiling” stating that “the data subject shall have the right not to be subject to a decision based solely on automated processing, including profiling, which produces legal effects concerning him or her or similarly significantly affects him or her” (art. 22(1)). RSs are based on profiling, so they can be considered within the regulatory scope of this article. However, there are three exceptions to the provision reported above, which “shall not apply if the decision: (a) is necessary for entering into, or performance of, a contract between the data subject and a data controller; (b) is authorised by Union or Member State law to which the controller is subject and which also lays down suitable measures to safeguard the data subject's rights and freedoms and legitimate interests; or (c) is based on the data subject's explicit consent” (art. 22(2)). When exceptions (a) and (c) apply, “the data controller shall implement suitable measures to safeguard the data subject's rights and freedoms and legitimate interests, at least the right to obtain human intervention on the part of the controller, to express his or her point of view and to contest the decision” (art. 22(3)). Moreover, according to the fourth paragraph of the article, sensitive data should never be collected for profiling. However, it often happens that sensitive data are inferred from non-sensitive data which act as proxies: for instance, income level could be inferred from household address.

[5] The complete list of amendments can be found at:https://www.europarl.europa.eu/doceo/document/TA-9-2023-0236_EN.html

Exception (a) could be claimed in all the cases in which users are asked to accept the terms of service of a platform, which define the contract between the data subject and the data controller. Exception (c) applies when the user is asked for online consent, for example for what concerns cookies. Therefore, it can be argued that automated recommendations comply with the GDPR requirements, given that, when accepting the terms of service, the user is often giving consent to profiling and inferences. On the one side, Article 27 of the DSA aligns with the rationale of GDPR by requiring that explanations of RSs are presented in the terms and conditions, which are not often read by users and therefore may not impact on their awareness of their rights. On the other side, Article 38 of the DSA complements the GDPR by requiring that very large online platforms keep a repository of targeted advertisements, so that users can view the outcome of legitimate profiling.

[13] point out that the nudging potential of automated decision-making systems may, in some cases, lead humans to conform uncritically to their assessments, thereby making the application of Article 22 of the GDPR controversial. In fact, the safeguards against decisions that do not involve humans in the loop are not clarified in Article 22, which does not state how users can determine whether a decision is completely automated. Instead, a hint in this direction is provided by articles 13 and 14, on the right to information, and 15, on the right to access, according to which the controller must give information about the existence of automated decision-making, including profiling, as referred to in Article 22, and, at least in such cases, meaningful information about the logic used, as well as the significance and the intended consequences of such processing for the data subject [13]. This is complemented by Recital 71, which suggests that profiling "should be subject to suitable safeguards, which should include specific information to the data subject and the right to obtain human intervention, to express his or her point of view, to obtain an explanation of the decision reached after such assessment and to challenge the decision". The right to explanation envisaged here is crucial to substantiate the safeguarding claims of the cited articles, but it is not described in further detail.

This lack of precision has been criticized by [33], who identify "several reasons to doubt both the legal existence and the feasibility of such a right": in fact, "the GDPR only mandates that data subjects receive meaningful, but properly limited, information (Articles 13–15) about the logic involved, as well as the significance and the envisaged consequences of automated decision-making systems". Moreover, "the ambiguity and limited scope of the 'right not to be subject to automated decision-making' contained in Article 22 (from which the alleged 'right to explanation' stems) raises questions over the protection actually afforded to data subjects" (ibidem). The DSA goes in the direction of implementing the right to explanation outlined in the GDPR, but the effectiveness of explanations in enhancing users' autonomy is still debated. Future empirical research should be aimed at establishing whether the presence of explanations would substantially contribute to substantiate the users' rights envisioned by these regulations.

4 CONCLUSION

Automated recommendations determine not only what we see on platforms, but also our potential interest for new or different categories of content. This influencing potential can be interpreted as an instance of the "new emerging grey power" of tech companies, which "is exercised about which questions can be asked, when and where, how and by whom and hence what answers can be received in principle" [12]. A platform like TikTok, which is mainly managed through RSs, is a prominent example of this tendency: as the interface is based on an endless flow of recommended content through which the user scrolls, the contents that the user ends up seeing more frequently are related to the single videos that he watches for a longer time. This exploitative policy has already caused harm [10] because, if a video on which a vulnerable person casually spends a few seconds concerns a dangerous activity, then that individual will see the same content more and more and may eventually be influenced by it. In this sense, platforms control the questions that users pose about their interests and, subsequently, the answers that they get: in this way, digital companies end up informing a substantial part of users' online, and sometimes offline, experience. Explanations may be a countermeasure to this harmful tendency of automated recommendations, as they have the potential to make users aware of some of the questions that platforms shape for them. I argue that, in order for this potential to be realized, explanations should be integrated as a readily available, standard feature of recommendations which people may choose to review when they want to, or that appear as a pop-up on the interface of online platforms. Thanks to such a policy, users could understand why they are targeted by specific content and, subsequently, become aware of the extent to which they are influenced by RSs.

The DSA will require digital companies that use RSs and targeted advertising to build mechanisms to grant transparency, in order to enhance users' self-determination and understanding of the systems they use. However, if users are not interested in receiving explanations, or if exposure to explanations does not influence users' perspective on algorithmic recommendations, the provisions of the DSA may not have the expected results. In fact, as [32] underline, "the explanations affect a user's mental model of the recommender system, and in turn the way they interact with the explanations". The contemporary trends of RSs outline an increasing focus on explorative policies, which are likely to shape the future ways of interacting online. This may seem an evolution towards more ethical platform environments, but this is not necessarily the case. Whilst exploitative policies are considered the negative side of automated recommendations because they may lead to filter bubbles, explorative policies can also give rise to risks that should not be left untouched by ethical concerns and regulatory attention. Indeed, from the perspective of digital companies, exploration is mainly a means to get to know users even better than they currently do, by gathering data on unexplored fields of potential interests and preferences. This can lead to an even deeper nudging, which is realized through incremental exposure to contents that can provide fine-grained information on how to induce users to like what they do not know they like yet: for this reason, explorative recommendations could contribute significantly to the grey power that VLOPs already have.

I argue that an effective right to explanation is a preliminary condition for users' self-determination in the platform environment. Explanations can be considered a means to mitigate the negative consequences of the power imbalance between platforms and users. Users cannot shape automated recommendations according to their interests and needs without firstly knowing how and why they are targeted and influenced by RSs: in fact, if someone doesn't know how a system works, they are unlikely to be able to make that system work better. Digital nudging may lead to undesirable outcomes, such as manipulation, if users' perception of the recommendation process is not informed by the knowledge of how it unfolds. In this regard, my contribution points to a prominent policy problem: as explanations are the building blocks of transparency, in order to support self-determination through transparent recommendations it is firstly necessary to educate users to understand not only "what recommenders recommend" [16], but also why they recommend what they recommend. If it is not properly met by regulators on time, this sociotechnical requirement may constrain the positive ethical and societal impact of the DSA provisions. In conclusion, I think that, in order to reduce the power imbalance between platforms and users and limit the influence that the former exert on the latter, policy-makers should: 1) enforce explanations as a user-friendly tool to foster awareness that users can experience on the interface and not only read in the terms and conditions; 2) grant users the possibility of intervening directly and substantially on the strategies through which RSs target them on the platform's interface.

ACKNOWLEDGMENTS

I would like to thank my supervisor professor Andrea Simoncini for his guidance throughout the development of the research resulting in this publication. I would also like to commend the essential support that my colleague Yuhui Zhu gave me in editing the final version of the manuscript.

REFERENCES

[1] [n. d.]. European Centre for Algorithmic Transparency website. https://algorithmic-transparency.ec.europa.eu/index_en Accessed 10-05-2023.

[2] [n. d.]. *Facebook preliminary views and comments on the Digital Services Act.* https://enterprise.gov.ie/en/consultations/consultations-files/facebook-dsa-submission.pdf Accessed 20-07-2022.

[3] [n. d.]. *Proposal for a regulation of the European Parliament and of the Council laying down harmonized rules on artificial intelligence (Artificial Intelligence Act) and amending certain Union legislative acts.* https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=CELEX:52021PC0206&from=EN Accessed 11-08-2022.

[4] [n. d.]. *REGULATION (EU) 2016/679 OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation).* https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32016R0679&from=EN Accessed 20-08-2022.

[5] [n. d.]. REGULATION (EU) 2022/2065 OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL of 19 October 2022 on a Single Market For Digital Services and amending Directive 2000/31/EC (Digital Services Act). https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32022R2065 Accessed 10-05-2023.

[6] Gediminas Adomavicius and Alexander Tuzhilin. 2005. Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions. *IEEE transactions on knowledge and data engineering* 17, 6 (2005), 734–749.

[7] Engin Bozdag. 2013. Bias in algorithmic filtering and personalization. *Ethics and information technology* 15 (2013), 209–227.

[8] Engin Bozdag and Jeroen Van Den Hoven. 2015. Breaking the filter bubble: democracy and design. *Ethics and information technology* 17 (2015), 249–265.

[9] Emma Llansó Caitlin Vogus and Samir Jain. 2023. *CDT and Technologists File SCOTUS Brief Urging Court To Hold that Section 230 Applies to Recommendations of Content.* https://cdt.org/insights/cdt-and-technologists-file-scotus-brief-urging-court-to-hold-that-section-230-applies-to-recommendations-of-content/ Accessed 05-02-2023.

[10] Jonathan Edwards. 2022. *Mother sues TikTok after 10-year-old died trying 'Blackout Challenge'.* https://www.washingtonpost.com/nation/2022/05/17/tiktok-blackout-challenge-lawsuit/ Accessed 11-08-2022.

[11] Matteo Fabbri. 2023. Social influence for societal interest: a pro-ethical framework for improving human decision making through multi-stakeholder recommender systems. *AI & SOCIETY* 38, 2 (2023), 995–1002.

[12] Luciano Floridi. 2015. The new grey power. *Philosophy & Technology* 28 (2015), 329–332.

[13] Giovanni Sartor Francesca Lagioia and Andrea Simoncini. 2018. *Commento all'articolo 22 del GDPR.*

[14] Michele Gorgoglione, Umberto Panniello, and Alexander Tuzhilin. 2019. Recommendation strategies in personalization applications. *Information & Management* 56, 6 (2019), 103143.

[15] Natali Helberger, Max Van Drunen, Sanne Vrijenhoek, and Judith Möller. 2021. Regulation of news recommenders in the Digital Services Act: Empowering David against the very large online Goliath. *Internet Policy Review* 26 (2021).

[16] Dietmar Jannach, Lukas Lerche, Iman Kamehkhosh, and Michael Jugovac. 2015. What recommenders recommend: an analysis of recommendation biases and possible countermeasures. *User Modeling and User-Adapted Interaction* 25 (2015), 427–491.

[17] Mathias Jesse and Dietmar Jannach. 2021. Digital nudging with recommender systems: Survey and future directions. *Computers in Human Behavior Reports* 3 (2021), 100052.

[18] Michael Levenson and April Rubin. 2022. *Parents Sue TikTok, Saying Children Died After Viewing 'Blackout Challenge'.* https://www.nytimes.com/2022/07/06/technology/tiktok-blackout-challenge-deaths.html Accessed 15-01-2023.

[19] James McInerney, Benjamin Lacker, Samantha Hansen, Karl Higley, Hugues Bouchard, Alois Gruson, and Rishabh Mehrotra. 2018. Explore, exploit, and explain: personalizing explainable recommendations with bandits. In *Proceedings of the 12th ACM conference on recommender systems.* 31–39.

[20] Silvia Milano, Brent Mittelstadt, Sandra Wachter, and Christopher Russell. 2021. Epistemic fragmentation poses a threat to the governance of online targeting. *Nature Machine Intelligence* 3, 6 (2021), 466–472.

[21] Silvia Milano, Mariarosaria Taddeo, and Luciano Floridi. 2020. Recommender systems and their ethical challenges. *Ai & Society* 35 (2020), 957–967.

[22] Silvia Milano, Mariarosaria Taddeo, and Luciano Floridi. 2021. Ethical aspects of multi-stakeholder recommendation systems. *The information society* 37, 1 (2021), 35–45.

[23] Efrat Nechushtai and Seth C Lewis. 2019. What kind of news gatekeepers do we want machines to be? Filter bubbles, fragmentation, and the normative dimensions of algorithmic recommendations. *Computers in human behavior* 90 (2019), 298–307.

[24] Supreme Court of the United States. 2022. *21-1333 GONZALEZ V. GOOGLE LLC.* https://www.supremecourt.gov/qp/21-01333qp.pdf Accessed 05-02-2023.

[25] Supreme Court of the United States. 2023. *21-1333 REYNALDO GONZALEZ, ET AL., PETITIONERS v. GOOGLE LLC ON WRIT OF CERTIORARI TO THE UNITED STATES COURT OF APPEALS FOR THE NINTH CIRCUIT.* https://www.supremecourt.gov/opinions/22pdf/21-1333_6j7a.pdf Accessed 25-06-2023.

[26] European Parliament. 2023. *Amendments adopted by the European Parliament on 14 June 2023 on the proposal for a regulation of the European Parliament and of the Council on laying down harmonised rules on artificial intelligence (Artificial Intelligence Act) and amending certain Union legislative acts.* https://www.europarl.europa.eu/doceo/document/TA-9-2023-0236_EN.html Accessed 06-07-2023.

[27] Tate Ryan-Mosley. 2023. *How the Supreme Court ruling on Section 230 could end Reddit as we know it.* https://www.technologyreview.com/2023/02/01/1067520/supreme-court-section-230-gonzalez-reddit/ Accessed 05-02-2023.

[28] Alisha Rahaman Sarkar. 2022. *TikTok's 'blackout' challenge linked to deaths of 20 children in 18 months, report says.* https://www.independent.co.uk/tech/tiktok-blackout-challenge-deaths-b2236669.html Accessed 15-01-2023.

[29] Cass Sunstein. 2018. *# Republic: Divided democracy in the age of social media.* Princeton university press.

[30] Richard Thaler and Cass Sunstein. 2008. *Nudge: Improving Decisions about Health, Wealth, and Happiness.* Yale University Press.

[31] Nava Tintarev and Judith Masthoff. 2007. A survey of explanations in recommender systems. In *2007 IEEE 23rd international conference on data engineering workshop.* IEEE, 801–810.

[32] Nava Tintarev and Judith Masthoff. 2012. Evaluating the effectiveness of explanations for recommender systems: Methodological issues and empirical studies on the impact of personalization. *User Modeling and User-Adapted Interaction* 22 (2012), 399–439.

[33] Sandra Wachter, Brent Mittelstadt, and Luciano Floridi. 2017. Why a right to explanation of automated decision-making does not exist in the general data protection regulation. *International Data Privacy Law* 7, 2 (2017), 76–99.

[34] Shoshana Zuboff. 2019. *The Age of Surveillance Capitalism.* Profile Books.