The International Library of Ethics, Law and Technology 40

Sergio Genovesi
Katharina Kaesling
Scott Robbins *Editors*

Recommender Systems: Legal and Ethical Issues

OPEN ACCESS

Springer

The International Library of Ethics, Law and Technology

Volume 40

Technologies are developing faster and their impact is bigger than ever before. Synergies emerge between formerly independent technologies that trigger accelerated and unpredicted effects. Alongside these technological advances new ethical ideas and powerful moral ideologies have appeared which force us to consider the application of these emerging technologies. In attempting to navigate utopian and dystopian visions of the future, it becomes clear that technological progress and its moral quandaries call for new policies and legislative responses. Against this backdrop, this book series from Springer provides a forum for interdisciplinary discussion and normative analysis of emerging technologies that are likely to have a significant impact on the environment, society and/or humanity. These will include, but be no means limited to nanotechnology, neurotechnology, information technology, biotechnology, weapons and security technology, energy technology, and space-based technologies.

Sergio Genovesi • Katharina Kaesling
Scott Robbins
Editors

Recommender Systems: Legal and Ethical Issues

Springer

Editors
Sergio Genovesi
Center for Science and Thought
University of Bonn
Bonn, Germany

Scott Robbins
Centre for Science and Thought
University of Bonn
Bonn, Germany

Katharina Kaesling
Institute of International Law, Intellectual
Property and Technology Law
TUD University of Technology
Dresden, Germany

(EN): Funded by the TRA Individuals, Institutions and Societies (University of Bonn) as part of the Excellence Strategy of the federal and state governments.

(D): Gefördert von der TRA Individuen, Institutionen und Gesellschaften (Rheinische Friedrich-Wilhelms-Universität Bonn) im Rahmen der Exzellenzstrategie von Bund und Ländern.

ISSN 1875-0044 ISSN 1875-0036 (electronic)
The International Library of Ethics, Law and Technology
ISBN 978-3-031-34803-7 ISBN 978-3-031-34804-4 (eBook)
https://doi.org/10.1007/978-3-031-34804-4

This Springer imprint is published by the registered company Springer Nature Switzerland AG
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

Contents

Part I Fairness and Transparency

Part II Manipulation and Personal Autonomy

Part III Designing and Evaluating Recommender Systems

Chapter 1
Introduction: Understanding and Regulating AI-Powered Recommender Systems

Sergio Genovesi, Katharina Kaesling, and Scott Robbins

Keywords Recommender systems · AI regulation · AI ethics

When a person recommends a restaurant, movie or book, he or she is usually thanked for this recommendation. The person receiving the information will then evaluate, based on his or her knowledge about the situation, whether to follow the recommendation. With the rise of AI-powered recommender systems, however, restaurants, movies, books, and other items relevant for many aspects of life are generally recommended by an algorithm rather than a person. This volume aims to shed light on the implications of this transnational development from both legal and ethical perspectives and to spark further interdisciplinary thinking about algorithmic recommender systems.

In the last years, scientific contributions analyzed challenges deriving from the introduction of recommender systems as a tool to support our decisions in many aspects of our everyday lives, from business and education to leisure and dating (Ricci et al. 2011; Milano et al. 2020). From an ethical perspective, Milano, Taddeo and Floridi (Milano et al. 2020) identified at least six areas of concern related to the use of recommender systems, namely the spread of inappropriate content, privacy violations, threats for individual autonomy and personal identity, system opacity, fairness, and possible negative social effects.

Looking closely at the functioning of a recommender system, it is possible to exemplify many of these concerns. Let's consider an application familiar to many travelers: a system to find, sort out and recommend possible accommodations for a holiday or business trip on a community-based online platform. This kind of system not only helps travelers find a place to stay, but also helps hosts to find new clients

S. Genovesi · S. Robbins
University of Bonn, Bonn, Germany
e-mail: genovesi@uni-bonn.de; srobbins@uni-bonn.de

K. Kaesling (✉)
TUD University of Technology, Dresden, Germany
e-mail: katharina.kaesling@tu-dresden.de

S. Genovesi et al. (eds.), *Recommender Systems: Legal and Ethical Issues*, The International Library of Ethics, Law and Technology 40,
https://doi.org/10.1007/978-3-031-34804-4_1

and, of course, is essential in order for the online platform to work. This means that, as already highlighted by Milano, Taddeo and Floridi (Milano et al. 2021), the study of a recommender system requires a multi-stakeholder analysis to understand the interests and needs of all the actors involved in its functioning and to highlight possible ways in which the system may represent a risk for one or more stakeholder groups, as well as for society at large.

Starting with fairness concerns, if the system performs differently across demographic groups based on users' personal attributes such as gender, spoken language, or nationality, a concrete discrimination risk exists both for consumers looking for a place to stay and for hosts providing accommodation (Burke 2017; Solans et al. 2021) – for example, certain users might not receive recommendations for certain accommodations based on wrong or biased automated predictions. Prescinding from the output's distribution, other fairness concerns may arise as well, among other things, concerning the inclusiveness of user experience, the application environment of the system (Grgić-Hlača et al. 2018), and the working conditions of people involved in the training, development, and supply processes of the system (Fuchs and Fisher 2015; Gray and Suri 2019).

Another much discussed ethical concern is the extent of the influence – if not even interference – with individual decision processes. In our example, the recommender system filters just a few options out of many based on the predicted relevance for the user. Other options are not hidden from the user, but are less visible – for example, they are not shown among the first results. So, even though users can access information about the available accommodation, in practice they are more likely to consider just first ones. The tendency of users to interact with items on top of a list with higher probability than with items at a lower position in the list, regardless of the items' actual relevance is called "position bias" and affects users of recommender systems (Collins et al. 2018). This is problematic for at least two reasons. On the one hand, if automated predictions are inaccurate or manipulated in order to promote other businesses' goals, users are being pushed toward options that don't reflect their interests and are therefore being distracted from their original intentions. This has led scholars to reflect about the manipulation threats related to digital technologies and AI systems in particular (Jongepier and Klenk 2022; Susser et al. 2019). On the other hand, irrespective of the predictions' accuracy, while interacting with the system users form their opinion based only on a selection of possible relevant items and might miss important pieces of information. This has led scholars to problematize whether pieces of information predicted to be not relevant should be included in the first result shown as well, a practice that has been called "serendipity by design" (Reviglio 2017). Moreover, in-built nudging techniques might influence individual decisions in a morally problematic way, for example causing users to make hasty decisions by pushing them to hurry up booking to not lose the accommodation.

Privacy issues of recommender systems are strictly related to fairness and individual autonomy concerns. Indeed, in order for recommender systems to predict what kind of content is relevant for a certain user, and therefore influence their decisions in a meaningful way, it is necessary to access, collect and process their data. Personal data is provided by users as part of agreements to use digital services, such as online platforms, and user behavior, such as history and navigation data, is tracked and amalgamated, often without proper consent or even awareness and beyond the purpose of providing the digital services sought by the user. The use of behavioral data as raw material for "prediction products" anticipating future behavior has aptly been described as surveillance capitalism (Zuboff 2019), causing power aggregation in the hands of big tech companies (Véliz 2020). These are often giving consumers no choice but to consent to collection and processing of personal data if they want to use digital services. Moreover, access to user data by untrusted parties or inappropriate use of this data can represent a serious threat to user privacy (Friedman et al. 2015).

Legal and ethical questions with regard to (meaningful) user consent and businesses' use of dark patterns have been discussed with regard to the EU General Data Protection Regulation (GDPR), which entered into force on 24 May 2016 and applies since 25 May 2018. With the rise of the platform economy (Acs et al. 2021), privacy concerns beyond subjective data rights, moved into the focus of attention, e.g., the impact of technology (including recommender systems) on decisional and intellectual privacy (Richards 2017). Big Data and artificial intelligence pose new challenges to the traditional understanding of privacy and data protection, prompting discussions on predictive privacy (Mühlhoff 2021). The EU Digital Services Act (DSA) of October 2022 addresses structural issues beyond subjective user rights in a number of novel ways (Kaesling 2023b). Inter alia, it contains special rules on recommender systems on online platforms and additional obligations of very large online platforms and very large online search engines with relation to their use of recommender systems (Article 27 DSA, Article 38 DSA). The role of recommender systems for systemic risks flowing from the design, functioning and use of their digital services is also addressed (Articles 34 and 35). The Digital Services Act specifically addresses the impact of recommender systems on the ability of recipients to retrieve and interact with information online, including to facilitate the search of relevant information for recipients of the service and contribute to an improved user experience, their role in the amplification of certain messages, the viral dissemination of information, and the stimulation of online behavior (Recital 70 DSA). The interpretation and impact of these new rules has yet to be determined (Janal 2021). This volume contains some of the first studies of these regulations and their relation to other regulatory approaches, while juxtaposing perspectives from legal and ethical studies with the same points of reference.

System opacity, meaning the lack of transparency in the decisional process and poor explainability of decisional outcome, is a problem that affects many AI-powered systems, including recommender systems. In our example, a possible case for opacity could be represented by the output of recommendations that are not explainable based on the search parameters. Moreover, requesting to input information that do not intuitively contribute to refine the search for accommodation without explaining how the system processes this information would be an example of a transparency and data protection issue at the same time. Opacity is an ethical issue because unexplainable decisions cannot be understood and therefore objected to by users and developers. This undermines user control over the system and human agency in general. In contrast, transparent systems and explainable decisions empower users by allowing them to contest decisions they perceive as wrong or unfair. In the field of computer science, it is much debated to what extent automated decision based, for instance, on machine learning (ML) could and should be made explainable and what analytic methods should be used to explain whole models or single decisions (Molnar 2022). The Digital Services Act demands recommender system transparency from online platforms, specifically with regard to the main parameters and options for users to modify or influence those main parameters (Article 27 DSA), i.e., linking transparency to users' choice, as will be analyzed in this volume.

Concerns regarding inappropriate content are common when considering recommender systems since misclassification of offensive and potentially harmful items has often occurred in the past. An infamous occurrence was the recommendation of disturbing videos portraying grotesque imitation of famous cartoon characters to children on YouTube Kids (Papadamou et al. 2020). In that case, the classifier failed at sorting out disturbing videos uploaded by trolls and labeled them as child-friendly content, causing psychological distress to many children. In our example, inappropriate content could be represented by fake accommodation posted by scammers, or by offensive content such as explicit text or images disguised as accommodation description or user profile.

Finally, concerning the potential negative impact on society of recommender systems much attention was raised in the last decade due to scandals involving the spread of disinformation and threats for democratic processes. This is addressed in the Digital Services Act as part of the systemic risks and their mitigation (Article 34 and 35 DSA) (Peukert 2021; Kaesling 2023a). Cambridge Analytica, arguably one of the most discussed cases, directly involved the use of recommender systems since content meant to influence voters was shown as recommended content on their social media news feed. Possible negative impact on society is not limited to disinformation. Considering our example one last time, we should question the impact that such an application could have on rental market in a city – for example, whether it will lead to price inflation or to scarcity of apartments available for long-term lease, and how this will impact the lives of inhabitants who cannot find affordable places anymore. Concluding this list of concerns on a positive note, investigating the impact of recommender systems on society also includes finding ways to employ them for social good, promoting sustainable development goals, individual flourishing and harm prevention (Hermann 2022; Taddeo and Floridi 2018).

These general issues regarding recommender systems also mirror the ethical concerns expressed by the High-Level Expert Group (HLEG) on AI appointed by the European Commission. Indeed, in formulating the key requirements AI systems should meet in order to be trustworthy, the HLEG pointed out seven main risk areas AI audit should focus on: human agency and oversight, safety, privacy, transparency, fairness, societal and environmental wellbeing, and accountability (HLEG on AI 2019).

Within the discipline of legal studies alone, several legal areas are touched in the context of algorithmic recommender systems. These areas include discrimination law, data protection law, unfair competition law, existing sector-specific platform regulation, such as P2B Regulation (Busch 2019), and contract law and its general principles, such as the private autonomy of contracting parties. In addition to the far-reaching regulation of recommender systems on online platforms in the Digital Service Acts, the proposed EU framework for Artificial Intelligence (Proposal for a Regulation laying down harmonized rules on artificial intelligence and amending certain Union legislative Acts of 21 April 2021) will have an impact upon its adoption, which is – in part – already anticipated in this volume.

The Digital Services Act contains a workable legal definition of recommender systems. According to Art. 3 lit. s. Digital Services Act, 'recommender system' means 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. The understanding, implementation and further development of the Digital Services Act's regulatory approaches in the context of recommender systems depends on interdisciplinary exchanges, which this volume aims to start and foster. Assessing the role of recommender systems for systemic risks within the meaning of Article 34 DSA, for example, presupposes the development of a system definition (Kaesling 2023a), which can be informed by the human-centric approach of the High-Level Expert Group on AI appointed by the European Commission, and specifically the seven above mentioned key requirements for trustworthy AI (HLEG on AI 2019). The new legal framework of the Digital Services Act gives ample space to build on interdisciplinary insights, as they can be found in this volume.

Contributions in this volume offer analyses from different perspectives and aim to enrich both the ethical debate and the discussion on the interpretation of new legal norms and their future developments. Legal and ethical issues of recommender systems will be addressed in three thematic clusters: Fairness and Transparency, Manipulation and Personal Autonomy, and Design and Evaluation of Recommender Systems.

The first section entitled "Fairness and Transparency" addresses legal and ethical issues related to discrimination and unfair treatment of individuals as an effect of the development and application of recommender systems, as well as further concerns related to the lack of transparency of the decisional processes behind automated recommendations and its moral and legal implications for users. *Susanne Gössl*, in her paper "Recommender Systems and Discrimination" deals with a much-debated topic from a legal point of view. *Gössl* not only examines data protection law, unfair competition law and general anti-discrimination law, but also exposes lacunes in that regulation and evaluates the potential of emerging regulation to close regulatory gaps, notably the information approach, which is centered around the best possible information about the parameters and the risks of a specific recommender system. This aspect is then continued in *Christoph Busch*'s contribution "Platform Regulation and Recommender Systems – From Algorithmic Transparency to Algorithmic Choice", in which he describes a paradigm shift and its consequences for the regulation of recommender systems on online platforms. *Gesmann-Nuissl and Meyer* analyze the specific lack of transparency on gaming platforms. In their contribution entitled "Black Hole instead of Black Box? – The Double Opaqueness of Recommender Systems on Gaming Platforms and its Legal Implications", they find that the mixing of different components, namely shopping, streaming and social media, leads to an exacerbation of the black-box problem. With a view to the Digital Services Act and the proposed Artificial Intelligence Act, they develop solutions fostering transparency regarding the platform users, platform operators, and software developers as stakeholders. *Sergio Genovesi* complements these viewpoints by considering the position of digital laborers in the value production and redistribution processes for recommender systems in his paper "Digital Labor as a Structural Fairness Issue in Recommender Systems".

The second section on "Manipulation and Personal Autonomy" focusses on the recommender system's influence on the formation of the human will and values. Drawing on both legal and philosophical backgrounds, *Karina Grisse* explores risks of manipulation by recommender systems and how EU law can mitigate them in her chapter "Recommender Systems, Manipulation and Private Autonomy – How European civil law regulates and should regulate recommender systems for the benefit of private autonomy". *Marius Bartmann* then argues, from an ethical point of view, that the identification of the recommendation rationale is vital for preserving autonomous human decision-making in his chapter entitled "Reasoning with Recommender Systems? Practical Reasoning, Digital Nudging, and Autonomy". *Scott Robbins,* in this section's last chapter entitled "Recommending Ourselves to Death: values in the age of algorithms" argues that recommendations are likely to be off track due to distorting forces that are inherent to evaluative recommendations. He goes further and argues that these incorrect recommendations will feedback into our own evaluative standards – wresting control over the evaluative from humans. He makes the case that this is a fundamental loss of meaningful human control.

The Section "Designing and Evaluating Recommender Systems" focusses on the practical implementation of general legal and ethical principles. In order to better understand and effectively address the risks associated with the use of recommender systems, the lack of transparency and the potential for manipulation, the design and constant (re-)evaluation of recommender systems in their specific context is paramount. With regard to the specific use case of a food recommender app, in their contribution "Ethical and Legal Analysis of Machine Learning Based Systems: A Scenario Analysis of a Food Recommender System" *Olga Levina* and *Saskia Mattern* exemplify how a combined ethical and legal assessment should be performed, highlighting the benefits of integrating such an assessment in the design process. In the chapter "Factors Influencing Trust and Use of Recommendation AI: A Case Study of Diet Improvement AI in Japan", *Arisa Ema* and *Takashi Suyama* present a survey conducted in Japan investigating users' trust in a recommender system for dietary habit improvement. The survey questions the impact that the usage of AI technologies, data management standards and purposes of use have on users' trust. *Lisa Roux* and *Thierry Nodenot*, in the last chapter entitled "Ethics of E-learning Recommender Systems: Epistemic Positioning and Ideological Orientation" investigate the ethical and practical implications of recommender systems' design in e-learning and show how system design can reflect ideological conceptions of science and techniques and specific visions of teaching and learning.

This volume documents some of the ideas developed in the framework of the editors' project "Recommender Systems: Legal and Ethical Issues", which was funded by the University of Bonn's Transdisciplinary Research Area 4 (TRA 4) "Individuals, Institutions and Societies", set up as part of the University of Bonn's excellency initiative. The editors would like to thank project assistants Luis Nussbauer and Marie Bente John for their valuable support as well as all contributors to the hybrid conference on the topic organized in December 2021 as a preliminary step to the publication, namely Marius Bartmann, Joanna Bryson, Christoph Busch, Vicky Charisi, Dagmar Gesmann-Nuissl, Susanne Gössl, Karina Grisse, Olga Levina, Stefanie Meyer, Silvia Milano, Julia Maria Mönig, Lisa Roux, Shannon Vallor and Aimee van Whynsberghe. The interdisciplinary academic discussions at the conference inspired and shaped many of the contributions in this volume.

References

Acs, Zoltan J., Abraham K. Song, László Szerb, David B. Audretsch, and Éva Komlósi. 2021. The Evolution of the Global Digital Platform Economy: 1971–2021. *Small Business Economics* 57 (4): 1629–1659. https://doi.org/10.1007/s11187-021-00561-x.

Burke, Robin. 2017. Multisided fairness for recommendation. https://arxiv.org/pdf/1707.00093.

Busch, Christoph. 2019. Mehr Fairness und Transparenz in der Plattformökonomie? Die neue P2B-Verordnung im Überblick. *Zeitschrift der Deutschen Vereinigung für gewerblichen Rechtsschutz und Urheberrecht* 8: 788–796.

Collins, Andrew, Dominika Tkaczyk, Akiko Aizawa, and Joeran Beel. 2018. A study of position bias in digital library recommender systems. https://arxiv.org/pdf/1802.06565.

Friedman, Arik, Bart P. Knijnenburg, Kris Vanhecke, Luc Martens, and Shlomo Berkovsky. 2015. Privacy Aspects of Recommender Systems. In *Recommender Systems Handbook*, 649–688. Boston, MA: Springer.

Fuchs, Christian, and Eran Fisher, eds. 2015. *Reconsidering Value and Labour in the Digital Age*, Dynamics of Virtual Work. 1st ed. London: Palgrave Macmillan UK. Imprint: Palgrave Macmillan.

Gray, Mary, and Siddharth Suri. 2019. *Ghost Work. How Amazon, Google, and Uber Are Creating a New Global Underclass*. 1st ed. Boston: Houghton Mifflin Harcourt Publishing.

Grgić-Hlača, Nina, Muhammad Bilal Zafar, Krishna P. Gummadi, and Adrian Weller. 2018. Beyond Distributive Fairness in Algorithmic Decision Making: Feature Selection for Procedurally Fair Learning. In Proceedings of the AAAI Conference on Artificial Intelligence, vol. 32, issue 1. https://doi.org/10.1609/aaai.v32i1.11296.

Hermann, Erik. 2022. Artificial Intelligence And Mass Personalization of Communication Content – An Ethical and Literacy Perspective. *New Media & Society* 24 (5): 1258–1277. https://doi.org/10.1177/14614448211022702.

HLEG on AI. 2019. Ethics guidelines for trustworthy AI. https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai

Janal, Ruth. 2021. Haftung und Verantwortung im Entwurf des Digital Services Acts. *Zeitschrift für Europäisches Privatrecht* 29: 227–275.

Jongepier, Fleur, and Michael Klenk (eds.). 2022. *The Philosophy of Online Manipulation*. Routledge Research in Applied Ethics. Erscheinungsort nicht ermittelbar. Routledge.

Kaesling, Katharina. 2023a. Commentary on Article 34 DSA. In *Digital Services Act: DSA: Gesetz über digitale Dienste*, ed. Franz Hofmann and Benjamin Raue, 1st ed. Baden-Baden: Nomos/Manz Verlag Wien/Helbing & Lichtenhahn.

———. 2023b. Preliminary Remarks on Article 33 ff (Additional Obligations for Very Large Online Platforms and Very Large Online Search Engines). In *Digital Services Act: DSA: Gesetz über digitale Dienste*, ed. Franz Hofmann and Benjamin Raue, 1st ed. Baden-Baden: Nomos/Manz Verlag Wien/Helbing & Lichtenhahn.

Milano, Silvia, Mariarosaria Taddeo, and Luciano Floridi. 2020. Recommender Systems and Their Ethical Challenges. *AI & SOCIETY* 35 (4): 957–967. https://doi.org/10.1007/s00146-020-00950-y.

———. 2021. Ethical Aspects of Multi-stakeholder Recommendation Systems. *The Information Society* 37 (1): 35–45. https://doi.org/10.1080/01972243.2020.1832636.

Molnar, Christoph. 2022. *Interpretable Machine Learning. A Guide for Making Black Box Models Explainable*. Munich: Christoph Molnar.

Mühlhoff, Rainer. 2021. Predictive Privacy: Towards an Applied Ethics of Data Analytics. *Ethics and Information Technology* 23 (4): 675–690. https://doi.org/10.1007/s10676-021-09606-x.

Papadamou, Kostantinos, Antonis Papasavva, Savvas Zannettou, Jeremy Blackburn, Nicolas Kourtellis, Ilias Leontiadis, Gianluca Stringhini, and Michael Sirivianos. 2020. Disturbed YouTube for Kids: Characterizing and Detecting Inappropriate Videos Targeting Young Children. In Proceedings of the International AAAI Conference on Web and Social Media, vol. 14, 522–533. https://doi.org/10.1609/icwsm.v14i1.7320.

Peukert, Alexander. 2021. Five reasons to be sceptical about the DSA. *Verfassungsblog*. https://doi.org/10.17176/20210831-233126-0.

Reviglio, Urbano. 2017. Serendipity by Design? How to Turn from Diversity Exposure to Diversity Experience to Face Filter Bubbles in Social Media. In *International Conference on Internet Science*, 281–300. Cham: Springer. https://doi.org/10.1007/978-3-319-70284-1_22.

Ricci, Francesco, Lior Rokach, Bracha Shapira, and Paul B. Kantor, eds. 2011. *Recommender Systems Handbook*. Boston, MA: Scholars Portal.

Richards, Neil. 2017. *Intellectual Privacy. Rethinking Civil Liberties in the Digital Age*. New York: Oxford University Press.

Solans, David, Francesco Fabbri, Caterina Calsamiglia, Carlos Castillo, and Francesco Bonchi. 2021. Comparing Equity and Effectiveness of Different Algorithms in an Application for the Room Rental Market. In Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society, 978–988. AIES'21: AAAI/ACM Conference on AI, Ethics, and Society, Virtual Event USA. 19.05.2021, 21.05.2021. New York: Association for Computing Machinery. https://doi.org/10.1145/3461702.3462600.

Susser, Daniel, Beate Roessler, and Helen Nissenbaum. 2019. Technology, autonomy, and manipulation. *Internet Policy Review* 8 (2).

Taddeo, Mariarosaria, and Luciano Floridi. 2018. How AI can be a Force for Good. Science (New York, N.Y.) 361 (6404): 751–752. https://doi.org/10.1126/science.aat5991.

Véliz, Carissa. 2020. *Privacy is power. Why and How you Should Take Back Control of your Data/Carissa Véliz*. London: Transworld.

Zuboff, Shoshana. 2019. *The Age of Surveillance Capitalism. The Fight for a Human Future at the New Frontier of Power*. New York: PublicAffairs.

Part I
Fairness and Transparency

Chapter 2
Recommender Systems and Discrimination

Susanne Lilian Gössl

Abstract The following article deals with the topic of discrimination "by" a recommender system. Several reasons can create discriminating recommendations, especially the lack of diversity in training data, bias in training data or errors in the underlying modelling algorithm. The legal frame is still not sufficient to nudge developers or users to effectively avoid those discriminations, especially data protection law as enshrined in the EU General Data Protection Regulation (GDPR) is not feasible to fight discrimination. The same applies for the EU Unfair Competition Law, that at least contains first considerations to allow an autonomous decision of the subjects involved to know about possible forms of discrimination. Furthermore, with the Digital Service Act (DSA) and the AI Act (AIA) there are first steps into a direction that can inter alia tackle the problem. Most effectively seems a combination of regular monitoring and audit obligations and the development of an information model, supported by information by legal design, that allows an autonomous decision of all individuals using a recommender system.

Keywords Algorithmic discrimination · Fairness · Digital Services Act · AI Act proposal

The following article deals with the topic of discrimination "by" a recommender system that is based on incomplete or biased data or algorithms. After a short introduction (I.), I will describe the main reasons why discrimination by such a recommender system can happen (II.). Afterwards I will describe the current legal frame (III.) and conclude on how the future legal frame could look like and how the legal situation might be further improved (IV.).

S. L. Gössl (✉)
University of Bonn, Bonn, Germany
e-mail: sgoessl@uni-bonn.de

S. Genovesi et al. (eds.), *Recommender Systems: Legal and Ethical Issues*, The International Library of Ethics, Law and Technology 40,
https://doi.org/10.1007/978-3-031-34804-4_2

2.1 Introduction

A recommender system gives recommendations based on an algorithm, often a machine learning algorithm (Projektgruppe Wirtschaft, Arbeit, Green IT 2013, 20). A machine learning algorithm basically works in the way that it takes a set of data and tries to find correlations between different data sets. If it finds enough correlations, it might derive a rule from those correlations. Based on the rule, the algorithm then makes a prediction about how a similar input might be handled in the future. Based on the prediction, the recommendation is made. For example, a machine learning algorithm that is supposed to classify cats, is trained on a certain number of pictures of cats and other animals. The algorithm then finds a correlation regarding the shape and size of ears, the tail, and whiskers. When novel pictures are used as inputs it checks for these features to conclude whether the picture shows a cat or not.

All these steps, be it the data set or data gathering, the finding of the correlations or, consequently, of the rules and predictions, can contain biases. As a result, the recommendation can contain those biases as well, which might lead to a discriminatory recommendation, e.g. that a machine gives a recommendation that is more favorable towards men than women or towards persons from a privileged social background than persons from another background (Alpaydın 2016, 16 ff.; Gerards and Xenidis 2021, 32 ff.; Kelleher 2019, 7 ff.; Kim and Routledge 2022, 75–102, 77 ff.; Vöneky 2020, 9–22, 21).

While some recommendations, e.g. the ranking of proposed items on a shopping website, (Wachter 2020, 367–430, 369 ff.) can be of a lesser fundamental rights relevance (Speicher et al. 2018), some recommender systems can be extremely relevant for the well-being of a person. E.g. a website can match employers and employees. If the website does not propose a possible employee for a job even though s/he would have been well-suited (Lambrecht and Tucker 2020, 2966–2981), that is not only a question of a bad-functioning algorithm but can touch the professional existence of the person left out (see, e.g., recital 71 of the Data Protection Directive – General Data Protection Regulation (GDPR 2016, 1)). Similarly, rankings of professionals (doctors, lawyers etc.) for somebody looking for the relevant service, are highly important as only the first few candidates have a chance to be chosen.[1]

2.2 Reasons for Discriminating Recommendations

There are several reasons why a recommendation can be discriminating. They can basically be distinguished into three categories: The data set from which the machine learning algorithm is trained and adjusted can lack the relevant diversity (1), the training data can contain conscious or unconscious bias of the people creating the

[1] See e.g. with a focus on scoring (Gerberding and Wagner 2019, 116–119, 188).

data (2) and, finally, the underlying algorithm can be modelled in a way that it enhances discriminations (3) (von Ungern-Sternberg forthcoming).

2.2.1 Lack of Diversity in Training Data

The level of diversity in training data is paramount for the outcome of the concrete recommendation. One famous example where the lack of diversity lead to discrimination of women, was the Amazon hiring tool (Gershgorn 2018): The hiring tool was supposed to make objective predictions of the quality and suitability of applying job candidates. Problematic was that the algorithm was "fed" by application data of the last decade – which included a significant higher proportion of male (and probably white) candidates. The training data, therefore, lacked diversity regarding women. As a consequence, the hiring tool "concluded" that women were less qualified for the job, resulting in discriminatory recommendations (Gershgorn 2018). Similarly, whenever training data is only taken from reality and not created artificially, there is a high probability that it will lack diversity – especially in jobs that have typically a higher number of men (as STEM areas - Science, Technology, Engineering, and Mathematics (Wikipedia 2022)) or women (as care and social work) or lack – so far – People of Colour (PoC) or candidates with an immigration, LGBTIAQ* or disability background, as in these jobs the representation of these groups might be extraordinarily higher or lower than the one of other groups (Reiners 2021; Sheltzer and Smith 2014, 10107–10112). The effect of missing diversity in training data was also shown in face recognition software using machine learning algorithms: Face recognition software that was trained mainly with photos from white and male people, afterwards had stronger problems to identify black or female and especially female black persons (Buolamwini and Gebru 2018).

2.2.2 (Unconscious) Bias in Training Data

The second, very influential reason why recommender systems often show discriminating results is the fact that the training data very often contains data from real life people and therefore, also reflects their conscious or unconscious bias. For example, there has been a study of the University of Bonn regarding "Gender Differences in Financial Advice" (Bucher-Koenen et al. 2021; Cavalluzzo and Cavalluzzo 1998, 771–792) analysing the recommendations financial advisors gave to different people looking for advice. The study shows that usually the recommendations women receive are more expensive than those of male candidates. There are several explanations, e.g. the fact that women very often are more risk adverse, resulting in more expensive but also safer investments. Another possible reason was that men often look for advice to get a "second opinion", while women do not consult other advisors and lack the information male candidates may already have (Bucher-Koenen et al. 2021, 12 ff., 14 ff.). An algorithm that "learns" from this data might conclude that women always should get the more expensive recommendations without looking at the concrete woman applying. The whole problem can be enhanced by data labelling practises. Training data usually gets labelled as "correct" or "incorrect" (or "good" or "bad") to enable the learning process of the algorithm. Whenever the decision whether a résumé or a person's performance is "good" is not only based on the hiring decision, but furthermore, a separate person labels it as "good" or "bad", the labelling decision can contain an additional (unconscious) bias of the labelling person (Calders and Žliobaitė 2013, 48 ff.). For example, there are algorithms that recommend professionals or professional services, often based on users' recommendations. Very often a ranking is made with those receiving the highest recommendation coming first (thus having the label "good").[2] This can discriminate e.g. women or members of minority groups: There is research that people rate members of these groups or women typically less favourable than a man not belonging to a minority group even though the performance is the same. Research shows, e.g., that equal résumés with a male or female name on it are evaluated differently, usually the female one less favourable (by male and female evaluators equally) (Moss-Racusin et al. 2012, 16474–16479; Handley et al. 2015, 13201–13206). The same applies to teaching materials in law schools (Özgümüs et al. 2020, 1074).

If a machine learning algorithm, thus, ranks the recommendation of professionals or professional services based on these user evaluations, the probability is high that these (unconscious) biases that led to a less favorable rating in the first place, will also lead to a lower ranking in the recommendation – with negative influence e.g., on the income and career of the professional. For instance, a study looking at the company Uber that based the ranking of its drivers on consumers' ratings, shows these biases clearly (Rosenblat et al. 2017, 256–279).

2.2.3 Modelling Algorithm

Finally, the algorithm can be modelled in a way that it enhances biases already contained in the training data. One reason can of course be the selection of the relevant features the algorithm uses for the selection – e.g., if a personalized ad algorithm filters ads only according to the gender of the user, the result might be that women always receive recommendations for sexy dresses and make-up while men might always receive recommendations for adventure trips, barbeque and home building tools (Ali et al. 2019, 1–30).

Flaws in the modelling can also have a tremendous impact depending on the area where they are used. Another example of discriminating results with regard to the programming of the algorithm could be found in the famous COMPAS program

[2] E.g. the page https://www.jameda.de/ for physicians in Germany.

used by several US-States (Angwin et al. 2016; Flores et al. 2016, 38–46; Martini 2019, 2 ff.). This program aimed at recommendations regarding the re-offense probability of criminal offenders. A high risk for future crime would lead to a less favorable treatment in detention – e.g. a higher bail or an exclusion of the possibility to be bailed out. The program reflected the unconscious bias the judges had towards Afro-American candidates, assuming their re-offense risk would be higher and towards Caucasian offenders, assuming their re-offense risk would be lower. These two flaws from the real world could have been at least mitigated by a calibration of the algorithm allocating different error rates to different groups within the training data, making the algorithm "learn" to avoid the same bias. This is important whenever different groups have different base rates, i.e., rates of positive or negative outcomes.

Therefore, one problem in the modelling algorithm was that the allocation of error rates analyzing the existing data was equal towards both groups, even though it should have included the fact that a Caucasian person had two biases in his/her favor (no assumption of higher re-offence risk and assumption of lower re-offence risk) while the Afro-American person had only one against him (the assumption of a higher re-offence risk), thus, different base rates. The probability that the outcome regarding of an Afro-American person would be a false (and negative) prediction, therefore, was higher. This should have been reflected in the error rate.

Therefore, an equal allocation of error rates even enhanced the biases already contained in the training data (Chouldechova 2017, 153–163; Rahman 2020; Barocas et al. 2018, 23, 31, 68). Problematic, on the other hand, is that different error rates assume that there are differences in groups, thus making a distinction even though a distinction was supposed to be avoided (Barocas et al. 2018, 47 ff.).

2.2.4 Interim Conclusion and Thoughts

Recommender systems can discriminate as they can reinforce and deepen stereotypes and biases already found in our society. Several problems can lead or enhance those outcomes: First, the data used from experience, is always from the past, thus reflecting biases and difficulties from the past. Thus, the person selecting the training data must always have in mind that the past has no perfect data to reflect the diversity of our society. So-called "counterfactuals" that have to be created artificially can help to avoid the lack of diversity (Cofone 2019, 1389–1443; Mothilal et al. 2020; Oosterhuis and de Rijke 2020).[3] Counterfactuals refer to artificially created data sets that can counterbalance the aforementioned lack of diversity in data stemming from reality – e.g., the aforementioned lack of female résumés in the STEM areas can be counterbalanced by introducing artificially created female résumés.[4]

[3] See also the solution proposed by (Blass 2019, 415–468).

[4] Regarding the use of counterfactuals and its risks e.g. see (Kasirzadeh and Smart 2021, 228 ff.).

Second, while it is easy to avoid differentiation features that are obviously discriminatory, such as “race” or “gender”, the compilation of data can have similar effects as such direct discriminating features (Ali et al. 2019, 1–30; Buolamwini and Gebru 2018, 12). For example, the postal code of a person in many countries is highly correlated with ethnicity or social background, thus, if an algorithm “learns” that résumés from a certain area are usually “bad”, this indirectly leads to discrimination based on the social or ethical background (Calders and Žliobaitė 2013, 4–49).

Third, because of these effects caused by certain data compilations that are difficult to predict ex ante, especially if the algorithm is self-learning, it is also difficult to predict under which circumstances discrimination will be caused by which reason. This unpredictability makes it necessary to monitor and adjust such algorithms on a regular basis.

2.3 Legal Frame

So far, there is no coherent legal frame to tackle discrimination by recommender systems. Nevertheless, certain approaches can be derived from the existing legal frame: Existing solutions are either based on agreement (1.), information (2.), or a combination of both approaches. Finally, the general rules of anti-discrimination law apply (3.).

2.3.1 Agreement – Data Protection Law

The first approach that is based on user agreement can be found in data protection law, especially Article 22 para. 1 GDPR (2016, 1). According to that rule, “[t]he 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.”

Recital 71 gives more specifications on the Article and makes clear that the controller of the algorithm should “prevent, inter alia, discriminatory effects on natural persons on the basis of racial or ethnic origin, political opinion, religion or beliefs, trade union membership, genetic or health status or sexual orientation, or processing that results in measures having such an effect.”

While this rule at first glance sounds like a clear prohibition to create recommendations based on exclusively algorithmic decisions, there are several problems in the application of the rule that make it questionable whether it is sufficient to resolve the problem. First, one can question in general whether data protection law is the proper venue to prevent discriminatory results. Data protection law primarily is intended to protect the personal data of natural persons and to give them control on how this data is used. It aims at the protection of the personality rights of such a person. Anti-discrimination law, on the other hand, tackles certain inequalities that exist in society, and protects the individual from discrimination – independently of the data used or concerned. While, of course, discrimination can also lead to the infringement of a personality right, the protective function is a different one.

Furthermore, literature disagrees under which circumstances there is a "decision" in the sense of Article 22 GDPR regarding recommender systems. While recital 71 clarifies that such a "decision" is the case when we have a refusal, e.g., "of an online credit application or a recruiting practice without any human intervention", the case becomes less clear when the algorithm only proposes a certain job opportunity (or not) (Lambrecht and Tucker 2020, 2966–2981) or ranking that afterwards will be subjected to the decision of a person. While some voices regard such a preliminary recommendation as excluded from the cope of Article 22 DGPR (German government 2000, 37; EU Commission 2012, 26 et seq.; Martini 2019, 173; see also OLG Frankfurt/M. 2015, 137), others limit the notion of decision to the exclusion of a person (from e.g. a ranking).[5]

Nevertheless, even if we apply Article 22 to all recommendations, a justification is possible if the controller uses the algorithm, *inter alia*, "to ensure the security and reliability of a service provided by the controller, or necessary for the entering or performance of a contract between the data subject and a controller, or when the data subject has given his or her explicit consent" (Recital 71, also Article 22 para. 2). Para. 3 then introduces some procedural safeguards for the protection of the personality rights of the person concerned. Nevertheless, the basic rule is that whenever the data subject has given the explicit consent for the processing of the data, the infringement within the meaning of Article 22 para. 1 is justified under Article 22 para. 2 GDPR (Vöneky 2020, 9–22, 13; Martini 2019, 171 ff.). This is problematic as research shows that the majority of internet users are willing to give their consent to proceed on a website without really dealing with the content of the agreement (Carolan 2016, 462–473; in detail see also Machuletz and Böhme 2020, 481–498). If an agreement is easily given without a conscious choice, Article 22 GDPR does not provide a very stable protection against discriminatory results.

2.3.2 *Information – Unfair Competition Law*

The second approach can be called an information-centered approach. The main measure consists in giving information to the user about the available ranking parameters and the reasons for the relative importance of certain parameters to others. We can see that approach on the Business-to-Business (B2B) level in Article 5 P2B Regulation (Regulation (EU) 2019/1150. 2019, 57 ff.) regarding online providers and businesses using their platforms. A similar rule has also been introduced into the UCP Directive (Directive 2005/29/EC 2005, 22) regarding the

[5] E.g. (von Lewinski 2021, para. 16, unclear at 16.1). To the whole discussion see von Ungern-Sternberg, Discriminatory AI and the Law: Legal Standards for Algorithmic Profiling. In *Responsible AI*, ed. Silja Vöneky et al., forthcoming. II. 2. b).

Business-to-Consumer (B2C) level in its 2019 amendment (Article 3 Nr. 4 lit. b) (Directive (EU) 2019/2161 2019, 7 ff.). Article 7 para. 4a of the UCP Directive provides that whenever a consumer can search for products offered by different traders or by general consumers "information […] on the main parameters determining the ranking of products presented to the consumer as a result of the search query and the relative importance of those parameters, as opposed to other parameters, shall be regarded as material," meaning that this information has to be part of the general information obligations towards the consumer. The effectiveness of these measures to combat discriminatory recommendations is doubtful.

First, both rules only contain information obligations, meaning that the effectiveness mainly depends on the attention of the user and his or her willingness to read the information, understand what the "relative importance of certain parameters" means for his or her concrete use of the platform and act upon that knowledge. Even if a trader or intermediary indirectly gives the information that the recommendation can be discriminatory, in most cases the platform or search possibility will most probably still be used as the majority of users will not notice it (Martini 2019, 188; Bergram et al. 2020). Furthermore, the information necessary to understand the logic of a discriminatory recommender system might not be part of the information that is part of the information obligation. The limit will most probably lie behind the protection of trade secrets of the provider of the algorithm – including the algorithm or at least some of its features. So, discrimination caused by a certain algorithm model will probably stay undetected despite the information obligation.

2.3.3 General Anti-discrimination Law

Specific rules regarding recommender systems or algorithms do not seem sufficient to tackle discriminatory recommendations. Nevertheless, they are not exhaustive in that area – also the general anti-discrimination rules apply and might sufficiently prevent discriminatory recommendations.

These rules, usually, on the national or EU level, e.g., forbid an unjustified unequal treatment according to certain personal features such as gender, race, disability, sexual orientation, age, social origin, nationality, faith or political opinion (list not exhaustive, depending on country or entity) (TFEU (EU) 2007, Art. 19; CFR (EU) 2012, Art. 21; Fundamental Law (Ger) 1949, Art. 3 para. 3; AGG(Ger) 2006 sec. 1). While many important features, therefore, are included, there is no general prohibition to treat people differently, e.g., for the region they live in or the dialect they speak or the color of their hair (Martini 2019, 238; Wachter forthcoming). Of course, those features can accumulate to features protected by anti-discrimination law, e.g., the region and the dialect of a person can allow conclusions regarding the ethical or social background (see above, Sect. 2.4.). But the general rule remains that discrimination is allowed as long as an explicitly mentioned feature is not the reason.

Applying anti-discrimination rules to the relationship between the provider of a recommender system and a user raises some further issues. First, those anti-discrimination rules primarily were drafted to protect the citizens against the State. If a public agency, for instance, uses a recommender system as a recruiting tool, anti-discrimination law applies directly.[6] On the other hand, the effect of these rules in private legal relationships, where the majority of recommender systems is used, is less easy to establish and highly disputed (Knebel 2018, 33 ff.; Perner 2013, 143 ff.; Schaaf 2021, 249; Neuner 2020, 1851–1855). Additionally, recommender systems are often used without the conclusion of a contract, thus, they move in the pre-contractual area where the parties' responsibility is traditionally harder to establish. Nevertheless, a tendency can be observed that the prohibition of discrimination slowly creeps into private relationships, especially contract law and employment law, at least in the EU (AGG (Ger) 2006, sec. 2, 7 para. 2, 21 para. 4; Hellgardt 2018, 901; Perner 2013, 145 ff.). Several EU anti-discrimination directives (Directive 2000/43/EC 2000, 22; Directive 2000/78/EC 2000, 16; Directive 2002/73/EC 2002, 15; Directive 2004/113/EC 2004, 37) as well as a constant flow of case law from the CJEU have enhanced this process and extended it to the pre-contractual level as well (CJEU 1976 Defrenne/SABENA, para 39; CJEU 2011 Test-Achats; Perner 2013, 157 ff.; Grüneberger and Reinelt 2020, 19 ff.). However, whether and to whom a provider of a recommender system is responsible if the recommender system is discriminatory, is unclear.[7]

Furthermore, there is the problem of indirect discrimination. As mentioned above, it will be easy to detect discrimination if the modelling algorithm uses a forbidden differentiation criterion. Nevertheless, a combination of other, not directly forbidden criteria, can lead to the same result (Ali et al. 2019, 1–30; Buolamwini and Gebru 2018, 12). Recruiting tools, for example, have often regarded résumés with longer periods without gainful employment as a sign of a weaker working performance. However, these periods can also be caused by breaks such as parental leaves or additional care obligations, typically involving more women than men. Thus, differentiating regarding that criterion, in consequence, can lead to the discrimination of women.

In anti-discrimination law it has been recognized that indirect discrimination can be forbidden as well (see Sect. 3 para. 2 AGG). The difference can become relevant for the requirements for the justification of unequal treatment. Unequal treatment can be justified if there are equally weighing values or interests on the other side to makeup the differentiation. This leads to a balancing of interests and risks of the people involved. Usually, direct discrimination weights more heavily and is almost impossible to justify, compared to an indirect one is (von Ungern-Sternberg forthcoming). Of course, the result also depends on the area of life where the

[6] E.g. Public Job Services, (see e.g. Allhutter et al. 2020).

[7] See, e.g. to the application of German Anti-discrimination law in the context of insurance recommendations Martini 2019, 234; see also to the problem of the scope of application of anti-discrimination law Hacker 2021 at fn. 88 to 98.

recommender system is used in. Thus, personalized ads are not as risky and relevant for the person involved as, for example, a job proposal or the exclusion of a job proposal.

Finally, the chain of responsibilities can be difficult. Often the recommender system is used by a platform but programmed by another business while the contract in question will be concluded between a user of the platform (e.g., an employer) and another user (e.g., the job seeker). Anti-discrimination law usually only has effects between the latter two, meaning that afterwards the possible employer must seek compensation from the platform provider who, in return, can seek compensation by the programmer. To ensure that the person of business finally responsible for the discriminatory algorithm is really forced to compensate the other parties, and, consequently, has an incentive to change the algorithm, is difficult in this way. Additionally, a justification might be possible if the functioning of the algorithm was not predictable to him or her as especially a self-learning algorithm is difficult to control regarding the data input and the improvement of the algorithm (black box problem).

2.3.4 Interim Conclusion

The legal frame only partly deals with discrimination by algorithms and is not sufficient to efficiently tackle it. Furthermore, the existing anti-discrimination law bears several uncertainties for all the parties involved.

2.4 Outlook

From these first conclusions, the next question is what should be done.

2.4.1 Extreme Solutions

One extreme possibility could be the prohibition to use machine learning algorithms in recommender systems at all. This would, of course, stop discriminations by recommendations, but also impede any progress regarding the use of machine learning algorithms or the development of recommender systems.

The other extreme solution could be a hands-off-approach and to leave it to the market powers to regulate the use of recommender systems. This approach also does not seem feasible as the past has shown that the mere play of market powers has been unable to prevent discrimination.[8]

[8] See e.g. regarding gender discrimination (Ekin 2018).

2.4.2 Further Development of the Information Approach

One possible solution between those two extreme positions could be a further development of the already existing information approach (Martini 2019, 187). Providers of recommender systems should also provide the necessary information for users to foresee and understand the risks of discrimination by a certain system in combination with an opt-out or opt-in possibility, meaning that they should not only have the choice to use the system or not, but also to use the system with the possible discrimination but also with alternatives. Furthermore, providers should be obliged to use a legal design that ensures that the people involved really read and understand the information (Martini 2019, 189; Kim and Routledge 2022, 97 ff.).[9]

This approach has also been chosen by a recent EU regulation, the Digital Service Act (DSA 2022/2065 (EU)). Article 27 para. 1 DSA states an explicit obligation for recommender systems proved by "online platforms" (not including "micro and small enterprises", Art. 19 DSA) to "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". Furthermore, according to Article 38 DSA, providers of very large online platforms that use recommender systems "shall provide at least one option for each of their recommender systems which is not based on profiling". Moreover, another proposed EU Act, the Artificial Intelligence Act (AIA 2021 (EU)), foresees that "AI" must be transparent and explainable for the user in Article 13 of the Commission Proposal (Kalbhenn 2021, 668).

This approach, in general, is a good step in the right direction. However, it has two flaws. First, Article 38 DSA only address "very large online platforms", platforms with more than 45 million recipients each month and designated as such by the Commission (Article 33 para. 1, 4 DSA). Recommender systems can, nevertheless, also be used in certain niche areas and be of high importance for the live of the parties involved, e.g., in certain job branches where highly specific people are recruited or searched. The AIA does not have this restriction. Besides, Article 38 only provides an "opt-out", meaning that users actively must choose not to use the proposed algorithm. The AIA does not provide any comparable consequences. Studies show that most users do not read the information but only continue to click to progress with the process they visited a certain platform for (Bergram et al. 2020; Martini 2019, 188). An opt-out possibility, therefore, is less efficient than an opt-in and nudges the users to just use what is already provided.

[9] Regarding the importance of design see e.g. (Machuletz and Böhme 2020, 481–498); see also the proposal to introduce counterfactual explanations as a complete information by (Wachter et al. 2018, 841–887).

2.4.3 Monitoring and Audit Obligations

The DSA also provides another interesting feature to control very large online platforms by establishing an obligation for regular audits (Article 37 DSA) to ensure that certain standards are met (Kalbhenn 2021, 671). Unfortunately, the audit obligation does not include recommender systems and possible discriminatory outcomes as mentioned in Articles 27, 38 DSA. An audit obligation, however, could be extended to possible discriminations, especially in areas where such discrimination can have massive effects on the life of the person involved, e.g. in questions of employment or job evaluation (Buolamwini and Gebru 2018, 12).

Therefore, it is no coincidence that another proposal for an EU Act, the Artificial Intelligence Act (AIA), also establishes an audit obligation for AI that is used in "high risk" areas, referring to areas that bear a high risk for the involved subjects. Contrary to the DSA, it applies no matter how many users a platform or provider has.

A similar approach can also be seen in other countries: The "Automated Employment Decision Tools" Bill by New York City (Law No. 2021/144 (Int 1894–2020) (NYC)) only allows the use of algorithms in employment decisions if the algorithm is subjected to a yearly audit. The advantage of such an audit is that the algorithm can be analyzed by specialists and nudge businesses to improve them (Raji and Buolamwini 2019). On the other hand, businesses only have to hand over trade sensitive information to those auditors, thus, their trade secrets can be respected and protected as well.

2.4.4 Interim Conclusion and Thoughts

To conclude, both (proposed) approaches of DSA and AIA, information/transparency and a regular audit obligation, should be combined for the use of recommender systems, at least in highly risky/sensitive areas for the person involved. An information obligation together with an opt-in possibility (rather than the opt-out-option provided in the DSA) and not limited to "very large platforms" would be feasible in those areas. Furthermore, a regular audit should be obligatory to ensure that possible discriminations in recommender systems can be found by the auditors and countered by them or others.

2.5 Conclusions

1. Recommender Systems based on algorithms can cause discrimination.
2. The existing legal framework is not sufficient to combat those discriminations. It is limited to certain information obligations and general non-discrimination rules that cannot provide the necessary legal certainty.

3. Information about the consequences of using a certain recommender system should be available for the people involved and phrased in a way that the users can understand it. Also, similar to Articles 27 para. 1, 38 DSA, at least an "opt-out" possibility should be provided, even though an opt-in possibility would be preferable.
4. A regular audit should be required, at least in areas that are highly sensitive to discrimination. This audit would allow the analysis by experts to find the reasons for discriminatory recommendations without endangering the trade secrets of the provider of the algorithm.

References

Act on Equal Treatment (*Allgemeines Gleichbehandlungsgesetz* – AGG). 2006. Available online: https://www.gesetze-im-internet.de/agg/index.html. Accessed on 06.09.2022.

Ali, Muhammad, Sapiezynski, Piotr, Bogen, Miranda, Korolova, Aleksandra, Mislove, Alan, and Rieke, Aaron. 2019. Discrimination Through Optimization. In Proceedings of the ACM on Human-Computer Interaction 3, CSCW 2019, 1–30.

Allhutter, Doris, Mager, Astrid, Cech, Florian, Fischer, Fabian, and Grill, Gabrial. 2020. Der AMS-Algorithmus: Eine Soziotechnische Analyse Des Arbeitsmarktchancen-Assistenz-Systems (AMAS). ITA-Projektbericht Nr.: 2020-02 2020.

Alpaydın, Ethem. 2016. *Machine Learning: The New AI. The MIT Press Essential Knowledge Series*. Cambridge, MA/London: MIT Press.

Angwin, Julia, Larson, Jeff, Mattu, Surya, and Lauren Kirchner. 2016. Machine Bias. *ProPublica*, May 23, Available online: https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing. Accessed on 07.09.2022.

Automated Employment Decision Tools Bill by New York City Law No. 2021/144 (Int 1894-2020) 2021 (New York City). Available online: https://www.assembly.ny.gov/leg/?bn=A07244&term=&Summary=Y&Actions=Y&Votes=Y&Memo=Y&Text=Y. Accessed on 06.09.2022.

Barocas, Solon, Hardt, Moritz, and Narayanan, Arvind. 2018. Fairness and Machine Learning: Limitations and Opportunities. 2018 (last update 2022). Online book available at https://fairmlbook.org/pdf/fairmlbook.pdf. Accessed on 07.09.2022.

Bergram, Kristoffer, Bezençon, Valéry, Maingot, Paul, Gjerlufsen, Tony, and Holzer, Adrian. 2020. Digital Nudges for Privacy Awareness: From Consent to Informed Consent?. In Proceedings of the 28th European Conference on Information Systems (ECIS), An Online AIS Conference, June 15–17, 2020. Available online https://aisel.aisnet.org/ecis2020_rp/64. Accessed on 06.09.2022.

Blass, Joseph. 2019. Algorithmic Advertising Discrimination. *Northwestern University Law Review* 114 (2): 415–468.

Bucher-Koenen, Tabea, Hackethal, Andreas, Koenen, Johannes, and Laudenbach, Christine. 2021. Gender Differences in Financial Advice. ECONtribute Discussion Paper no. 095 2021.

Buolamwini, Joy, and Gebru, Timnit. 2018. Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification. In Proceedings of Machine Learning Research, vol. 81, issue 1.

Calders, Toon, and Indrė Žliobaitė. 2013. Why Unbiased Computational Processes Can Lead to Discriminative Decision Procedures. In *Discrimination and Privacy in the Information Society*, ed. Bart Custers et al., vol. 3, 43–57. Berlin/Heidelberg: Springer.

Carolan, Eoin. 2016. The Continuing Problems with Online Consent Under the EU's Emerging Data Protection Principles. *Computer Law & Security Review* 32 (3): 462–473.

Cavalluzzo, Ken S., and Cavalluzzo, Linda C. 1998. Market Structure and Discrimination: The Case of Small Businesses. Journal of Money, Credit and Banking 30(4): 771–792. Available online: https://EconPapers.repec.org/RePEc:mcb:jmoncb:v:30:y:1998:i:4: p. 771-92. Accessed on 06.09.2022.

Chouldechova, Alexandra. 2017. Fair Prediction with Disparate Impact: A Study of Bias in Recidivism Prediction Instruments. *Big Data* 5 (2): 153–163.

Court of Justice of the European Union (CJEU). Defrenne/SABENA. ECLI:EU:C:1976:56.

———. Test-Achats. ECLI:EU:C:2011:100.

Cofone, Ignacio N. 2019. Algorithmic Discrimination is an Information Problem. *Hastings Law Journal* 70: 1389–1443.

Directive 2000/43/EC of 29 June 2000 implementing the principle of equal treatment between persons irrespective of racial or ethnic origin, OJ L 180, 19.7.2000.

Directive 2000/78/EC of 27 November 2000 establishing a general framework for equal treatment in employment and occupation, OJ L 303, 2.12.2000.

Directive 2002/73/EC of 23 September 2002 Amending Council Directive 76/207/EEC on the implementation of the principle of equal treatment for men and women as regards access to employment, vocational training and promotion, and working conditions, OJ L 269, 5.10.2002.

Directive 2004/113/EC of 13 December 2004 implementing the principle of equal treatment between men and women in the access to and supply of goods and services, OJ L 373, 21.12.2004.

Directive 2005/29/EC of 11 May 2005 concerning unfair business-to-consumer commercial practices in the internal market and amending Council Directive 84/450/EEC, Directives 97/7/EC, 98/27/EC and 2002/65/EC of the European Parliament and of the Council and Regulation (EC) No 2006/2004 of the European Parliament and of the Council ('Unfair Commercial Practices Directive'), OJ L 149, 11.6.2005: 22.

Directive (EU) 2019/2161 of 27 November 2019 amending Directive 93/13/EEC and Directives 98/6/EC, 2005/29/EC and 2011/83/EU as regards the better enforcement and modernisation of Union consumer protection rules, OJ L 328, 18.12.2019.

Ekin, Annette. 2018. Quotas get more women on boards and stir change from within. *Horizon*. The EU Research and Innovation Magazine. Available online: https://ec.europa.eu/research-and-innovation/en/horizon-magazine/quotas-get-more-women-boards-and-stir-change-within.

EU Commission. 2012. EU Commission regarding the predecessor rule Article 15 DPD, COM(92) 422 final – SYN 287.

Flores, Anthony W., Kristin Bechtel, and Christopher T. Lowenkamp. 2016. False Positives, False Negatives, and False Analyses: A Rejoinder to Machine Bias: There's Software Used Across the Country to Predict Future Criminals. And It's Biased Against Blacks. *Federal Probation* 80 (2): 38–46.

Fundamental Law (*Grundgesetz* – German Constitution) 1949 (Germany). Available online: https://www.gesetze-im-internet.de/gg/index.html. Accessed on 06.09.2022.

Gerards, Janneke, and Raphaële Xenidis. 2021. *Algorithmic Discrimination in Europe: Challenges and Opportunities for Gender Equality and Non-Discrimination Law: A Special Report*. Luxembourg: Publications Office of the European Union.

Gerberding, Johannes, and Gert G. Wagner. 2019. Qualitätssicherung Für "Predictive Analytics" Durch Digitale Algorithmen. *Zeitschrift für Rechtspolitik* 2019: 116–119.

German Government. 2000. Gesetzentwurf der Bundesregierung. Entwurf eines Gesetzes zur Änderung des Bundesdatenschutzgesetzes und anderer Gesetze. Bundestags-Drucksache 14/4329. Available online: https://dserver.bundestag.de/btd/14/043/1404329.pdf. Accessed on 06.09.2022.

Gershgorn, Dave. 2018. Amazons "holy grail" recruiting tool was actually just biased against women. *Quartz*, October 10, Available online: https://qz.com/1419228/amazons-ai-powered-recruiting-tool-was-biased-against-women. Accessed on 07.09.2022.

Grünberger, Michael, and André Reinelt. 2020. *Konfliktlinien im Nichtdiskriminierungsrecht: Das Rechtsdurchsetzungsregime aus Sicht soziologischer Jurisprudenz*, 2020. Tübingen: Mohr Siebeck.

Hacker, Philipp. 2021. A Legal Framework for AI Training Data – From First Principles to the Artificial Intelligence Act. *Law, Innovation and Technology* 13 (2): 257–301.

Handley, Ian M., Elizabeth R. Brown, Corinne A. Moss-Racusin, and Jessi L. Smith. 2015. Quality of Evidence Revealing Subtle Gender Biases in Science is in the Eye of the Beholder. *Proceedings of the National Academy of Sciences of the United States of America* 112 (43): 13201–13206.

Hellgardt, Alexander. 2018. Wer Hat Angst Vor Der Unmittelbaren Drittwirkung? *Juristen Zeitung* 73 (19): 901.

Kalbhenn, Jan C. 2021. Designvorgaben Für Chatbots, Deepfakes Und Emotionserkennungssysteme: Der Vorschlag Der Europäischen Kommission Zu Einer KI-VO Als Erweiterung Der Medienrechtlichen Plattformregulierung. *Zeitschrift für Urheber- und Medienrecht* 2021: 663–674.

Kasirzadeh, A., and A. Smart. 2021. The Use and Misuse of Counterfactuals in Ethical Machine Learning. In Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, 228–236. https://doi.org/10.1145/3442188.3445886

Kelleher, John D. 2019. *Deep Learning. The MIT Press Essential Knowledge Series*. Cambridge, MA/London: MIT Press.

Kim, Tae W., and Bryan R. Routledge. 2022. Why a Right to an Explanation of Algorithmic Decision-Making Should Exist: A Trust-Based Approach. *Business Ethics Quarterly* 32 (1): 75–102.

Knebel, Sophie V. 2018. *Die Drittwirkung Der Grundrechte Und -Freiheiten Gegenüber Privaten*, 2018. Baden-Baden: Nomos.

Lambrecht, Anja, and Catherine Tucker. 2020. Algorithmic Bias? An Empirical Study of Apparent Gender-Based Discrimination in the Display of STEM Career Ads. *Management Science* 65 (7): 2966–2981.

Machuletz, Dominique, and Böhme, Rainer. 2020. Multiple Purposes, Multiple Problems: A User Study of Consent Dialogs After GDPR. In Proceedings on Privacy Enhancing Technologies, 2, 481–498. Available online: http://arxiv.org/pdf/1908.10048v2. Accessed on 06.09.2022.

Martini, Mario. 2019. *Blackbox Algorithmus – Grundfragen Einer Regulierung Künstlicher Intelligenz*. Berlin/Heidelberg: Springer.

Moss-Racusin, Corinne A., John F. Dovidio, Victoria L. Brescoll, Mark J. Graham, and Jo Handelsman. 2012. Science Faculty's Subtle Gender Biases Favor Male Students. *Proceedings of the National Academy of Sciences of the United States of America* 109 (41): 16474–16479.

Mothilal, Ramaravind K., Amit Sharma, and Chenhao Tan. 2020. Explaining Machine Learning Classifiers Through Diverse Counterfactual Explanations. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, ed. Mireille Hildebrandt et al., 607–617. New York: ACM.

Neuner, Jörg. 2020. Das BVerfG Im Labyrinth Der Drittwirkung. *Neue Juristische Wochenschrift* 26: 1851–1855.

OLG Frankfurt/M. 2015. ZD.

Oosterhuis, Harrie, and Maarten de Rijke. 2020. Taking the Counterfactual Online: Efficient and Unbiased Online Evaluation for Ranking. In *Proceedings of the 2020 ACM SIGIR on International Conference on Theory of Information Retrieval*, ed. Krisztian Balog et al., 137–144. New York: ACM.

Özgümüs, Asri, Holger A. Rau, Stefan T. Trautmann, and Christian König-Kersting. 2020. Gender Bias in the Evaluation of Teaching Materials. *Frontiers in Psychology* 11: 1074.

Perner, Stefan. 2013. Grundfreiheiten, Grundrechte-Charta und Privatrecht. Beiträge zum ausländischen und internationalen Privatrecht 98. Tübingen: Mohr Siebeck.

Projektgruppe Wirtschaft, Arbeit, Green IT der Enquete-Kommission Internet und digitale Gesellschaft. 2013. Achter Zwischenbericht der Enquete-Kommission “Internet und digitale Gesellschaft” Wirtschaft, Arbeit, Green IT. Bundestagsdrucksache 17/12505.Proposal for a Regulation of the European Parliament and of the Council on a Single Market For Digital Services (Digital Services Act) and amending Directive 2000/31/EC, COM(2020) 825 final; consolidated text as adopted by the European Parliament P9_TC1-COD(2020)036. Available online: https://eur-lex.europa.eu/procedure/EN/2020_361. Accessed on 06.09.2022.

Rahman, Farhan. 2020. COMPAS Case Study: Fairness of a Machine Learning Model. *Towards Data Science*, September 7. Available online: https://towardsdatascience.com/compas-case-study-fairness-of-a-machine-learning-model-f0f804108751. Accessed on 31.05.2022

Raji, Inioluwa D., and Buolamwini, Joy. 2019. Actionable Auditing: Investigating the Impact of Publicly Naming Biased Performance Results of Commercial AI Products. AIES-119 Paper N0. 223. Available online: https://dam-prod.media.mit.edu/x/2019/01/24/AIES-19_paper_223.pdf. Accessed on 06.09.2022.

Regulation on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, General Data Protection Regulation (GDPR) (EU) 2016, OJ L 119, 4 May 2016.

Regulation (EU) 2019/1150 of 20 June 2019 on promoting fairness and transparency for business users of online intermediation services, OJ L 186, 11.7.2019OJ L 186, 11.7.2019.

Reiners, Bailey. 2021. 57 Diversity in the Workplace Statistics You Should Know, *Builtin*, October 20. Available online https://builtin.com/diversity-inclusion/diversity-in-the-workplace-statistics. Accessed 07.09.2022.

Rosenblat, Alex; Levy, Karen E. C.; Barocas, Solon; Hwang, Tim. 2017. Discriminating Tastes: Uber’s Customer Ratings as Vehicles for Workplace Discrimination. Policy & Internet 9 (3): 256–279.

Schaaf, Henning. 2021. Drittwirkung Der Grundrechte – Dogmatik Und Fallbearbeitung. *JURA – Juristische Ausbildung* 43 (3): 249–257.

Sheltzer, Jason M., and Joan C. Smith. 2014. Elite Male Faculty in the Life Sciences Employ Fewer Women. *Proceedings of the National Academy of Sciences of the United States of America* 111 (28): 10107–10112.

Speicher, Till, Muhammad Ali, Giridhari Venkatadri, Filipe N. Ribeiro, George Arvanitakis, Fabrício Benevenuto, Krishna P. Gummadi, Patrick Loiseau, and Alan Mislove. 2018. Potential for Discrimination in Online Targeted Advertising. *Proceedings of Machine Learning Research* 81 (1).

von Ungern-Sternberg, Antje. forthcoming. Discriminatory AI and the Law: Legal Standards for Algorithmic Profiling. In *Responsible AI*, ed. Silja Vöneky et al.

Vöneky, Silja. 2020. Key Elements of Responsible Artificial Intelligence – Disruptive Technologies, Dynamic Law. *Ordnung der Wissenschaft* 1: 9–22.

Wachter, Sandra. 2020. Affinity Profiling and Discrimination by Association in Online Behavioral Advertising. *Berkeley Technology Law Journal* 35 (2): 367–430.

———. forthcoming. The Theory of Artificial Immutability: Protecting Algorithmic Groups under Anti-Discrimination Law. *Tulane Law Review* 97: 2022–2023.

Wachter, Sandra, Brent Mittelstadt, and Chris Russell. 2018. Counterfactual Explanations without Opening the Black Box: Automated Decisions and the GDPR. *Harvard Journal of Law & Technology* 31 (2): 841–887.

Wikipedia. 2022. Science, technology, engineering, and mathematics. Available online: https://en.wikipedia.org/wiki/Science,_technology,_engineering,_and_mathematics. Accessed on 05.09.2022.

Chapter 3
From Algorithmic Transparency to Algorithmic Choice: European Perspectives on Recommender Systems and Platform Regulation

Christoph Busch

Abstract Algorithmic recommendations and rankings have become a key feature of the user experience offered by digital platforms. Recommender systems determine which information and options are prominently presented to users. While there is abundant technical literature on recommender systems, the topic has only recently attracted the attention of the European legislator. This chapter scrutinizes the emerging European regulatory framework for algorithmic rankings and recommendations in the platform economy with a specific focus on online retail platforms. Surveying the new rules for rankings and recommender systems in consumer contract law, unfair commercial practices law, and platform regulation, it identifies shortcomings and inconsistencies and highlights the need for coherence between the different regulatory regimes. The Digital Services Act could change the regulatory trajectory by introducing (albeit hesitantly and incompletely) a new regulatory model that shifts the focus from algorithmic transparency to algorithmic choice. More importantly, a choice-based approach to recommender governance and a market for third-party recommender systems ("RecommenderTech") could also be facilitated by the new interoperability requirements introduced by the Digital Markets Act.

Keywords Algorithmic transparency · Digital Services Act · P2B regulation · Recommender systems

C. Busch (✉)
University of Osnabrück, Osnabrück, Germany
e-mail: christoph.busch@uos.de

S. Genovesi et al. (eds.), *Recommender Systems: Legal and Ethical Issues*, The International Library of Ethics, Law and Technology 40,
https://doi.org/10.1007/978-3-031-34804-4_3

3.1 Introduction

Algorithmic rankings and recommendations constitute an essential element of the architecture of digital platforms (see Jannach and Adomavicius 2016). Recommender systems facilitate shopping online (Amazon), booking holiday rentals (Airbnb), discovering new movies (Netflix) or even dating (Tinder). By determining which information and options are prominently presented on a platform and which content remains hidden, automatic recommendations and rankings affect the choice architectures for consumers (see Hildebrandt 2022). Although recommender systems assist consumers in filtering information and may help to improve overall decision quality (Häubl and Trifts 2000), overdependence on algorithmic recommendations and rankings can reduce competition and harm consumers (see Banker and Khetani 2019). Moreover, recommender systems are a key source of platform power and a tool for private ordering by platform operators (see e.g., Leerssen 2020a; Cobbe and Singh 2019). In order to mitigate risks for competition and consumers, legislators at EU level and member state level have started to introduce new regulatory requirements for algorithmic rankings and recommendations on digital platforms.

Against this background, this chapter will scrutinize the emerging regulatory framework for algorithmic recommender systems in the European Union. Much of the academic and political debate has focused primarily on recommender systems used by social media platforms such as Facebook, Twitter, and TikTok and their societal effects (see e.g., Leerssen 2020a; Helberger et al. 2018; Milano et al. 2020). In contrast, this paper will seek to fill a research gap by focusing mainly on product recommendations on online retail platforms. In doing so, the chapter makes three contributions to the literature on platform regulation and recommender systems: First, it surveys the new rules for rankings and recommender systems in consumer contract law, unfair commercial practices law, and platform regulation. Second, it identifies gaps and inconsistencies and highlights the need to ensure coherence between the different regulatory regimes. Third, it argues that the European legislator should go beyond the current regulatory model based on algorithmic transparency and embrace new regulatory tools which enable users to control the functioning of rankings and recommendations and to choose between different competing recommender systems.

The rest of the chapter is organized as follows: Part II sets the scene by giving a brief overview of transparency requirements regarding ranking criteria as well as paid search results and paid rankings under the Unfair Commercial Practices Directive (UCPD),[1] the Consumer Rights Directive (CRD)[2] and the Platform-to-Business (P2B) Regulation.[3] In addition, this Part offers some terminological

[1] Directive 2005/29/EC concerning unfair business-to-consumer commercial practices in the internal market [2005] OJ L149/22 (Unfair Commercial Practices Directive).

[2] Directive 2011/83/EU on consumer rights [2011] OJ L304/64 (Consumer Rights Directive).

[3] Regulation (EU) 2019/1150 on promoting fairness and transparency for business users of online intermediation services [2019] OJ L186/57 (P2B Regulation).

clarifications and explains what EU law means when it speaks about rankings and recommender systems. Building on this overview, Part III provides a more detailed comparative analysis of the relevant provisions and scrutinizes emerging differences and commonalities in the field of EU recommender governance. Part IV then turns the focus to the Digital Services Act (DSA),[4] the latest addition to the emerging framework for recommender governance in the EU. This Part explains that the DSA (albeit hesitantly and incompletely) introduces a new regulatory paradigm that shifts the focus from algorithmic transparency to algorithmic choice. Part V looks beyond the DSA and argues that a choice-based approach to recommender governance and a market for "RecommenderTech" could also be facilitated through new interoperability requirements introduced by the Digital Markets Act (DMA).[5] Finally, Part VI offers some conclusions.

3.2 Recommender Governance in the EU Platform Economy

Until very recently, there were no specific rules for algorithmic rankings and recommendations at EU level. The main legal requirements regarding the transparency of recommender systems stemmed essentially from general rules of unfair commercial practices law, in particular the prohibition of misleading practices under Art. 6 and 7 UCPD. Within the timespan of only a few years, the situation has fundamentally changed. Instead of too few, there may now be too many rules that are not sufficiently coordinated. Since 2019, the European legislator has enacted several new regulations aimed at increasing the transparency of rankings and algorithmic recommendations on digital platforms. As a result, the regulatory framework currently presents itself as a complex and fragmented landscape of partially overlapping transparency rules. This Part will briefly map out the relevant rules and provide some terminological clarifications before the next Part will analyze in more detail the differences and similarities between the different rules applicable to algorithmic rankings and recommendations.

3.2.1 Mapping the Regulatory Landscape

For reasons of clarity, the following overview will focus on three legal instruments, that are most relevant for algorithmic transparency regarding online retail platforms: the P2B Regulation, the UCPD, and the CRD. The DSA, which will introduce a shift towards a new regulatory model combining algorithmic transparency and

[4] Regulation (EU) 2022/2065 on a Single Market for Digital Services and amending Directive 2000/31/EC (Digital Services Act).

[5] Regulation (EU) 2022/1925 on contestable and fair markets in the digital sector and amending Directives (EU) 2019/1937 and (EU) 2020/1828 (Digital Markets Act).

algorithmic choice, will be addressed separately in Part V. It should be noted that in some cases there may be also overlaps with transparency obligations stemming from the field of media law, such as Sect. 93 of the German Interstate Media Treaty (*Medienstaatsvertrag*),[6] which imposes algorithmic transparency requirements for "media intermediaries". This category includes search engines and social media platforms.[7] With the growing convergence of social media and e-commerce ("social commerce") the dividing line between e-commerce law and media law is getting more and more blurred and we will most likely see also growing overlap between media law and e-commerce regulation (see e.g., Svirskis 2020). This will even further increase the complexity of EU recommender governance. The regulatory landscape may even become more complex with the forthcoming AI Act,[8] which will add further transparency requirements for algorithmic systems.[9]

This being said, there are currently mainly three legal instruments that define the regulatory framework for rankings and recommendations on online retail platforms: the P2B Regulation, the CRD, and the UCPD. Art. 5(1) P2B Regulation requires providers of "online intermediation services" (e.g., online retail marketplaces) to "set out set out in their terms and conditions the main parameters deter mining ranking and the reasons for the relative importance of those main parameters as opposed to other parameters". Art. 5(2) P2B Regulation stipulates a similar transparency rule for online search engines. Art. 5(3) to (7) P2B Regulation further spells out the details of the transparency duty. Of particular interest here is Art. 5(3) P2B Regulation, which lays down specific disclosure duties for cases where the position in the ranking can be influenced by direct or indirect payments, such as "ranking boosters" or preferred partner programs.

While the P2B Regulation aims at promoting fairness and transparency between platforms and business users of intermediation services offered by the platforms, the two other transparency requirements for rankings have been introduced in the context of the recent reform of EU consumer law. As part of the "New Deal for Consumers", the Modernisation Directive 2019/2161/EU has added in 2019 two new information requirements regarding online rankings to the CRD and the UCPD. According to Art. 6a(1) CRD, operators of online marketplaces have to provide consumers with general information on the "main parameters" determining the ranking of offers presented to the consumer as a result of a search query and the "relative importance of those parameters as opposed to other parameters". A similar information requirement is set out in Art. 7(4a) UCPD, which also stipulates a duty to inform about the main parameters of the ranking and their relative importance.

[6] State Treaty on the Modernisation of the Media Order in Germany of 28 October 2020 (German Interstate Media Treaty).

[7] See Art. 2(2) No. 16 German Interstate Media Treaty.

[8] European Commission, Proposal for a Regulation on artificial intelligence, COM(2021) 206 final (AI Act Proposal).

[9] See Arts. 13, 52 AI Act Proposal.

This provision is complemented by the new No. 11a of Annex I to the UCPD.[10] According to the new provision, it is under all circumstances prohibited to provide "search results in response to a consumer's online search query without clearly disclosing any paid advertisement or payment specifically for achieving for achieving higher ranking of products within the search results".[11]

3.2.2 Layers of Terminology in EU Law: "Rankings" and "Recommender Systems"

Before taking a closer look at the emerging European regulatory framework for recommender systems and rankings on digital platforms, some terminological clarifications may be necessary. From a technical perspective, the term "recommender systems" refers to software tools "that provide suggestions for items that are most likely of interest to a particular user" (Ricci et al. 2022). For this purpose, recommender systems may use collaborative filtering techniques, content-based filters, knowledge-based filtering mechanisms, or hybrids between these models (See e.g. Aggarwal 2016; Ricci et al. 2022). Simply put, the task of a recommender system is to help users "find good items or predict an item's relevance to a user" (Jannach and Adomavicius 2016).

While there is abundant technical literature on recommender systems (see, ex multis Ricci et al. 2022; Aggarwal 2016), the terms "recommender system" and "ranking" have only recently entered the vocabulary of the European legislator. One of the earliest EU references to rankings in the context of electronic commerce is not to be found in a Directive or Regulation, but in a somewhat apocryphal text, the "Key principles for comparison tools" of May 2016, which have been elaborated by a multi-stakeholder group (including consumer and business associations, providers of online comparison tools and national authorities) under the auspices of the European Commission (European Commission 2016a, b). These principles, which have been drafted to facilitate the application of the UCPD to online comparison tools such as Verivox or Yelp, do not define the term "ranking", but stipulate that "criteria used for the rankings should be clearly and prominently indicated, as well as, where relevant to ensure that consumers are not misled, general information about any specific methodology used" (European Commission 2016a, b). This non-binding requirement is explicitly referred to in the (equally non-binding) Commission Guidance on the implementation and application of the UCPD, which was also published in May 2016.[12] The brief look at these early traces of "recommender

[10] Annex I of the Directive contains a "blacklist" of commercial practices which are prohibited under all circumstances and do not require a case-by-case assessment regarding the likely impact of the practice on the average consumer's economic behaviour.

[11] No. 11a Annex I UCPD.

[12] European Commission, Guidance on the implementation and application of Directive 2005/29/EC on Unfair Commercial Practices, SWD(2016) 163 final, p. 123.

governance" in EU law underlines that the topic is rooted in unfair commercial practices law and goes back to the early days of EU platform regulation.

One of the first attempts to explicitly regulate recommender systems in the context of platform regulation was the P2B Regulation. Interestingly, neither the proposal[13] for the Regulation, which was published in April 2018, nor the final text of June 2019 uses the term "recommender system". Instead, the term "ranking mechanism" is used.[14] According to the definition in Art. 2(1)(m) P2B Regulation, "ranking" means

> the *relative* prominence given to the goods or services offered through online intermediation services, or the relevance given to search results by online search engines, as presented, organised or communicated by the providers of online intermediation services or by providers of online search engines, respectively, irrespective of the technological means used for such presentation, organisation or communication.[15]

This definition was copied almost verbatim in Art. 2(m) UCPD, which was added to the UCPD in November 2019 by the Modernisation Directive.[16] The definition is not only relevant for the new transparency requirements for rankings under Art. 7(4a) UCPD. The UCPD definition is also referred to explicitly in the transparency rule for rankings in Art. 6a(1) CRD.[17] It is interesting to note that neither the P2B Regulation nor the UCPD uses the term "recommender system". The alternative term "ranking" shifts the focus away from the technological innards of the recommender system to the *output* of the system. How this output is produced is irrelevant to the applicability of the transparency rules.

The definitions in Art. 2(1)(m) P2B Regulation and Art. 2(m) UCPD are also technology-neutral in another respect. Both definitions cover rankings of products presented, organized or communicated to platform users "irrespective of the technological means used for such presentation, organization or communication." Recital 19 of Directive (EU) 2019/2161 further explains that rankings can result "from the use of algorithmic sequencing, rating or review mechanisms, visual highlights, or other saliency tools, or combinations thereof". This underlines that a ranking is not necessarily a list of items, but can also take the form of a display of varying prominence on a map or in a word cloud. The common feature in each case is that the

[13] European Commission, Proposal for a Regulation on promoting fairness and transparency for business users of online intermediation services, COM(2018) 238 final.

[14] Recitals 24, 25 P2B Regulation.

[15] Art. 2(1)(m) P2B Regulation.

[16] See Directive (EU) 2019/2161 of the European Parliament and of the Council of 27 November 2019 amending Council Directive 93/13/EEC and Directives 98/6/EC, 2005/29/EC and 2011/83/EU of the European Parliament and of the Council as regards the better enforcement and modernization of Union consumer protection rules [2019] OJ L328/7 (Modernisation Directive), Art. 3(1)(b) ("'ranking' means the relative prominence given to products, as presented, organised or communicated by the trader, irrespective of the technological means used for such presentation, organisation or communication").

[17] The Commission proposal for the Modernisation Directive (COM(2018) 185 final) of 11 April 2018 did not contain a definition of the term "ranking". The definition was only added later to the final version.

individual items differ in their "relative prominence", as it is called in the two definitions.

The term "ranking" is also used in the draft Digital Markets Act (DMA). The definition used in the DMA is recognizably modeled on Art. 2(1)(m) P2B Regulation.[18] Recital 52 DMA underlines that the concept of ranking is meant to "cover all forms of relative prominence, including display, rating, linking or voice results". In this context, the DMA adds an interesting clarification by stating that a ranking can "include instances where a core platform service presents or communicates only one result to the end user". At first glance, this might seem paradoxical. However, this is probably due to the fact that the "core platform services" covered by the DMA also include virtual assistants.[19] When consumers use virtual assistants for online shopping, voice-controlled devices such as Amazon's Alexa do not read out long lists of ranked items. Rather they offer only a single recommendation ("Amazon's choice"). In a way, this takes the idea of a ranking to the extreme.

Interestingly, the Digital Services Act (DSA) utilizes a different terminology and uses the technical "recommender system". Art. 2(s) DSA defines "recommender systems" 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 prioritise 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.[20]

Despite the different terminology, the substance is more or less the same as in the other legal texts that use the term "ranking". As in the other legal instruments, the DSA provision essentially aims at systems that determine a "relative order of prominence" and thus establishes a choice architecture for the platform users. Thus, in summary, it can be said that despite the differences in terminology, the different legal instruments share one important characteristic: They do not distinguish between different types of recommender systems such as content-based systems, knowledge-based systems, collaborative systems, or hybrid systems. They are agnostic with regard to the technology used in the filtering process and focus on the output of the systems, i.e., the "relative order" or "prominence" they produce.

[18] See Art. 2(22) DMA ('ranking' means the relative prominence given to goods or services offered through online intermediation services, online social networking services, video-sharing platform services or virtual assistants, or the relevance given to search results by online search engines, as presented, organised or communicated by the undertakings providing online intermediation services, online social networking services, video-sharing platform services, virtual assistants or online search engines, irrespective of the technological means used for such presentation, organisation or communication and irrespective of whether only one result is presented or communicated).

[19] Art. 2(2)(h) DMA.

[20] Art. 2(s) DSA.

3.3 Five Axes of Algorithmic Transparency: A Comparative Analysis

This section will zoom in on the transparency rules for algorithmic rankings set out in Sect. 3.2. In doing so, we will analyze differences and similarities between the rules along five axes: (1) purposes of transparency, (2) audiences of transparency, (3) addressees of the duty, (4) content of the disclosure, and (5) modalities of disclosure.

3.3.1 Purpose of Transparency

The criteria and the "hidden logics" used to create the ranking and how they are weighted usually remain in the dark. From the point of view of both consumers and professional platform users, the ranking algorithms are black box systems (see Pasquale 2015). However, the transparency problem presents itself somewhat differently from the consumer and trader perspective.

From the consumer's point of view, the primary concern is to prevent unfair influence through a biased ranking. Empirical research suggests that "consumers are more likely to select options near the top of a list of results, simply by virtue of their position and independent of relevance, price or quality of the options" (UK Competition and Markets Authority 2021; see also Ursu 2018; De los Santos and Koulayev 2017). This position bias (or "ranking effect") may induce platform providers to exploit consumers by giving a higher position to products that are more profitable for the platform, but which are not necessarily the best choice for the consumer. Personalized rankings could have even more harmful effects and lead to similar results as personalized pricing if more expensive products are presented specifically to consumers with a higher willingness-to-pay ("price steering") (UK Competition and Markets Authority 2021). Reducing such risks for consumers through meaningful transparency of ranking criteria is the purpose of the transparency requirements stipulated by Art. 7(4a) UCPD and Art. 6a(1)(a) CRD.

From the perspective of traders who distribute their goods or services via an online marketplace, the transparency problem is different. Professional platform users are interested in learning more details about the functioning of the ranking mechanism in order to improve the visibility of their products on the platform ("ranking optimization"). If, for example, a hotel booking site informs their users how they can achieve a better position in the ranking if they display high-quality photos in their profile, a hotel owner can make a business decision on whether to invest in better photos. Similarly, detailed and truthful information about "ranking boosters" or "preferred partner programs" offered by the platform which promise to improve the position in the ranking will enable businesses to take an informed decision on how much they spend on such offers (see, e.g. Bundeskartellamt 2019).

Addressing these concerns is the main purpose of the transparency requirements under Art. 5 P2B Regulation.

The transparency rule set out by Art. 27(1) Draft DSA has a hybrid function. On the one hand, it protects the autonomy of platform users by providing them with information on how information is prioritized for them. At the same time, the purpose of the provision transcends the platform-user-relationship. This is underlined by Recital 70 DSA which explicitly mentions that recommender systems "play an important role in the amplification of certain messages, the viral dissemination of information and the stimulation of online behaviour." The Recital concludes: "Consequently, 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." In this sense, the DSA adds a broader, societal dimension to recommender governance.

3.3.2 Audiences of Disclosure

Closely linked to the purpose of the transparency requirement is the question to whom transparency shall be offered. In this sense, transparency is a relational concept that is defined by the audiences the disclosure duties serve (Leerssen 2020a). For transparency rules regarding social media recommender systems, a tripartite distinction has been suggested: (1) disclosures for users, (2) disclosures for public authorities, and (3) disclosures for academia and civil society (Leerssen 2020a). While this tiered approach could also be applied to transparency rules for rankings on online retail platforms, a slightly different approach seems preferable. Leaving the forthcoming DSA rules aside, the current EU regulatory framework for rankings and recommendations in the field of online retail clearly focuses on disclosures towards platform users. With regard to ranking transparency, the two other audiences are neither addressed in the P2B Regulation, the UCPD nor the CRD. However, the distinction between the two relevant sub-groups of "users" is of key importance for the effectiveness of ranking transparency in the e-commerce sector. In order to be effective, the information provided on ranking parameters has to be adjusted to the relevant target group.

For the transparency requirements under Art. 7(4a) UCPD the "average consumer test" applies.[21] Therefore, as a general rule, the information provided about the main parameters of the ranking and their relative importance must be intelligible for an average consumer who is "reasonably well-informed and reasonably observant and circumspect".[22] However, this standard has to be adjusted if the ranking

[21] See Art. 5(2)(b) UCPD.

[22] See Recital 18 UCPD; see also CJEU, 16.7.1998, C-210/96, ECLI:EU:C:1998:369 (Gut Springenheide); Schebesta and Purnhagen 2020; see also Weber 2020.

and its explanation are specifically aimed at a particular group of consumers or if the trader can foresee that the ranking will materially distort the economic behavior of an identifiable group of vulnerable consumers (e.g. children).[23] It is difficult to see, how a platform operator will be able to match this high standard given the technical complexity of many ranking mechanisms and whether such information will in effect be digestible for the consumer.

In contrast, the disclosures required by the P2B Regulation are directed at "business users" who offer their goods and services via online platforms (Art. 5(1) P2B Regulation) or "corporate website users" whose websites are ranked by search engines (Art. 5(2) P2B Regulation. In both cases, the audience of the transparency requirements consists of professionals who are interested in improving their online visibility for potential customers. As explained in the Commission's Guidelines on ranking transparency, in order to be meaningful for the professional audience, explanations about the main ranking parameters "should take account of the nature, technical ability and needs of 'average' users of a given service, which may vary considerably between different types of services".[24] In other words, in the P2B Regulation an "average business test" replaces the "average consumer test".[25]

3.3.3 Addressees of the Duty to Disclose

The transparency requirements differ not only in terms of their audiences, but with regard to those who are required to provide information about rankings. Among the legal instruments under comparison, Art. 6a(1) CRD has the narrowest scope. The provision only applies to "online marketplaces" i.e. websites and applications which allow consumers to conclude distance contracts with other traders or consumers.[26] This includes online retail marketplaces (e.g. Amazon.com) as well as hotel booking platforms (e.g. Booking.com), but excludes online search engines (e.g. Google.com) and price comparison tools (e.g. Shopping.com) which redirect consumers to the trader's website in order to conclude a contract. However, in the face of rapidly changing business models, the contours of the term "online marketplace" are not entirely clear. In particular, it is an open question under which conditions social

[23] See Art. 5(3) UCPD.

[24] European Commission, Guidelines on ranking transparency pursuant to Regulation (EU) 2019/1150 of the European Parliament and of the Council, 2020 O.J. (C 424/1) at para. 17.

[25] See also European Commission, Guidelines on ranking transparency pursuant to Regulation (EU) 2019/1150 of the European Parliament and of the Council, 2020 O.J. (C 424/1) at para. 105 ("In certain cases, more technical descriptions may be appropriate and required, bearing in mind that the descriptions are intended for professional users. Without prejudice to the requirement of using plain and intelligible language, professionals may in principle be assumed to require and be able to understand more detailed and more technical information than consumers.")

[26] See Art. 2(17) CRD (defining online marketplace as "a service using software, including a website, part of a website or an application, operated by or on behalf of a trader which allows consumers to conclude distance contracts with other traders or consumers").

media platforms that offer shopping features (e.g. Shoppable Posts on Instagram) fall under Art. 6a(1) CRD. While the answer seems to be negative if the contract is concluded entirely outside the social media app, it might be positive if the shopping website is displayed within the application.

Art. 7(4a) UCPD has a broader scope and applies regardless of where the contract is eventually concluded. Hence, the information requirements under Art. 7(4a) UCPD not only applies to online marketplaces, but also to price comparison sites and similar online tools.[27] In contrast, search engines (as defined in Art. 2(6) P2B Regulation) are explicitly excluded from the scope of Art. 7(4a) UCPD. Similarly, the UCPD provision does not apply where traders provide consumers with a possibility to search only among their own offers of different products.[28] Furthermore, the information requirement under Art. 7(4a) UCPD only applies where the ranking is displayed to the consumers on the basis of a search query (e.g. in the form of a keyword, phrase or other input). Therefore, it does not apply to the default display on the online interface that is shown to the consumer and that is not the result of a specific search query.[29]

The business-facing transparency rule under Art. 5 P2B Regulation has an even a wider scope. Art. 5(2) P2B Regulation applies to "online intermediary services". This term refers to online services that allow business users to offer their products to consumers "with a view to facilitating the initiating of direct transactions between those business users and consumers, irrespective of where those transactions are ultimately concluded".[30] However, it is required that the intermediation service is provided on the basis of a contractual relationship between the platform operator and the business user. Similar to Art. 7(4a) UCPD, the broad definition of "online intermediary services" does not only cover online marketplaces (where contracts are concluded), but also other websites or online interfaces (e.g. apps, virtual assistants) that "facilitate the initiating of direct transactions". Here, social media platforms that offer shopping features clearly seem to be covered.

The broadest scope among the four legal instruments that stipulate transparency requirements for rankings and recommendations is found in Art. 27 DSA. The provision applies to all "providers of online platforms" that use recommender systems. According to Art. 2(i) DSA the term "online platform" means "a hosting service that, at the request of a recipient of the service, stores and disseminates information to the public, unless that activity is a minor and purely ancillary feature of another service or a minor functionality of the principal service and, for objective and technical reasons, cannot be used without that other service, and the integration of the feature or functionality into the other service is not a means to circumvent the

[27] European Commission, Guidance on the interpretation and application of Directive 2011/83/EU [2021] OJ C525/01 (CRD Guidance), at 3.4.1.

[28] European Commission, Guidance on the interpretation and application of Directive 2005/29/EC [2021] OJ C526/01 (UCPD Guidance), at 4.23.

[29] Ibid.

[30] Art. 2(2)(b) P2B Regulation.

applicability of this Regulation". In other words, not only marketplaces but also communication platforms are covered.

3.3.4 Content of the Disclosure

While the scope of application of the two transparency rules in the CRD and the UCPD is different, the content of the information duties is more or less the same. Both provisions require businesses to provide consumers only with information on the "main parameters" determining the ranking of offers. In other words, there is no obligation to disclose all ranking parameters. Requiring the platform operator to indicate all factors that influence the ranking would most likely lead to an information overload as most ranking mechanisms take into account a rather large number of factors that are weighted according to a complex formula (Alexander 2019). Therefore, limiting the transparency requirement to "general information" about the "main parameters" seems reasonable.

For example, the short-term rental platform Airbnb states that their "search algorithm considers more than 100 signals to decide how to order listings in search results".[31] Among the factors that are taken into account by the ranking algorithm are guest reviews, competitive pricing, availability for instant booking, host response time, and superhost status (Airbnb UK 2022). These factors seem to be linked to the overall attractiveness of a listing. By considering these factors the ranking of listings is likely to match the preferences of users who are looking for an attractive offer. However, the design of Airbnb's ranking algorithm is not solely based on the preferences of guests. It also reflects the interests of prospective hosts and the platform provider itself. In this sense, Airbnb states that to "help hosts get started, the algorithm is designed to make sure new listings show up well in search results" (Airbnb Ireland 2022). Giving priority to listings simply because they are new is not necessarily in the interest of platform users who are looking for high-quality offers. However, it is probably in the own interest of Airbnb to give a boost to new listings in order to keep new hosts satisfied with the platform.

The Modernisation Directive, which introduces the new transparency requirements into the CRD and UCPD, does not provide very detailed and actionable guidance on how to determine the "main parameters" of the ranking mechanism. As explained in the Recitals of the Directive, the term "parameters" refers to "any general criteria, processes, specific signals incorporated into algorithms or other adjustment or demotion mechanisms used in connection with the ranking".[32] This list of synonyms is not very helpful in elucidating the meaning of the term (Peifer 2021). Among the input variables which impact the ranking, "main parameters" are those

[31] Airbnb UK 2022. Interestingly, the information provided on the Irish Airbnb website (Airbnb.ie) is vaguer and only indicates that their "algorithm considers many factors to determine how to order search results but some factors have a larger impact than others" (Airbnb Ireland 2022).

[32] Recital 22 Modernisation Directive.

"which individually or collectively are most significant in determining ranking".[33] It remains an open question how exactly the "most significant" factors shall be determined. Are these the factors that individually or collectively account for 50% of determining the ranking position? What if different combinations of criteria change the weighting of the individual criteria? What about dynamic systems that apply temporary changes (e.g. for Black Friday or Christmas shopping) or use A/B testing in order to optimize the ranking mechanism? Arguably, it would be disproportionate to require real-time adjustment of information in cases where businesses use dynamic ranking systems (Peifer 2021).

Once the main parameters have been determined, the CRD and the UCPD require only "general information" about their influence on the ranking. Traders are not required to disclose the detailed functioning of their ranking mechanism or even the underlying algorithm.[34] It is also not necessary to present the information about the ranking in a customized manner for each individual search query.[35] In this sense, transparency is limited to a rather general "model explanation" and does not require individualized "outcome explanation".[36] This is particularly important for personalized rankings which are based on the customer's purchasing history or other elements of an individual customer profile (see, generally, Kant 2020; Cohn 2019). In such a case the trader is not obliged to provide personalized information about the personalized ranking but can limit herself to a general description of the parameters used for personalization (Alexander 2019).

Both the CRD and UCPD also require traders to indicate the "relative importance" of the main parameters as opposed to other parameters. However, the two Directives give no details as to how the relative importance of the main parameters should be indicated. One option could be to indicate the weighting of the main parameters in percentage points, as in the following example: "The top five parameters are weighted with 70% while the remaining fifteen parameters are only weighted with 30% percent." An alternative could be to use standardized information about ranking factors similar to the nutrition fact labels which indicate the percentage of different ingredients in food products (see Stoyanovich et al. 2018). Such a standardized "Nutrition Label for Rankings" could in particular facilitate the comparison of different ranking mechanisms. In order to increase the comprehensibility of the label, graphic elements and pictograms could also be used.[37]

On the business side, Art. 5 P2B Regulation also requires information about the "main parameters" which determine the ranking. With regard to further details about the functionality of the ranking mechanism, the P2B distinguishes between "online intermediation services" and "online search engines". Providers of search

[33] Recital 21 Modernisation Directive.

[34] Recital 23 Modernisation Directive.

[35] Ibid.

[36] See, e.g. Grochowski et al. 2021, discussing different dimensions and addressees of transparency and explainability.

[37] See also Art. 12(7) GDPR which suggests that information about the processing of personal data may be provided in combination with standardized and machine-readable icons.

engines have to provide information about the “relative importance” of the main parameters.[38] In contrast, providers of online intermediation services shall explain “the reasons for the relative importance” of the main parameters.[39] While the former follows the model used in the CRD and the UCPD, the latter deviates from this model. One may wonder whether this terminological discrepancy is an expression of an underlying substantive difference between the two transparency regimes. From this perspective, Art. 5(2) P2B Regulation could be understood in the sense that online search engines only have to indicate the weighting of the ranking parameters (as a percentage), but are not obliged to give a reason for the chosen weighting. Ultimately, however, the terminological differences appear to be an editorial inaccuracy that has no deeper impact on the scope of the disclosure obligation. This reading is also supported by the (non-binding) Commission Guidelines on ranking transparency. There, Art. 5(1) and (2) P2B Regulation are mentioned in the same breath, without addressing a possible differentiation of the information obligations: “The descriptions given by providers in accordance with Article 5 should provide real added-value to the users concerned. Articles 5(1) and (2) require that providers give information not only of the main parameters but also the reasons for the relative importance of those main parameters as opposed to other parameters.”[40] Therefore, providers of online intermediation services and online search engines have to “go beyond a simple enumeration of the main parameters”[41] and provide a “second layer”[42] of explanatory information that explain what objective the ranking mechanism has been optimized for.

Given the diversity of ranking algorithms, the content of the explanations required under Art. 5(1) and (2) P2B Regulation may vary significantly depending on the design of the ranking mechanism. Art. 5(3) and (5) P2B Regulation, therefore, specify the requirements for the content of the explanations. This establishes a mandatory minimum content of the disclosures. On the one hand, Art. 5(3) P2B Regulation states that providers must inform about the possibility that users can influence the ranking through direct or indirect fees, e.g. through temporary “ranking boosters” or a preferred partner program. On the other hand, Art. 5(5) P2B Regulation defines a list of mandatory disclosures. Among other things, it must be indicated whether and to what extent the ranking mechanism takes into account the characteristics of the products offered (lit. a) and the relevance of these characteristics for consumers (lit. b). Providers of online search engines must also indicate the extent to which design characteristics of the website (e.g. page speed, optimization for mobile devices) influence the ranking of the website (lit. c). These mandatory

[38] Art. 5(2) P2B Regulation.

[39] Art. 5(1) P2B Regulation.

[40] European Commission, Guidelines on ranking transparency pursuant to Regulation (EU) 2019/1150 of the European Parliament and of the Council, 2020 O.J. (C 424/1) at para. 22.

[41] Ibid.

[42] Ibid.

disclosures aim to achieve a certain standardization and thus improve the comparability of the ranking practices of different providers.

Finally, Art. 5(6) P2B Regulation defines the limits of ranking transparency. According to this provision, platform operators are not required to disclose algorithms or any information that would enable manipulation of the ranking mechanism. However, Art. 5(6) P2B Regulation does not allow platform operators to limit transparency on the blanket grounds that this is necessary to prevent "gaming of the system". Instead, they have to provide evidence that further disclosure "with reasonable certainty" would open the door for manipulation of the search results and thus create harm for consumers.

3.3.5 Modalities of Disclosure

In addition to the content of the information duties, the *modalities* of disclosure are relevant in order to assess the effectiveness of the transparency requirements. In this respect, too, there are differences between the regulations under comparison. Basically, two different models can be distinguished:

The UCPD and the CRD require that the information about the main ranking parameters and their relative importance "made available in a specific section of the online interface that is directly accessible from the page" where the search query results or the offers are presented.[43] In other words, the information has to be provided at the place where the ranking is displayed to the consumer. Apparently, the EU legislator is guided by the idea that the consumer consults the information at the moment of the purchasing decision. Whether this is a realistic assumption seems rather doubtful in view of the complexity of the information.

In contrast, Art. 5(1) P2B Regulation stipulates that the explanation of the main parameters determining the ranking and the reasons for the relative importance of those parameters as opposed to other parameters are set out in the terms and conditions of the provider of online intermediation services. As can be seen from Art. 3(1)(b) P2B Regulation, the terms and conditions – and thus also the explanation of the rankings – must be made available to business users before they create an account on the platform. This shows that the information shall enable business users to make an informed decision about whether to use a given platform as a distribution channel. Furthermore, businesses shall be enabled to improve their visibility through "ranking optimization". For online search engines, Art. 5(2) P2B Regulation applies a slightly different model. As "corporate website users" do not have to conclude a contract with the provider of the online search engine for their website to be ranked, the information does not have to be included in the terms and conditions of the search engine. Instead, Art. 5(2) P2B Regulation requires that the provider of the search engine provides an "easily and publicly available description" of the main

[43] Art. 7(4a) UCPD and Art. 6a(1)(a) CRD.

ranking parameters. This shall enable the corporate website users to engage in meaningful (and legally acceptable) forms of "search engine optimization".

3.4 The Digital Services Act: From Algorithmic Transparency to Algorithmic Choice?

The most recent layer of EU recommender regulation has been added by the Digital Services Act (DSA) which was formally adopted in October 2022 and will apply from 17 February 2024. In a sense, the DSA marks the transition towards a new regulatory model that goes beyond algorithmic transparency and makes a first tentative step towards *algorithmic choice*.

3.4.1 Extension of Transparency Rules

The DSA not only introduces a definition of "recommender systems" to the EU regulatory framework but also extends the existing rules on algorithmic transparency. Notwithstanding the minor differences in their respective scope of application, the transparency rules for recommender systems stipulated by the P2B Regulation, the CRD, and the UCPD primarily focus on digital platforms that facilitate the initiating of transactions between platform users.

The DSA goes one step further and introduces transparency requirements that apply to all online platforms that use recommender systems regardless of whether the recommendations are meant to facilitate any transactions between platform users. Therefore, the new transparency rules also apply to platforms such as Twitter, Spotify, or Tinder. The Commission's original proposal for the DSA provided that the transparency rules would only apply to very large platforms (VLOPs) with more than 45 million monthly active users.[44] However, during the trilogue negotiations, this regulation was extended to all online platforms, regardless of the number of users.[45] The final version of Art. 27(1) DSA now requires all online platforms that use recommender systems to "set out in their terms and conditions, in plain and intelligible language, the main parameters use in their recommender systems, as well as any options for the recipients of the services to modify or influence those main parameters".

If one compares this provision with the consumer-facing transparency rules of Art. 7(4a) UCPD and Art. 6(1)(a) CRD, it comes a bit as a surprise that the DSA allows platform providers to hide the information about the main parameters in the

[44] Art. 29(1) DSA Commission Proposal.

[45] Art. 24a(1) DSA Provisional Agreement.

small print of the platform's terms and conditions.[46] It seems that even the drafters of the provisions did not really believe that platform users would be inclined to read such detailed information. At first glance, it is therefore surprising that Art. 27(2)(b) DSA additionally requires that the "reasons for the relative importance of those parameters" must also be stated. In this respect, the DSA even goes beyond the transparency requirements of the UCPD and CRD, which only require an indication of "relative importance", but not an indication of any "reasons" for the weighting chosen by the platform provider. Perhaps the true character of Art. 27 DSA becomes clear if one assumes that the information about the functioning of the recommender system is not intended to be read and evaluated by the individual platform users. The information in the terms and conditions may rather serve for documentation purposes and as a starting point for investigations by the competent authorities or further research by civil society organizations. Such a reading of Art. 27 DSA would be in line with the additional transparency requirements in the DSA directed at public authorities. For example, according to Art. 40(3) DSA the competent national Digital Service Coordinator or the Commission may ask the platform provider to "explain the design, the logic, the functioning and the testing of their algorithmic systems, including their recommender systems". Such explanations offered by the platform providers could then be compared with the information on "main parameters" provided under Art. 27(1) Draft DSA.

One question that still needs further clarification is how Art. 27 DSA relates to the other transparency requirements from the UCPD, the CRD, and the P2B Regulation. On the surface, Art. 2(4)(e) and (f) DSA seems to provide a simple answer to this question by stating that the DSA is "without prejudice" to the P2B Regulation and Union law on consumer protection. What is less clear, however, is what the formula "without prejudice" means in this context. Does it mean that UCPD, CRD and P2B take precedence over the horizontal DSA as vertical *leges speciales*? But does this also apply where certain topics – such as algorithmic choice – are not being addressed in the *leges speciales*? In other words, since the UCPD and the CRD do not contain separate rules on "recommender switchboards", would it be conceivable that Art. 27(3) DSA also applies to online marketplaces covered by the UCPD and the CRD? Such a "combined" application of the different rules would mean that the content of the information on "main parameters" would be governed by UCPD and CRD, but the user interface design of the "recommender switchboard" would be governed by DSA. These considerations show that there is still a considerable need for coordination in the increasingly complex regulatory landscape for recommender systems.

[46] During the trilogue negotiations, the EU Council had suggested to require VLOPs to "make this information directly and easily available on the specific section of the online interface where the information is being prioritized according to the recommender system", but this proposal was not taken adopted for the provisional agreement, see EU Council, General Approach, p. 161.

3.4.2 User Control Over Ranking Criteria

The DSA does not limit itself to extending the transparency requirements for recommender systems but makes a first tentative step towards a new regulatory model that seeks to enable *algorithmic choice* for platform users. In this sense, the Commission Proposal for the DSA stipulated that providers of VLOPs should inform their users about "any options for the recipients to modify or influence those parameters that they may have made available, including at least one option which is not based on profiling"[47] within the meaning of Art. 4(4) GPDR. However, in order to provide users with effective control over the functioning of recommender systems, information about available options is not sufficient. It must also be possible for users to change the parameters easily. Whether the available options are used in practice very much depends on the user interface design. Therefore, the Commission's proposal for the DSA required providers of VLOPs which offer several options to modify or influence ranking parameters to "provide an easily accessible functionality on their online interface allowing the recipient of the service to select and to modify at any time their preferred option for each of the recommender systems that determines the relative order of information presented to them".[48] From a practical perspective, this meant that providers of VLOPs would have to offer a sort of "recommender switchboard" (or control panel) that allows users to modify the functioning of the recommender system. Many platforms already today offer a number of options for adjusting the ranking criteria on a voluntary basis.

During the trilogue negotiations, the scope of the provision on algorithmic choice was partially extended beyond VLOPs. In particular, the duty to inform about any options to modify or influence the parameters of the recommender system has been extended to all online platforms.[49] Similarly, the duty to provide an easily accessible functionality to change the parameters (if such an option is provided) has also been extended to all online platforms.[50] However, the duty to provide a "profiling-free" ranking as an option still applies only VLOPs.[51] Therefore, the transition from a transparency-based model to a choice-based model remains rather limited.

It is doubtful, whether such a limited approach to algorithmic choice is sufficient to ensure autonomy and informed choice.[52] While the DSA gives users a certain degree of control over the functioning of recommender systems, the proposed regulations still leave it in the hands of the platform provider to decide which modifications of the recommender system are made available. The only real choice that must be made available by VLOPs to their users is a "profiling-free" ranking. Other user

[47] Art. 29(1) DSA Commission Proposal.

[48] Art. 29(2) DSA Commission Proposal.

[49] Art. 27(1) DSA.

[50] Art. 27(3) DSA.

[51] Art. 38 DSA.

[52] See also Leerssen 2020b, criticizing that the DSA is based on a rather narrow understanding of user choice as a matter of selecting algorithmic weightings.

preferences, such as a more prominent display based on environmental or social ranking criteria, do not need to be offered. Moreover, as mentioned before, the effectiveness of the "recommender switchboard" very much depends on the user interface design. It must be ensured that platform operators do not attempt to influence and impair the selection of ranking parameters in an unfair manner by using manipulative design choices or "dark patterns". In this sense, the EU Council rightly suggested during the trilogue negotiations to add a provision that explicitly stipulates that providers of VLOPs "shall not seek to subvert or impair the autonomy, decision-making, or choice of the recipient of the service through the design, structure, function or manner of operating of their online interface" when presenting options regarding the functioning of the recommender system.[53] However, this proposal did not make it into the final text of the DSA, which only mentions "dark patterns" in a more general context in Recital 67. An important contribution to the development of a uniform and user-friendly "recommender switchboard" could be provided by voluntary standards, which are to be developed by the relevant European and international standardization bodies (CEN, ISO) at the suggestion of the Commission.[54]

3.5 Third Party Recommender Systems: Towards a Market for "RecommenderTech"

While the DSA only takes rather hesitant steps towards algorithmic choice, a more radical solution could have been possible. Instead of leaving the range of available options in the hands of the platforms, the DSA could have allowed platform users to choose between different competing third-party recommender systems. Such a solution would give users more control over what information they see on digital platforms. A practical proposal to this effect was recently put forward by a group of scholars led by Stanford political scientist Francis Fukuyama (Fukuyama et al. 2020). In essence, they propose that users should be able to select an alternative recommender system, that works as an external filtering device on top of a given platform. Such a third-party software, which Fukuyama and his co-authors refer to as "middleware", would interact with the data provided by the platform via an application programming interface (API). With the help of middleware, users of an online retail marketplace could decide to choose a filter that displays only products that are environmentally friendly or from producers that comply with high social standards (Fukuyama 2021).

A similar approach has also been put forward in a proposal for an amendment to the DSA, which was tabled in the IMCO Committee of the European Parliament in

[53] EU Council, General Approach, p. 161.

[54] Art. 44(1)(i) DSA.

July 2021. The proposal suggested adding the following passage to Art. 29 of the Commission's proposal for the DSA:

> In addition to the obligations to all online platforms, very large online platforms shall offer the recipients of the service the choice of using recommender systems from third party providers, where available. Such third parties must be offered access to the same operating system, hardware or software features that are available or used in the provision by the platform of its own recommender system.[55]

This proposal effectively aimed at unbundling content hosting and content curation by introducing an interoperability requirement that would open up online platforms for third-party recommender systems. The underlying idea was to increase consumer choice and competition by creating a market for third-party recommender systems. Similar solutions have been applied successfully in other fields where interoperability has helped to boost innovation with regard to complementary products and services.[56] One prominent example that could serve as a model also for recommender systems is the unbundling of accounts from financial services ("open banking"). Here mandatory interoperability has created a market for third-party payment service providers such as payment initiation providers (e.g. iDeal, Sofort Überweisung, Trustly) or account information providers (e.g. Zuper, Outbank, Numbrs).

Similar to the blossoming market for "FinTech" businesses, a market for providers of "RecommenderTech" – or "middleware" in the terminology suggested by Francis Fukuyama – could be created. It is unclear, however, whether providers of "middleware" will emerge that are able and willing to offer alternative recommender systems that counterbalance the economic dominance of major online platforms (Ghosh and Srinivasan 2021).

It will be necessary to explore how such an unbundling of platform and recommender systems could be technically feasible. In particular, issues of security and privacy have to be solved. As a general rule, interoperable systems with a higher level of interconnectedness may lead to higher risks regarding reliability and security (Keber and Schweitzer 2017). The crucial question, however, might be whether third-party recommender systems will be economically viable and what are possible business models (Keller 2021). Should such a product be funded via advertising or based on a subscription model? How much would consumers be willing to pay for fair and unbiased rankings? More fundamentally, is it right that only consumers who can afford to pay for "RecommenderTech" solutions benefit from fair and independent rankings, while financially vulnerable consumers must continue to use the (potentially biased) bundle of hosting and ranking?

[55] European Parliament, IMCO Committee, Draft Report on the Digital Services Act (PE693.594v01-00), 8 July 2021, Amendment 1703.

[56] See Kerber and Schweitzer 2017, explaining that greater interoperability may lead to more innovation and competition with regard to complementary products, but could lead to less innovation and competition with regard to the standards and interfaces themselves. See also Bourreau and Buiten 2022.

While the IMCO proposal cited above did not make its way into the final version of the DSA, its twin, the Digital Markets Act (DMA), could open the door for providers of "Recommender Tech". Art. 6(1)(f) of the Commission's proposal for the DMA stipulated that providers of core platform services who have been designated as gatekeepers by the European Commission shall "allow business users and providers of ancillary services access to and interoperability with the same operating system, hardware or software features that are available or used in the provision by the gatekeeper of any ancillary services". In this context, the term "ancillary service" means "services provided in the context of or together with the core platform services".[57] The Commission proposal explicitly mentions payment services, fulfillment, identification, and advertising services as examples. One could well imagine that third-party recommender services also fall under the term ancillary services and are thus covered by the mandatory interoperability requirement.

Following the provisional agreement on the DMA, reached in March 2022, the wording of the provision has changed and the term "ancillary services" is no longer used (see Gerpott 2022, providing an overview of the DMA based on the provisional trilogue agreement). In substance, however, the interoperability obligation was retained. The final version of the provision now requires gatekeepers to "allow business users and alternative providers of services provided together with, or in support of, core platform services, free of charge, effective interoperability" with the gatekeeper's operating system, hardware, or software features.[58] It seems that the DMA provisions on interoperability for alternative providers' ancillary services have opened the door (at least a little bit) for third-party recommender systems. Whether this door leads to an attractive market for providers of "RecommenderTech" and whether it will be possible to develop viable business models in this field that will experience growth comparable to that of the "FinTech" ecosystem remains to be seen.

3.6 Conclusion

The European regulatory framework for algorithmic rankings and recommendations in the platform economy has developed rapidly in a short period of time. Until very recently, there were no specific rules for recommender systems at the European level. With the entry into force of the P2B Regulation and the DSA and the recent reform of UCPD and CRD, this situation has changed. Instead of too few, there are now maybe too many regulations that are not sufficiently coordinated. On a more fundamental level, one may ask whether transparency requirements alone are sufficient to ensure unbiased recommendations and consumer autonomy. Against this background, it is to be welcomed that the DSA takes a first step (albeit hesitantly

[57] Art. 2(11) DMA Commission Proposal.

[58] Art. 6(7) DMA.

and incompletely) from a regulatory model based on algorithmic transparency towards a new regulatory model based on algorithmic choice. In the medium term, the DMA could even have a greater impact if it succeeds in creating a market for "RecommenderTech". In such a scenario, third-party recommender systems could offer consumers a real alternative to the rankings and recommendations currently provided by large platforms. While the economic viability and technical feasibility of such a decentralized regulatory model are not yet entirely clear, a choice-based approach to recommender governance may indeed be the wave of the future.

References

Aggarwal, C.C. 2016. *Recommender Systems*. Cham: Springer. https://doi.org/10.1007/978-3-319-29659-3.

Airbnb Ireland. 2022. How search results work. https://www.airbnb.ie/help/article/39/how-search-results-work. Accessed 13 Feb 2022.

Airbnb UK. 2022. How Airbnb search works. https://www.airbnb.co.uk/resources/hosting-homes/a/how-airbnb-search-works-44. Accessed 13 Feb 2022.

Alexander, C. 2019. Neue Transparenzanforderungen im Internet – Ergänzungen der UGP-RL durch den "New Deal for Consumers". *Wettbewerb in Recht und Praxis* 10: 1235–1241.

Banker, S., and S. Khetani. 2019. Algorithm Overdependence: How the Use of Algorithmic Recommendation Systems Can Increase Risks to Consumer Well-Being. *Journal of Public Policy & Marketing* 38 (4): 500–515.

Bourreau, M., and M. Buiten. 2022. Interoperability in digital markets. CERRE report. https://cerre.eu/wp-content/uploads/2022/03/220321_CERRE_Report_Interoperability-in-Digital-Markets_FINAL.pdf

Bundeskartellamt. 2019. Sektoruntersuchung Vergleichsportale, Abschlussbericht. https://www.bundeskartellamt.de/SharedDocs/Publikation/DE/Sektoruntersuchungen/Sektoruntersuchung_Vergleichsportale_Bericht.pdf

Cobbe, J., and J. Singh. 2019. Regulating Recommending: Motivations, Considerations, and Principles. *European Journal of Law and Technology* 10 (3) http://www.ejlt.org/index.php/ejlt/article/view/686.

Cohn, J. 2019. *The Burden of Choice: Recommendations, Subversion, and Algorithmic Culture*. New Brunswick: Rutgers University Press.

De Los Santos, B., and S. Koulayev. 2017. Optimizing Click-Through in Online Rankings with Endogenous Search Refinement. *Marketing Science* 36 (4): 542–564.

European Commission. 2016a. Guidance on the implementation and application of Directive 2005/29/EC on Unfair Commercial Practices, SWD(2016) 163 final. https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:52016SC0163

———. 2016b. Key principles for comparison tools. https://ec.europa.eu/info/sites/default/files/key_principles_for_comparison_tools_en.pdf

Fukuyama, F. 2021. Making the Internet Safe for Democracy. *Journal of Democracy* 32 (2): 37–44.

Fukuyama, F., B. Richman, A. Goel, R.R. Katz, D. Melamed, and M. Schaake. 2020. Middleware for dominant digital platforms: A technological solution to a threat to democracy. Stanford Cyber Policy Center White Paper. https://fsi-live.s3.us-west-1.amazonaws.com/s3fs-public/cpc-middleware_ff_v2.pdf

Gerpott, T.J. 2022. Das Gesetz über digitale Märkte nach den Trilog-Verhandlungen. *Computer und Recht* 38 (6): 409–416.

Ghosh, D., and R. Srinivasan. 2021. The Future of Platform Power: Reining In Big Tech. *Journal of Democracy* 32 (3): 163–167.

Grochowski, M., A. Jabłonowska, F. Lagioia, and G. Sartor. 2021. Algorithmic Transparency and Explainability for EU Consumer Protection: Unwrapping the Regulatory Premises. *Critical Analysis of Law* 8 (1): 43–63.

Häubl, G., and V. Trifts. 2000. Consumer Decision Making in Online Shopping Environments: The Effects of Interactive Decision Aids. *Marketing Science* 19 (1): 4–21.

Helberger, N., K. Karppinen, and L. D'Acunto. 2018. Exposure Diversity as a Design Principle for Recommender Systems. *Information, Communication & Society* 21 (2): 191–207.

Hildebrandt, M. 2022. The Issue of Proxies and Choice Architectures: Why EU Law Matters for Recommender Systems. *Frontiers in Artificial Intelligence* 5 (2022): 789076.

Jannach, D., and G. Adomcavicius. 2016. Recommendations with a Purpose. In *Proceedings of the 10th ACM Conference on Recommender Systems*, 7–10. New York: Association for Computing Machinery. https://doi.org/10.1145/2959100.2959186.

Kant, T. 2020. *Making it Personal: Algorithmic Personalization, Identity, and Everyday Life*. New York: Oxford University Press.

Keller, D. 2021. The Future of Platform Power: Making Middleware Work. *Journal of Democracy* 32 (3): 168–172.

Kerber, W., and H. Schweitzer. 2017. Interoperability in the Digital Economy. *Journal of Intellectual Property, Information Technology and E-Commerce Law* 8 (1): 39–58.

Leerssen, P. 2020a. The Soapbox as a Blackbox: Regulating Transparency in Social Media Recommender Systems. *European Journal of Law and Technology* 11 (2) http://www.ejlt.org/index.php/ejlt/article/view/786.

———. 2020b. Algorithmic Centrism in the DSA's Regulation of Recommender Systems. *Verfassungsblog*,March29,2020.https://verfassungsblog.de/roa-algorithm-centrism-in-the-dsa/

Milano, S., M. Taddeo, and L. Floridi. 2020. Recommender Systems and Their Ethical Challenges. *AI & Society* 35: 957–967. https://doi.org/10.1007/s00146-020-00950-y.

Pasquale, F. 2015. *The Black Box Society*. Cambridge: Harvard University Press.

Peifer, N. 2021. Die neuen Transparenzregeln im UWG (Bewertungen, Rankings und Influencer). *Gewerblicher Rechtsschutz und Urheberrecht* 123 (12): 1453–1461.

Ricci, F., L. Rokach, and B. Shapira. 2022. Recommender Systems: Introduction and Challenges. In *Recommender Systems Handbook*, ed. F. Ricci, L. Rokach, and B. Shapira, 1–35. Boston: Springer. https://doi.org/10.1007/978-1-4899-7637-6_1.

Schebesta, H., and K. Purnhagen. 2020. Island or Ocean? Empirical Evidence on the Average Consumer Concept in the UCPD. *European Review of Private Law* 28 (2): 293–310. https://doi.org/10.54648/erpl2020015.

Stoyanovich, J., K. Yang, A. Asudeh, B. Howe, H.V. Jagadish, and G. Miklau. 2018. A Nutritional Label for Rankings. In *Proceedings of the ACM SIGMOD'18 Conference*, 1773–1776. New York: Association for Computing Machinery. https://doi.org/10.1145/3183713.3193568.

Svirskis, A. 2020. Instagram Shopping – A New Dawn for Social Commerce. *Forbes,* May 26, 2020. https://www.forbes.com/sites/anthonysvirskis/2020/05/26/instagram-shoppinga-new-dawn-for-social-commerce

United Kingdom. UK Competition and Markets Authority. 2021. Algorithms: How they can reduce competition and harm consumers. https://www.gov.uk/government/publications/algorithms-how-they-can-reduce-competition-and-harm-consumers/algorithms-how-they-can-reduce-competition-and-harm-consumers

Ursu, R. 2018. The Power of Rankings: Quantifying the Effect of Rankings on Online Consumer Search and Purchase Decisions. *Marketing Science* 37 (4): 530–552.

Weber, F. 2020. Das Verbraucherleitbild des Verbrauchervertragsrechts – im Wandel? *Verbraucher und Recht* 35 (1): 9–15.

Chapter 4
Black Hole Instead of Black Box?: The Double Opaqueness of Recommender Systems on Gaming Platforms and Its Legal Implications

Dagmar Gesmann-Nuissl and Stefanie Meyer

Abstract Recommender systems that support us in our everyday lives are becoming more precise and accurate in terms of the appropriateness of recommendations to users' needs – with the result that the user often follows these recommendations. This is mainly due to the filtering methods and various algorithms used. In our paper, we will look specifically at the recommender systems on gaming platforms. These consist of different components: a shopping component, a streaming component and a social media component. The recommender systems of these components, when considered individually, have certain characteristics in terms of the machine learning and filtering methods used, which are mixed by combining them on one platform. As a result, it is unclear which of the information collected about the user at any time is lost and disappears into obscurity, and what information is used to generate recommendations. The frequently discussed "black box" problem exacerbates at this point and becomes a "black hole." With the interests of platform users, platform operators, and software developers in mind, we examine the legal provisions that have been established to address this opaqueness: transparency obligations. Derived from the Digital Services Act and the Artificial Intelligence Act, we present various legally valid solutions to address the "black hole" problem and also lead them to practical suggestions for implementation.

Keywords Multidimensional decision algorithms · Transparency · Digital Services Act · Artificial Intelligence Act · Hybrid Societies

D. Gesmann-Nuissl (✉) · S. Meyer
Chemnitz University of Technology, Chemnitz, Germany
e-mail: dagmar.gesmann@wiwi.tu-chemnitz.de; stefanie.meyer@wiwi.tu-chemnitz.de

S. Genovesi et al. (eds.), *Recommender Systems: Legal and Ethical Issues*, The International Library of Ethics, Law and Technology 40,
https://doi.org/10.1007/978-3-031-34804-4_4

4.1 Introduction

It was a few years ago – in March 2020 – when the gaming platform *Steam* (© 2022 Valve Corporation) promoted its new *Steam* recommendation service (Steam 2020). It featured interactivity and was advertised as an absolute novelty. Users receive recommendations based on the games they have already played – a pool of potentially interesting games is generated and selected based on other players who have similar interests to the user. In addition, other recommendations can also be made: based on the games that friends play or by hints from curators. What exactly is new about this recommendation system only becomes clear with a look behind the screen, at the special filtering methods of the recommendation system and the algorithms used.

The idea of recommender systems in general is not new: The first recommender system was developed in 1992, but had little practical application due to insufficient computer processing power and limited data sources (Gahier and Gujral 2021). With the availability of higher-quality technologies that can process large amounts of data and the digitization of society, the recommender system has now spread to many areas of daily life: They have established themselves as robo-advisors in securities trading, accompany selection processes in human resources management and manage investments in media products (Linardatos 2020; Maume, 2021; Isaias et al. 2010; Barreau 2020; Fleder et al. 2010). Especially in the latter application area, the system is supposed to make a prediction about how strong a user's interest in a (virtual) product is in order to recommend to the user exactly those products from the set of all available products that are likely to interest them the most (Mahesh and Vivek 2021). However, both the mass of people using the offerings and the number of objects to be recommended have increased in recent years; moreover, the interests of the providers of objects join the user interests: They want to be seen (Goanta and Spanakis 2020). In the field of computer games, especially the offerings of smaller game developers (so-called indie game developers) are brought to the screen of many users by the recommendation algorithms. The problem that arises due to the multitude of products and information, due to different filters and ambitions, is the prototypical "black box" (or sometimes "white box") problem.

4.2 The Black Box-Problem of AI Applications

An overarching goal of recommender systems is to make a prediction that quantifies how strong a user's interest in an object is, in order to recommend to the user exactly those objects from the set of all available objects in which the user is most likely to be interested in (Ziegler and Loepp, 2019). The quality of a recommender system as perceived by the user depends not only on the predictive quality of the algorithms, but also to a large extent on the usability of the system (Knijnenburg et al. 2012). Therefore, to determine the appropriate recommendations, the service uses machine learning and information retrieval methods (Zednik 2021; Mohanty et al. 2020; Silva 2019). Although users are usually interested in having a movie, product, or service recommended to them that is tailored to their interests (Schmidt et al. 2018), so they don't have to search for it on their own and search out their own preferences from the almost infinite number of objects, they are usually not aware of why AI systems make decisions or engage in certain behaviors. In fact, most of these procedures lead to "black box" phenomena (Zafar et al. 2017): That is, knowledge – a model – is established through machine learning processes, but it is not explainable or comprehensible to the users of the systems, or at least only with great difficulty (Niederée and Nejdl 2020). Opacity can affect users' trust in the system and lead to rejection of the systems, especially in contexts where the consequences are significant (Raj 2020; Burrell 2016; Ribeiro et al. 2016). For example, if AI is used in medicine and supports the attending physician in evaluating CT or MRI scans, the algorithm learns and can analyze the data faster than a human, flag tumors and suggest results of a therapy. However, the positive effect of constantly and rapidly growing knowledge and accurate classification of symptoms in real time (especially in cancer detection) also raises issues of trust. It remains opaque why AI can distinguish harmless cysts from malignant cancer – at least from the user's perspective (which may include both the patient and the treating physician). Due to this opacity of the detection process and the risk of serious health consequences if misdiagnosed, patients do not trust AI. Instead, they prefer to trust the professional opinion of a human physician and their assessment of the need for therapy. Nonetheless, opacity is always agent-dependent, meaning that a computer system is not opaque in and of itself, but in relation to the actor using it (Humphreys 2008). The developer of an algorithm can understand its operation better than a user. The degree of opacity also depends on what kind of algorithms are used to generate the output. There are several possibilities of using search algorithms, such as linear models (i.e., logistic regressions), generalized additive models (i.e., GAM), decision trees, clustering (i.e. key nearest neighbors), kernel based methods (i.e., support vector machine), ensemble methods (i.e., random forest, XGBoost), neural network (i.e., CNN, RNN) (Niederée and Nejdl 2020).[1] The manifold possibilities of algorithms are also legally relevant and are put into a legislative context, such as in Annex I of the Artificial Intelligence Act (COM/2021/206 final).

4.2.1 Transparency and Explainability: An Introduction

In order to approach the "black hole" problem of gaming platforms, it is necessary to explain the "black box" phenomenon. In this respect, a distinction has to be made already between the opacity of the recommender systems on the one hand and the general opacity of AI systems on the other hand, each of which has

[1] A detailed description of the methods can be found, for example, at: Abdullah et al. 2021.

a different dimension. For this purpose, transparency requirements and explanation intentions have to be set in relation to the current technical situation. It is obvious that the demand for transparency for all affected groups of people is not appreciated from every technical perspective. Nevertheless, the debate is important if the AI system is designed for interactivity – if users are allowed and expected to participate responsibly, they also need to have a basic understanding of the process.

4.2.2 Efficiency vs. Explainability of Machine Learning

The various applicable algorithms, are of varying effectiveness and also have varying degrees of transparency. While linear models (such as rule-based systems) or decision trees can be explained comparatively well (as Fig. 4.1 shows), the accuracy potential of these approaches is comparatively lower, neural networks are potentially more accurate. Of course, this depends on the application context – but in general this can be demonstrated in research (Körner 2020; Abdullah et al. 2021). There are several reasons: (1) Expressiveness: Similar algorithms can be used for an increasing number of domains and problems. For example, certain neural network architectures can be used for prediction, autonomous driving, pharmaceutical research, and particle physics alike. (2) Versatility: This allows different types of data to be used together and even multimodal approaches where different types of data are processed simultaneously. (3) Adaptability: Some of the approaches can be

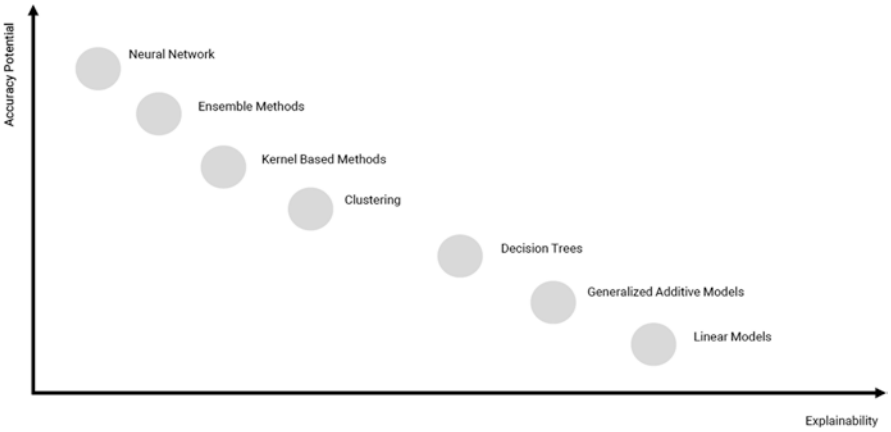


Fig. 4.1 Different approaches to AI, as measured by their explanatory power and accuracy potential

transferred with little effort.[2] (4) Efficiency: Special hardware has enabled corresponding models to be trained faster and more efficiently (Körner 2020).

However, these more advanced and precise methods of machine learning in general and neural networks in particular are less comprehensible than the simpler forms. The more simple approaches, such as the methods of rule-based systems or decision trees, can be explained in principle. For example, if a simple algorithm such as a decision tree is built into an autonomous vehicle that recognizes that you need to stop at a red light and reports this accordingly, most people can understand that. Sometimes such methods are referred to as a "white box", although they are of course not comprehensible in detail to everyone – (for this reason Pasquale still refers to a "black box" in terms of the user perspective; Pasquale 2015).[3] Theoretically, the input and output data would usually be known to the user. In addition, the systems are comprehensible to the user to a certain extent due to their internal knowledge structure and the rules used for decision-making (see Fig. 4.2). Moreover, this applies regardless of the expertise of those on whom the system acts

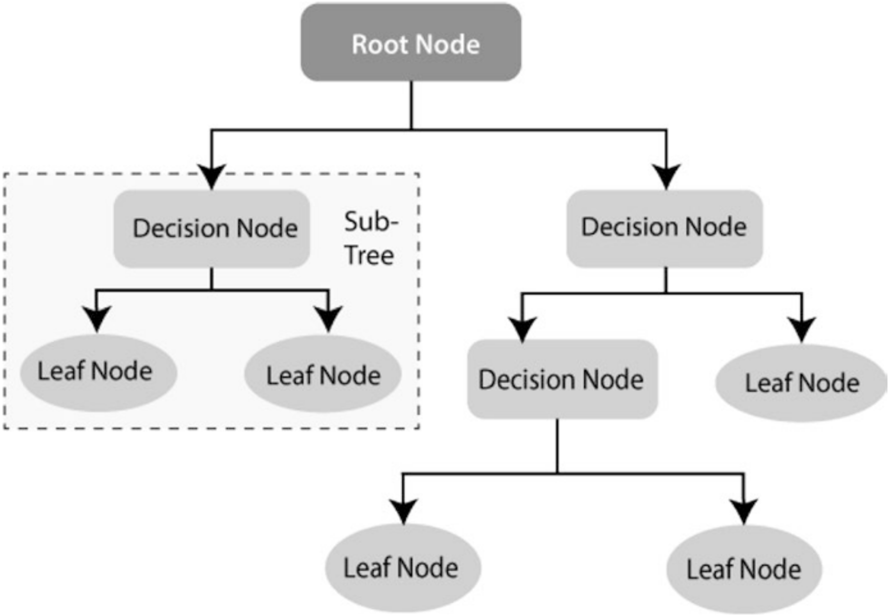


Fig. 4.2 Example of a decision tree structure. (Sarker 2021)

[2] However, the assertion of adaptability needs to be considered with caution, as it has been established by researchers (such as Körner in: Kaulartz and Braegelmann 2020) in relation to the application of an AI under regular conditions. Extreme situations (such as the Corona pandemic spreading globally) and the subsequent adapted behavior of the general population have shown that adaptability is not always provided under exceptional circumstances (Nielson and Killeen 2022, Sousa and Barrata 2021).

[3] This term is not always used consistently, so its significance often remains unclear.

(Niederée and Nejdl 2020). With respect to neural networks, this is referred to as the “black box” problem. They are opaque and hardly explainable to the user, since a multitude of different paths is conceivable for the algorithms during decision making (see Fig. 4.3). However, this can lead to remarkable problems. Neural networks are being used in an evolving range of domains, and the resulting decisions and assessments are increasingly impacting critical areas relevant to the lives of the people involved, such as in medicine. Better understanding and explaining the results of machine learning have several benefits (Holzinger 2018). For example, it is interesting to know on which data the AI system’s decision is based – how reliable are they and of what quality? Also, how exactly the patient data (which one?) was matched with the training data. This would make it possible to check and evaluate machine decision proposals and assessments for their credibility. While symbolic systems can be examined line by line, instruction by instruction, in neural networks the symbolic representation of the knowledge and the start-up control disappear. The knowledge and behavior stored in the neural network can now only be inferred indirectly through experimentation (Ebers 2020). There are several reasons for this: (1) The strength of these types of networks is their ability to learn. Given a training data set that contains the correct answers, they can gradually improve their performance by optimizing the strength of each connection until even their top-level outputs are correct. This process, which simulates how the brain learns by strengthening or weakening synapses, eventually leads to a network that can successfully classify new data that was not part of its training set. Thus, they are not limited to human perceptual and communication patterns. This type of learning is partly why they are so powerful, but also why the information in the network is so diffuse: Similar to the brain, memory is encoded in the strength of multiple connections rather than stored in specific locations as in a traditional database (Castelvecchi 2016; Robbins 2019). (2) Furthermore, another property of deep neural networks is that they can also learn the features they use to learn for themselves; however, this extends the “black box” problem to the features they use and further complicates explainability (Niederée and Nejdl 2020).

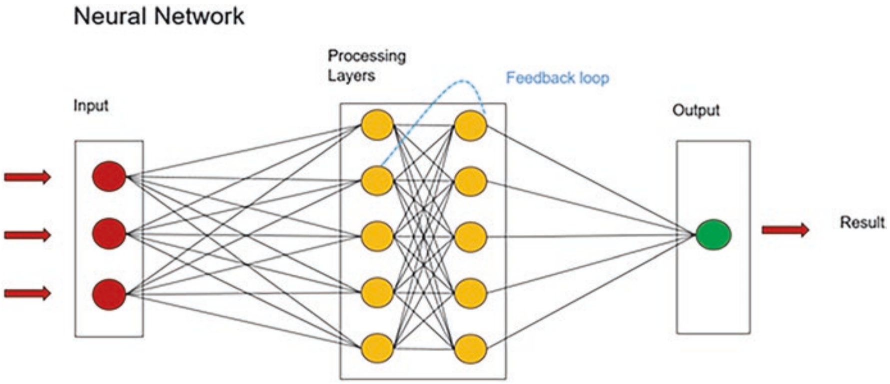


Fig. 4.3 Basic neural network layout. (Uzair and Jamil 2020)

4.2.3 Background of the Transparency Requirement

Regulators are increasingly focusing on the objective of transparency of machine learning systems in general. With the draft of the European Artificial Intelligence Act published in April 2021 (COM/2021/206 final), the requirement of transparency and explainability was taken up again and readdressed in comprehensive regulations. Lipton (2016) has devised a taxonomy of methods and approaches that can be categorized within the realm of explainability and can facilitate a basic understanding of the terms used here. This can often lead to a linguistic imprecision regarding the meaning of explainability on the one hand and transparency on the other. According to Lipton, explainability is the overarching term for the two concepts of transparency and interpretability. The concept of transparency focuses strongly on the technology and algorithms involved, while the concept of interpretability is less technological and more likely to be found in specific contexts (Waltl 2019). Here, the focus is on human perception. Thus, when we speak of explainability in this context, it includes both transparency of the technical components and interpretability by the individuals using the system (Lipton 2016). In terms of the individuals involved, i.e., the addressees of transparency obligations, transparency becomes relevant at different levels (Anand et al. 2018): (1) Software developers and vendors need to understand how the concerned system works in order to fix any bugs and improve the system (Hohman et al. 2018). (2) Individuals affected by an algorithmic decision want to know and understand why the system reached a particular judgment, as this is the only way to detect any errors (in decision-making, in the basis for the decision, or in the evaluation of the decision). (3) Transparency allows legislators, regulators, certifiers, experts, courts or other neutral parties to assess the fundamental process and technical products (Rieder and Simon 2017; Ebers 2020). In addition, the technical consideration of the individual process steps is also important. “Procedure” in this context means procedures of automated decision making – so-called automated decision-making processes. In this procedure, the transparency of the processes becomes relevant on three levels: the process level, the model level and the classification level. The process level refers to various steps that an AI system needs to go through for training. This process usually includes five steps that immediately follow each other: Data acquisition, data preparation for the purpose of correcting incomplete or erroneous data, data transformation to unify them, training the AI model by optimizing mathematical functions and approximating the training data, and post-processing them (Ebers 2020). In terms of the requirement for transparency, it is important to know and understand each of these individual steps in order to understand the associated algorithmic decision-making process. The model level refers to the different types of machine learning methods used to make decisions (Ebers 2020). Analyzing this level is important because the different models have different levels of transparency. The classification level provides information about which attributes or features are used in the model and the weighting given to each attribute (Ebers 2020).

If transparency is to be explored in relation to a particular system and in relation to users, four questions should be asked: What is happening? Why is it happening? How does it work? and where does the process take place? (Zednik 2021; Tomsett et al. 2018; Marr 1982). Those who review, approve, and certify systems to ultimately determine functional safety need to know what is about to happen. They have to ask about what a particular system does – they are interested in what the system as a whole does and how algorithmic decision making occurs (Zednik 2021). Those whose data are processed and learned in a decision-making process, as well as those who are told that a decision is being made, want to know why it is happening. Those who use an algorithmic system for decision making (e.g., bankers who assess creditworthiness using an automated system) need to know why a system does what it does. For them, the interpretation of the behavior in relation to the specific facts of the case is relevant (Zednik 2021). Those who review, approve, and certify systems should know both why something happens and what happens. They and those who design, improve, and maintain the systems have to ask, at the algorithmic level, how a system works; at the implementation level, they have to ask how the program tries to realize the algorithms (Marr 1982). Thus, when approaching the question of the right level of transparency and finding the right way to inform and educate, these different interests need to be considered in relation to the different points of comprehensibility. Although the opacity of computer systems programmed with machine learning has traditionally been seen as the "black box" problem, in this sense it is perhaps more appropriate to speak of many "black box" problems. Depending on the perspective and the nature of the interaction with the machine learning program, the program will be opaque for different reasons and will need to be made transparent in different ways (Zednik 2021; see for a detailed analysis: Burrell 2016).

4.2.4 *Criticism*

Nevertheless, there are critical voices regarding the designation of AI-driven systems as "black boxes" and the associated demand for explanations and transparency (see Bryson 2019; in a normative sense: Robbins 2019). The demand for transparency is dismissed by arguing that an explanation of "how" a decision is reached is not helpful to a user, since the explanation of how the algorithmic decision is reached is difficult to understand anyway (Anderson 1972). It is necessary, but also sufficient, for those who program and use AI software to keep detailed records of how it works. They just need to ensure that appropriate care is taken (Bryson 2019).

However, that is not the point. Even if the algorithm has been learned thoroughly and, in this sense, complies with the normative requirements, the results it generates in adaptive neural networks such as recommender systems are not always predictable. In particular, when multidimensional learning data is the basis, new and different information that tends to be added randomly in application practice can be incorporated into the learning data that eventually affects the result in a lasting way that is unpredictable for the programmer (Zech 2019). This danger applies all the more if, as in the present case, the users participate in determining the decision parameters.

This goes along with the legally necessary requirements for the safety of products. For example, according to the New Legislative Framework system applicable in Europe (consists of Regulation (EU) 765/2008, Decision 768/2008 and Regulation (EU) 2019/1020.), safety-relevant products may only be brought onto the market if they have been tested and assessed as sufficiently safe by the manufacturer; beyond this, there is always an assessment of the legitimacy of such systems (Zech 2019). However, such an assessment requires prior understanding of the system used. Generally, this includes explaining the machine as such, but if AI is implemented, i.e., a tool or algorithm that controls the outcome of the machine, it also includes explaining those algorithms. Statements made by the manufacturer without knowledge of the system cannot be tolerated in the case of safety-relevant systems. However, the requirement of product safety certainly does not end with the mere development process up to placing the product on the market. The development of a normatively traceable product may relieve the developer, but it doesn't indicate anything regarding the further consequences that may arise for third parties and the environment. Even if the developer of a system has operated in accordance with the regulations, this does by no means release them from further monitoring, e.g. of the consequences. This may, for example, be an expression of the product monitoring obligation as provided for in tort law (Ensthaler et al. 2012). Furthermore, explainability is not only important for the person placing the system on the market, but also for market surveillance. There, too, the processes need to be comprehensible if they want to check, among other things, whether the functional safety of the systems is guaranteed, which can be punished with fines. In this respect, explainability is essential not only at the level of the programmer, but also at the level of all other market participants, which is why the current EU legislations (COM/2021/206 final, Regulation (EU) 2022/2065) gives high priority to transparency of machine systems. Due diligence of users of AI systems is an important part of safety, but explainability is an additional – and above all essential – part. Thus, the legislature has made it clear that self-regulation is not effective and is not sufficient to protect users.

4.2.5 In Terms of Recommender Systems

Compared to the one-dimensional AI systems in the form of neural networks described so far, the associated "black box" phenomenon is amplified in recommender systems. The weightings made by algorithms can additionally be affected by user signals – and conversely, the recommendation system also has the power to shape and control user preferences and habits (Leerssen 2020). For this reason, it is particularly important in this context to look beyond pure algorithms and understand the complex interactions between technology and users. Several questions arise: (1) What algorithms are used to generate recommendations, i.e., what filtering method is used? (2) What recommendations are made? (3) What user content, metadata, or behavioral data feeds into the system? And (4) What human actors or organizational structures are involved in the process that will never be seen?

Especially in the area of recommender systems, providers have an interest in being cautious about the algorithms and filters they use and in not making these methods transparently available to users. Platform operators provide various reasons: (1) The platforms argue that the design of recommender systems involves commercially valuable trade secrets and that they would suffer economic disadvantages by publishing the methods. (2) Keeping the algorithms secret may be necessary in some cases to prevent users from undermining the gatekeeping function (e.g., abuse by spamming if the keyword blacklist is published). (3) User privacy may be compromised to the extent that the algorithm is developed based on user data (Leerssen 2020). The last point in particular draws attention to the socio-technical perspective: The meaning of algorithms is highly context-dependent, as the results of the system are co-determined by user behavior. Therefore, it is argued that in terms of transparency, the output should be explained first and from there the further recommendation patterns should be considered (Rieder et al. 2018). This includes, for example, the type of recommendations as well as user content or metadata or behavioral data, etc. In short, the user should know which of his personal data is linked to which data in the algorithm and through which linkage of these two pillars the recommendation emerges. However, this poses a particular challenge because the output of the system is not generalizable. Thus, in the context of recommender systems, it is not only unclear why a decision is made, but also which decisions are made (Leerssen 2020). In recommender systems, not only the code or data needs to be kept transparent, but also human and non-human actors need to be involved (Ananny and Crawford 2018).

4.3 The Black Hole-Problem of Gaming Platforms

The reason why we speak of “black hole” instead of “black box” in relation to gaming platforms is because of the multidimensionality of the gaming platform. Steam’s interactive recommender system, described at the beginning of this paper, says that it is no longer affected by tags or reviews alone, but learns through the games (Steam 2020). It analyzes the games users play and compares them to the gaming habits of other users. Based on this, the selection is to be tailored even more precisely to the user – alongside the possibility that the users themselves can add parameters by selecting that they would also like to receive recommendations based on friend preferences and those of curators. Basically, gaming platforms behave no differently than other platforms (such as Amazon or YouTube, see Covington et al. 2016; Davidson et al. 2010) in the way they are perceived externally. The distinguishing feature, as will be mentioned below, is the combination of different types of algorithms used in these cases and the disregard of some data in certain contexts. The platform offers three different components from which it draws and evaluates its information: a shopping component, a streaming component and a social media component. In the store, users can buy the respective games and then find them in the library. From there, they can be downloaded to their device. In addition, it is possible to watch live streaming offers for some games, which are reminiscent of video or streaming platforms such as YouTube (©Google Ireland Limited) or Twitch (©Twitch Interactive, Inc.). It is also possible to network with friends, which is necessary to play cooperative games together; however, it is also possible to network independently and follow the activities. All of this information that the recommendation system can receive from the different types of platforms, it processes to generate the recommended output. Recommendation systems on shopping, streaming and social media platforms are not new territory – but linking what is known on a platform offers special challenges. Here, we are dealing with learning, evolving and, above all, interacting neural networks.

4.3.1 Types of Recommender Systems

Recommendation systems can basically be divided into two types, provided that filtering methods are used as a distinguishing criterion: content-based and collaborative filtering systems. These are extended by a third category, that of hybrid methods.

4.3.1.1 Content-Based Filtering Methods

Content-based filter systems recommend to the user those offerings that are similar to those that the user has preferred in the past (Adomavicius and Tuzhilin 2005). If these filtering methods are to be classified into the categories mentioned above, this applies to the area of video and streaming platforms. In most cases, the system here consists of two neural networks – one to generate a "pool" of possible recommended content and one to further assess and rank the individual content from this pool. This two-step approach allows recommendations to be made from a very large corpus of videos or streams, while being sure that the small portion of output that is eventually displayed to the individual user is personalized and appealing to the user.[4] The main task of ranking in the second step is to specialize and calibrate candidate predictions (Covington et al. 2016). The main advantage of using deep neural networks for candidate generation is that new interest categories and new

[4] Sometimes collaborative filtering methods are used for the first of the two stages - the details of the design cannot be discussed in this paper. For further information see: Deng et al. 2015, Wang et al. 2013, Covington et al. 2016, Davidson et al. 2010, Zhou et al. 2010, Rappaz et al. 2021, Hamilton et al. 2014.

appearances can be added continuously. As accurate as deep neural networks may be, they are not just opaque due to fitting errors, but also require explanation.[5]

The filtering method is primarily based on the comparison of articles and user information (Gahier and Gujral 2021). Based on this comparison, this method is further divided into user-centered and object-centered methods. On the one hand, the recommendations should reflect the user's behavior and activities, but on the other hand, in the interest of visibility among providers, a set of content unknown to the user should be displayed (Davidson et al. 2010). This content-based system has some drawbacks: (1) Recommendations are also limited by the associated features. (2) Two videos with the same features are difficult to distinguish from each other (and often the videos are divided into multiple parts). (3) Only similar items can be recommended (Adomavicius and Tuzhilin 2005). Moreover, these data are highly noisy. The problem encompasses a variety of circumstances: Video metadata may be nonexistent, incomplete, outdated, or simply wrong. User data often capture only a fraction of a user's activities and have limited ability to capture and measure engagement and satisfaction. The length of videos can affect the quality of recommendations derived from these videos (Davidson et al. 2010). In the context of live streaming, stream providers are additionally not available indefinitely (Rappaz et al. 2021). These circumstances may result in the recommendation being inaccurate after all.

4.3.1.2 Collaborative Filtering Methods

Collaborative filtering methods recommend objects in which users with similar evaluation behavior have the greatest interest (Adomavicius and Tuzhilin 2005). Here, no further knowledge about the object is required; the algorithm here refers to the user or acts element-based. The former algorithms operate memory-based, i.e., they make a rating prediction for the user based on the previous ratings. This prediction is evaluated as a weighted average of the ratings given by other users, where the weight is proportional to the similarity between the users. The model-based or element-based algorithms, on the other hand, attempt to model users based on their past interests and use these models to predict ratings for unseen items. These algorithms typically span multiple interests of users by classifying them into multiple clusters or classes (Das et al. 2007).

This method is commonly used in shopping platforms. Unlike the content-based filtering method, it does not necessarily rely on deep neural networks, but also works on the basis of linear models or symbolic AI. It can be guided by the user's implicit feedback, e.g., transaction or browsing history, or by explicit interactions, e.g., previous ratings. However, this also highlights the drawbacks of this method: (1) New users of a platform do not yet have a basis on which to identify their

[5] Compare, for example, the case where the algorithm is supposed to filter images of cats. However, the algorithm associates the property of the cat that it likes to be pictured with a ball of wool in such a way that it also displays images on which only a ball of wool is pictured as "cat".

interests, i.e., recommendations are not personalized. (2) New items are difficult to recommend if they have not been previously evaluated (Adomavicius and Tuzhilin 2005). However, this is inconsistent with the intentions of lesser-known vendors, such as indie game developers, described at the beginning of this paper. If there are no ratings for the unknown games, they are less likely to be recommended – but this is exactly why a recommendation system is used on game platforms, namely to not only focus on the big providers.

4.3.1.3 Hybrid Filtering Methods

Hybrid filtering methods first use the collaborative and content-based filtering methods just described separately and make predictions about user behavior based on each method. In a second step, the content-based features are included in the collaborative approach and the collaborative features are included in the content-based approach. In this way, the user of this method can create a general unified model that includes both content-based and collaborative features. These hybrid systems can also be augmented with knowledge-based techniques, such as case-based reasoning, to improve recommendation accuracy and solve some of the problems of traditional recommender systems (Adomavicius and Tuzhilin 2005).

Such a hybrid form of recommendation can be found in the field of social media platforms. Recommendations in these online networks differ from the previously mentioned platforms in that not only content recommendations, but also the social behavior of users is taken into account (Wang et al. 2013). Arguably, the most powerful influencing factor on social media platforms are the content recommendation systems that determine the ranking of content presented to users. These have a powerful gatekeeping function that is the subject of widespread public debate (Cobbe and Singh 2019). The system's recommendations appear on the start page, disguised among friends' posts – but the order of the news feed is determined by the ranking algorithms (Leerssen 2020). Two fundamentally different policies are followed: interest-based and influence-based recommendations. Interest-based recommendations aim to evaluate the relevance between a user and a piece of content, so that the content that is likely to interest the user the most is recommended (this is the focus of content-based recommendation). Influence-based recommendation examines what content is shared to maximize influence (this focuses on ideas of collaborative filtering) (Wang et al. 2013). Since the content recommender system mixes these two focuses as well as the underlying filtering methods, it is referred to as hybrid filtering. However, the problems and inaccuracies of each method described above exist here as well, so they may be exacerbated.

4.3.2 Black Hole Phenomenon

Gaming platforms combine the different platform types of shopping, streaming and social media platforms. This also connects the different types of filtering methods for their recommender systems. The "black box" problems that arise with recommender systems in general are amplified – it is unclear what the input or output is. It is also unclear what type of recommendation method currently has the upper hand in developing a game recommendation.

This is what we call a "black hole": It is unclear in these evolving, interacting neural networks what inputs are being examined by the recommender system: Is it preferences for publishers or for genres? Is it reviews that users have submitted? Is it purchases and clicks? Is it what friends think is good or primarily what users think is good for them? What kind of output is being generated? What kind of recommendation system is being focused on? – The collaborative one of the shopping component, the content-based one of the streaming interface, or the hybrid one of the social media button? Which algorithms are preferred?

Or even: What information collected at any point is lost and disappears into obscurity, and what information is used to generate recommendations? Much remains unclear, opaque and therefore unexplainable to the user. However, this lack of transparency affects not only the user, but also, in a weakened manner, the platform operator who sets up the recommendation system on their platform. Unlike the user, the operator can estimate which algorithms are used in a specific case, but the learning ability of the systems and the varying weighting of the available data resulting from this also remain a "black hole" for them. Nevertheless, the legislator's primary focus in the laws to be described in more detail below, the Proposal for an Artificial Intelligence Act and the Proposal for a Digital Services Act (Regulation (EU) 2022/2065), is to protect the user. For this reason, and since users are even more affected by opacity, the following considerations focus on them in order to solve this problem.

4.4 Legal Bases and Consequences

"Where opacity is the problem, transparency is the solution" (Zednik 2021). In recent years, voices have already been raised in the legal and policy literature taking this approach and proposing ways to eliminate opacity (e.g., Leerssen 2020). More recently, the call for transparency with respect to AI systems has also found its way into recent legislative proposals, as described above. Regarding the regulation of AI in particular, the European legislator has taken a major step in recent years and has been active legislatively by means of two legal acts that are worth taking a closer look at: the Digital Services Act (Regulation (EU) 2022/2065), and the Proposal for an Artificial Intelligence Act (COM/2021/206 final). However, it is the requirements of these two legislations that require a closer look in a second step in terms of appropriate implementation – especially with regard to the multidimensional opaque gaming platforms.

As mentioned at the beginning, the generic term of explainability can be divided into transparency on the one hand and interpretability on the other, with transparency referring in particular to the technical components and the algorithms (Lipton 2016). In their legislative proposals, the European institutions also refer to a concept of transparency that is in some respects prior to explainability (Berberich and Seip 2021). Since the legislator particularly focuses on the protection of the users of the platforms, we refer to user-oriented transparency in this context, which can admittedly only represent one aspect of the explicability of the entire system.

4.4.1 Legal Acts

Specifically relevant to recommender systems are two current legislative developments: the Digital Services Act (Regulation (EU) 2022/2065), which came into force on 16th November 2022 and will be fully applicable as of 17th February 2024, and the draft Artificial Intelligence Act (COM/2021/206 final). While the Digital Services Act broadly addresses transparency and impact of systems, in addition to definitions of terms, due diligence requirements, and enforcement mechanisms, the Artificial Intelligence Act focuses on the design and development of systems. Together, the two approaches can help address the black hole problem by developing solutions based on the requirements of the laws.

The legislation and its recitals make it clear, that European legislators believe that recommender systems have a significant impact on the ability of recipients to access or interact with information online. They play an important role in reinforcing certain messages, spreading information virally, and encouraging online behavior (see recital 70 of the Digital Services Act). With the establishment of recommendation systems, a completely new set of problems has arisen with regard to the amount of information conveyed online. This particularly applies because the system offers a large attack surface in terms of possible interference and therefore the risk of misuse is particularly obvious.

4.4.2 Digital Services Act

The Digital Services Act (Regulation (EU) 2022/2065) contains two fundamental assertions about recommender systems.[6] First, the already mentioned recital 70 clarifies that the recommendation system has a central function: “A core part of a

[6] For a detailed analysis of the Digital Services Act regulations, see: Gerdemann and Spindler 2023a; and Gerdemann and Spindler 2023b.

very large online platform's business is the manner in which information is prioritised and presented on its online interface to facilitate and optimise access to information for the recipients of the service. This is done, for example, by algorithmically suggesting, ranking and prioritising information, distinguishing through text or other visual representations, or otherwise curating information provided by recipients." Second, Article 3(s) of the Digital Services Act provides, for the first time, a legal definition of recommender systems: "'Recommender system' means a fully or partially automated system used by an online platform to suggest in its online interface specific information to recipients of the service, including as a result of a search initiated by the recipient or otherwise determining the relative order or prominence of information displayed."

4.4.2.1 Problem Description

Recommender systems have opened up a completely new field of problems: In addition to the undisputed benefits that recommender systems offer, they also open up the possibility of disseminating disinformation (which is not illegal in as such) and increasingly disseminating it to end-users by exploiting algorithmic systems (Schwemer 2021). The main addressees of the due diligence obligations of the Digital Services Act are therefore the recipients of the services provided by the recommender systems, i.e. the end-users of the platform. This is also supported by the wording of recital 70 of the Digital Services Act.

As positive as it is that the European Union recognizes the potential of recommendation systems, it is to be criticized that the regulations are to apply only to "large online platforms". The legislator defines what online platforms are in Art. 3(i) of the Digital Services Act. According to this, other Internet intermediaries are exempt from the application of the standards for the protection of users, on the one hand, and platforms with less than 45 million users per month, on the other (cf. Art. 33(1) of the Digital Services Act). This may apply to gaming platforms such as Steam (Statista 2022), as it is the largest gaming platform in Europe, but not to other providers in this field. The concept of information in Art. 3(s) Digital Services Act is to be understood broadly, however. With the help of the wording of recital 70, the application of the law refers to the algorithmic suggestion, ordering and prioritization of information.

4.4.2.2 Regulatory Content Related to Recommender Systems

Large online platforms within the meaning of the Digital Services Act are thus subject to various transparency requirements with regard to the most important parameters of the automated, possibly AI-supported, decision. With regard to the details, a look at Art. 14, 27, 34, 35 and 38 of the Digital Services Act is recommended. Art. 38 (supplemented by Arts. 14 and 27) of the Digital Services Act requires that users of recommendation systems of large online platforms be provided with alternative options for the most important parameters of the system. This includes, in particular, options that are not based on profiling of the recipient. However, this key requirement necessarily assumes that the user knows and understands both the processes and the alternative options. It is unclear to what extent, with regard to a more detailed explanation of the circumstances, the obligation under Art. 14 of the Digital Services Act also applies. Accordingly, information about content moderation practices has to be provided, e.g., with regard to algorithmic decision making and human verification. In addition, the intermediaries addressed in Art. 14 Digital Services Act also have to act diligently, objectively in a proportionate manner, and with appropriate consideration of the rights and legitimate interests of all parties involved (including, in particular, the fundamental rights of users). However, and this argues opposed to the assumption of obligations with respect to recommender systems, this concerns the restrictions imposed, i.e., content blocking (Schwemer 2021). According to Art. 34 of the Digital Services Act, major online platforms are required to conduct an annual risk assessment to evaluate any significant systemic risks arising from the operation and use of their services in the European Union. In doing so, the large online platforms are required to consider in particular how their recommendation systems impact any of the systemic risks, including the potentially rapid and widespread dissemination of illegal content and information consistent with their terms and conditions, cf. Art. 34(2) of the Digital Services Act. Based on this risk assessment, Art. 35 of the Digital Services Act requires the large online platform to take appropriate, proportionate and effective measures to mitigate the risk, including adapting the recommendation systems.

4.4.3 Artificial Intelligence Act

The Artificial Intelligence Act (COM/2021/206 final) specifically addresses the regulation of artificial intelligence systems. This proposed legislation could also become relevant to recommender systems-particularly in light of the discussion about fairness, accountability, and transparency of certain recommender systems (Schwemer 2021). The proposed legislation follows up on the European Commission's White Paper on AI by setting policy requirements to achieve the dual goal of promoting the use of AI and addressing the potential risks associated with it.

4.4.3.1 Purpose of the Draft Act

The Artificial Intelligence Act aims to establish harmonized rules for the development, marketing, and use of AI systems that differ in their characteristics and risks, including prohibitions and a conformity assessment system aligned with the European Product Safety Act (Council Directive 85/374/EEC). The majority of the wording of the Artificial Intelligence Act derives from a 2008 decision (Decision No. 68/2008/EC of the European Parliament and of the Council of 9 July 2008 on a common framework for the marketing of products, and repealing Council Decision 93/465/EEC, OJ L 218/82.) that established a framework for certain product safety regulations. The principal enforcement authorities used to review the requirements of the Artificial Intelligence Act – market surveillance authorities – are also common in European product law (Veale and Borgesius 2021).

The Artificial Intelligence Act defines AI system in Article 3(1) of the Artificial Intelligence Act as "software developed using one or more of the techniques and concepts listed in Annex I that is capable of producing results such as content, predictions, recommendations, or decisions that influence the environment with which it interacts with respect to a set of human-determined goals." In addition, the European Commission distinguishes four levels of AI risk: (1) AI systems with unacceptable risks, which are prohibited; (2) AI systems with high risks, which are permitted but subject to certain obligations; (3) AI systems with limited risks, which are subject to certain transparency obligations; and (4) AI systems with minimal risks, which are permitted (Schwemer et al. 2021).

4.4.3.2 Regulatory Content Related to Recommender Systems

The proposal's definition of artificial intelligence in Art. 3 No. 1 Artificial Intelligence Act is drafted quite broadly, so that at first glance recommendation systems also fall within its scope. Due to the risk management system pursued by the proposal, in which foreseeable and other emerging risks are to be assessed (cf. Art. 9 Artificial Intelligence Act), the question arises as to whether a recommendation system would be classified as high-risk. This question is addressed by Art. 6 of the Artificial Intelligence Act in conjunction with Annex III of the draft act: There, eight areas are listed in which the use of AI systems is considered risky. Insofar as a recommendation system is used in the context of legal information, this can probably be affirmed on the basis of the legal requirements. In the case of media and shopping platforms, however, rather not.

Nevertheless, there are transparency obligations throughout – regardless of which risk level an AI system belongs to. The provision of Art. 52 Artificial Intelligence Act is pertinent, which establishes the obligation to inform natural persons that they are interacting with an AI system, unless this is evident from the circumstances or context of use. For example, an advanced chatbot is required to carry the information that the interaction is not with a human being, but with the AI system (Schwemer et al. 2021). In addition, Art. 14 of the Artificial Intelligence Act requires that (at least for high-risk AI systems) human supervision should be present to prevent or at least minimize risks to the health, safety, and fundamental rights of data subjects.

4.4.4 Dealing with Legal Requirements

A consideration of the two draft acts illustrates that both address the concept of responsibility of recommender systems in terms of fairness, accountability and transparency – however, they are weighted and considered differently. The question therefore arises as to how sufficient transparency, measured against the legal requirements, can be ensured, particularly in the case of multidimensional platforms, which we have described with the metaphor of the "black hole". How can platform providers disclose which form of recommendation systems are in the foreground and which type of algorithms bring the greatest possible success without putting themselves on display and without disclosing operational and also success secrets?

Not least with recourse to the European Union Regulation establishing a general framework for securitization and a specific framework for simple, transparent and standardized securitization (Regulation (EU) 2017/2402), a number of industry proposals have emerged on how transparency for recommender systems could be presented.

4.4.4.1 User-Oriented Transparency

User-oriented transparency could serve as the first proposed solution. This form of transparency aims to direct information to the individual user in order to empower him or her with regard to the recommendation system for its content. The overall goal of this form of transparency is to raise user awareness and inform them of the options available. This should help them develop their own preferences and consider personal values such as individual autonomy, agency, and trust in their decisions (Van Drunen et al. 2019; Leerssen 2020). This consideration takes into account the legal requirements of Art. 38 of the Digital Services Act, which requires transparency about key parameters and the possibility of alternatives, and underscores the "What does the system do?" question of a system's user and stakeholders addressed above (Zednik 2021). Thus, in terms of the recommendation system on gaming platforms, users need to be informed whether the recommendation is derived from previous shopping, streaming, or social media activities, which is accompanied by information about what machine learning methods are used.

A similar type of transparency is also found in the General Data Protection Regulation (GDPR) in Articles 5, 12, 13, 14 and 22 GDPR. Specifically, it addresses the right to be informed about the parties that affect editorial decisions and the profiles that are created about groups of data subjects based on the data fed by algorithms, as well as the relevant metrics and factors of those algorithms (Van Drunen et al. 2019; Leerssen 2020). This addresses the very basis that also constitutes the "black hole" problem.

However – and this is also correctly pointed out by critics of the transparency requirement – this demand for user-oriented transparency is not entirely without problems. Due to the platforms' users' prior knowledge and understanding of how the recommendation system works and the algorithms and filtering methods behind it, it is difficult to present the explanations in a complete and comprehensible manner. According to the requirements of Art. 38 of the Digital Services Act, the user has to be provided with an alternative if he or she does not agree with the parameters used for the recommendation. Furthermore, it is certainly an ambitious goal to include personal values in the selection of recommendations – after all, these are subjective in nature. There is no question of reflecting public values. Nevertheless, especially in a niche like online gaming, it is a good start to involve users in deciding on good recommendations, to give them control, and to actively shape their rights to information about the process.

4.4.4.2 Government Oversight

Another conceivable option for enforcing transparency is government oversight. In this proposed solution, a public body would have the task of monitoring recommender systems for compliance with the transparency standards set by the legal framework or making proposals for their design (Leerssen 2020).

The idea of government oversight and regulation is not new: It can be found in the area of data protection and competition issues in many European countries. State supervision with regard to non-discrimination is also established in the media landscape, for example in the German Media State Treaty. With regard to gaming platforms, however, the question arises as to whether state supervision can be crowned with success, especially in such a peripheral area of media activity. Looking at the state regulation of German gambling law, there have been immense problems with the recognition of the regulations during implementation. Whether state supervision is therefore suitable for the niche area of gaming may be doubted.

In addition to the intentions of the aforementioned legislation (protection of users of online platforms), other interests can also be enforced here, such as public interests and concerns for the protection of minors. The approach of state supervision would therefore have the undeniable advantage that, due to the multitude of state resources, a body with sufficient expertise could be formed to adequately address the multitude of problems. However, this presupposes a statutory reporting obligation on the part of platform operators, which is also difficult to implement in other factual constellations (e.g., plagiarism control, defective products in online commerce). In particular, due to the special role granted to platform operators in the area of telecommunications law, such reporting obligations are difficult to implement (see Gielen and Uphues 2021; Spindler 2021).

4.4.4.3 Combination of the Two Approaches with Additional Experts

A third approach combines the two ideas above by ensuring the transparency of recommender systems on gaming platforms and having them jointly supervised and monitored by representatives of academia and parts of civil society. This allows for research into the use of recommender systems as well as their practical criticism and questioning (Leerssen 2020). This idea links to the aforementioned problems of the other two approaches and tries to reconcile them: The lack of user expertise is replaced by the insights that research partners bring to the field. However, the practical needs, especially in such a niche industry, are determined by the users of the platform. With the insights gained from information about how the system works, regulatory and recommendatory interventions can be provided.

However, this approach also has a drawback: The law itself limits the effectiveness of this method. Science needs a lot of data to effectively control and develop the system, so, easy access to data and openness in processing information is desired. At the same time, it is the intention of the legislator (see the General Data Protection Regulation, GDPR) to keep access to data to a minimum. The generated data may only be accessed with the appropriate legal permission. Without the consent of all users of the gaming platform, it will be as difficult as for other users to find out which filtering system and which algorithm combination leads to (which?) result. While the ideal of openness and transparency is advocated and is also important to strengthen trust in the system – it is impossible to look behind the scenes.

4.5 Implementation of the Proposed Solutions

However, despite all the theoretical considerations and the question of the appropriate group of experts for implementation, it should not be forgotten that a practical solution is also required. In the following, possible approaches are presented which have already been discussed in other application environments and which also represent valuable considerations for the case described here.

4.5.1 Standardization

One of the most important keywords that can be mentioned with regard to a possible solution is that of standardization, both in the sense of technical standardization, but also in the sense of legal standardization and regulation. If it would be possible to design a technically comprehensible solution that explains in an understandable way in which such multidimensional platforms as the gaming platform are recommended, and if it would be correspondingly clear what kind of filtering process – content-based, collaborative or hybrid – would be used, a first step would be taken. Both users and indie game developers would then be able to understand how the user interface is constructed. The “black hole” would then become a “black box” again – and even if this is not a satisfactory state, it is easier to manage because of the existing research results. Harmonization – standardization – of recommender system technology in this area of application would be extremely helpful. It would also support our argument for putting the users of the system at the forefront in terms of information and transparency – because then the system would be at least a bit easier to understand for the users. Standardization is also a suggestion that could and should be considered on the legal side. The advantage of standardization in application is obvious: Recommendation systems for platforms that combine multidimensional decision algorithms would, in this case, work the same way, proceed the same way, and platform operators would be equally committed to standardized transparency. Another advantage is related to the emergence of these standards: The expert group developing the standardization framework is composed of people with appropriate expertise; this group, supplemented by users of the platforms with appropriate expertise, can profitably monitor and evaluate the security of the IT infrastructure. The more diverse perspectives represented within the team, the more likely it is that the team's work product will address all technically, ethically, and legally relevant aspects. Given these conditions, better standards can be set on the basis of the different know-how standards.

4.5.2 Control Mechanisms

The problem remains that every AI application remains a "black box", even if the recommendation system is one-dimensional or technically standardized in a way that at least makes the uncertainty of the system's nature clear to the user. This makes it all the more important to nevertheless control and understand the unpredictability of AI to some extent and thus make it manageable. For example, an internal control system could be built in to minimize risk (Bittner et al. 2021). This addresses platform operators' worries that the recommender system can be abused if they publish its capabilities. It also raises user awareness by signaling, for example, that the system performs frequent backups. Misuse by users could be counteracted by platform operators regularly participating in training or using tools to control the misuse.

Many AI systems have different requirements. Since recommender systems have been recognized as an important tool by the European Union, a verification process could be developed to prevent transparency and protection against abuse of market power by the platform operator. Mathematical-statistical models can be used to detect and analyze errors and deviations in the model (Bittner et al. 2021). In this context, it is mandatory to adhere to the multiaudience principle during the development of the software in order to sufficiently ensure quality (Lindstrom and Jeffries 2003). At the same time, the continuous comprehensibility of the AI algorithm needs to be ensured: How are the algorithms structured? According to which rules can it learn? And – particularly relevant in the case of neural networks – how quickly can it evolve and are the control mechanisms then sufficient?

Assuming that these extensive considerations cannot be enforced globally, but that the guidelines should be at the national level, there is another helpful support: liability regulations. Although many discussions occur in the context of public law, the advantages of private law should not be dismissed. For example, a liability law framework can control risks to some extent. However, there are broader problems associated with this consideration, particularly issues of conflict of laws (Lutzi 2020). The extent to which liability law or even competition law regulations could support the above approaches (in the sense of reciprocal fallback regulations, see Gesmann-Nuissl 2020) remains to be examined.

4.6 Conclusion

It is well known that a certain opacity is inherent in an AI system based on deep neural networks. Already at this stage, the demand for explicability of such systems becomes loud. This is also a problem with regard to recommendation systems that recommend suitable (digital) products to the user, also by means of AI systems, and the legislator has also recognized this problem. In the Digital Services Act (Regulation (EU) 2022/2065), recommendation systems are explicitly named, defined and the need for transparency is explicitly demanded by law. The problem we call “black hole” represents a form of multidimensionality. By combining different filtering methods used in known forms of recommender systems, namely content-based, collaborative and hybrid filtering methods on a single (gaming) platform, we exacerbate the phenomenon of opacity of AI systems sometimes known as the “black box” problem. It is particularly important to look beyond the opacity of individual algorithms and understand the complex interactions between technology and users. The solution to this “black hole” problem needs to focus on all levels of transparency, which the European Union addresses in its legislations. This includes the algorithms used (simple decision trees or deep neural networks), the filtering methods used (content-based or collaborative) and, in particular, the type of recommendation and the content and data used for this purpose. A balanced approach between users, manufacturers, regulators, government and research is needed to address the problem of double opacity and ultimately to increase the confidence of users, but also of platforms, in this technology – which, after all, brings many benefits.

Knowing what is technically conceivable, and knowing that it is feasible to also technically implement and legally secure the specifications required by the legislative proposals, the Digital Services Act and the Artificial Intelligence Act (Regulation (EU) 2022/2065, COM/2021/206 final), will help us to design and standardize guidelines for transparent AI. All of this is also in the interest of legislators. The Artificial Intelligence Act explicitly requires transparency of any system, regardless of the risk level to which it is assigned to. This would be a first step to regulate and certify multidimensional recommender systems.

Legal requirements impose certain transparency requirements. To meet these minimum requirements, AI systems in general and recommender systems in particular (especially since they are mentioned by name in the Digital Services Act) need to have a certain level of security that grants transparency and, accordingly, explainability to the user. These legal requirements, which will come into force in the near future with the AI Act and the Digital Services Act, can be implemented in various ways for recommender systems. On the one hand, user-oriented transparency is conceivable, which has already been implemented to some extent on gaming platforms. This type of transparency is intended to empower the user to control and manage the content of the recommender system, allowing individual values to be better taken into account – but there is the problem that the user cannot fully grasp how the system works. Alternatively, a government authority could also exercise oversight (similar to data protection or competition law). However, the past has shown that specific areas of application (such as the area of recommendation systems in this case) are difficult to regulate, especially since this involves certain reporting obligations. Another solution would be to combine the aforementioned approaches – and to combine user (interests) and state supervision (interests). This would strengthen trust in the guiding hand of the state and the application-oriented representation of interests by joint expert committees. These bodies would then also be in a position to implement the issues addressed here, for example by means of standardization. This would help to meet the different requirements of the users and users of the platforms.

Acknowledgements Funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – Project-ID 416228727 – SFB 1410.

* The original manuscript was submitted in January 2022 and has since been updated to reflect legal developments.

References

Abdullah, T.A.A., M.S.M. Zahid, and W. Ali. 2021. A Review of Interpretable ML in Healthcare: Taxonomy, Applications, Challenges, and Future Directions. *Symmetry* 2021 (13): 2439. https://doi.org/10.3390/sym13122439.

Adomavicius, G., and A. 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): 734–749. https://doi.org/10.1109/TKDE.2005.99.

Anand, A., K. Bizer, A. Erlei, U. Gadiraju, C. Heinze, L. Meub, W. Nejdl, and B. Steinröotter. 2018. Effects of Algorithmic Decision-Making and Interpretability on Human Behavior: Experiments using Crowdsourcing. In *Proceedings of the sixth AAAI Conference on Human Computation and Crowdsourcing (Hcomp-18)*. Zurich: AAAI Press.

Ananny, M., and K. Crawford. 2018. Seeing Without Knowing: Limitations of the Transparency Ideal and its Application to Algorithmic Accountability. *New Media & Society* 20 (3): 973–989. https://doi.org/10.1177/1461444816676645.

Anderson, P.W. 1972. More is Different: Broken Symmetry and the Nature of the Hierarchical Structure of Science. *Science* 177 (4047): 393–396.

Barreau, B. 2020. Machine Learning for Financial Products Recommendation. Computational Engineering, Finance, and Science [cs.CE]. Université Paris-Sacla.

Berberich, M., and F. Seip. 2021. Der Entwurf des Digital Services Act. *GRUR-Prax*: 4–7.

Bittner, J., N. Debowski, M. Lorenz, H.G. Taber, H. Steege, and K. Teile. 2021. Recht und Ethik bei der Entwicklung von Künstlicher Intelligenz für die Mobilität. *Neue Zeitschrift für Verkehrsrecht* 34 (10): 505–513.

Bryson, J. 2019. Six kinds of explanation for AI (one is useless). https://joanna-bryson.blogspot.com/2019/09/six-kinds-of-explanation-for-ai-one-is.html

Burrell, J. 2016. How the Machine 'Thinks': Understanding Opacity in Machine Learning Algorithms. *Big Data & Society* 3 (1). https://doi.org/10.1177/2053951715622512.

Castelvecchi, D. 2016. Can We Open the Black Box of AI? *Nature* 538 (7623): 20–23. https://doi.org/10.1038/538020a.

Cobbe, J., and J. Singh. 2019. Regulating Recommending: Motivations, Considerations and Principles. *European Journal of Law and Technology* 10 (3). https://doi.org/10.2139/ssrn.3371830.

Covington, P., J. Adams, and E. Sargin. 2016. Deep Neural Networks for YouTube Recommendation. In *Proceedings of the 10th ACM Conference on Recommender Systems*, 191–198. Boston: Association for Computing Machinery. https://doi.org/10.1145/2959100.2959190.

Das, A., M. Datar, A. Garg, and S. Rajaram. 2007. Google News Personalization: Scalable Online Collaborative Filtering. In *Proceedings of the 16th International conference on World Wide Web*, 271–280. Alberta: Association for Computing Machinery. https://doi.org/10.1145/1242572.1242610.

Davidson, J., B. Liebald, J. Liu, P. Nandy, T. Van Vleet, U. Gargi, S. Gupta, Y. He, M. Lambert, B. Livingston, and D. Sampath. 2010. The YouTube Video Recommendation System. In *Proceedings of the 2010 ACM Conference on Recommender Systems*, 293–296. Barcelona: Association for Computing Machinery. https://doi.org/10.1145/1864708.1864770.

Deng, J., F. Cuadradro, G. Tyson, and S. Uhlig. 2015. Behind the Game: Exploring the Twitch Streaming Platform. In *Proceedings of 2015 International Workshop on Network and Systems Support for Games (NetGames)*, 1–6. Zagreb: IEEE. https://doi.org/10.1109/NetGames.2015.7382994.

Ebers, M. 2020. Regulierung von KI und Robotik. In *Künstliche Intelligenz und Robotik*, ed. M. Ebers, C. Heinze, T. Krügel, and B. Steinrötter, 82–140. München: Beck.

Ensthaler, J., D. Gesmann-Nuissl, and S. Müller. 2012. *Technikrecht – Rechtliche Grundlagen des Technologiemanagements*. Berlin: Springer. https://doi.org/10.1007/978-3-642-13188-2.

European Union. 1985. Council Directive 85/374/EEC of 25 July 1985 on the approximation of the laws, regulations and administrative provisions of the Member States concerning liability for defective products. https://eur-lex.europa.eu/eli/dir/1985/374/oj

———. 2017. Regulation (EU) 2017/2402 of the European Parliament and of the Council of 12 December 2017 establishing a general framework for securitization and creating a specific framework for simple, transparent and standardized securitization. https://eur-lex.europa.eu/eli/reg/2017/2402/oj

———. 2021. Proposal for a Regulation of the European Parliament and of the Council Laying Down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act) and Amending Certain Union Legislative Acts, COM/2021/206 final. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex%3A52021PC0206

———. 2022. 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/eli/reg/2022/262/oj

Fleder, D., K., Hosanagar, and A. Buja. 2010. Recommender Systems and their Effects on Consumers: The Fragmentation Debate. ACM 978-1-60558-822-3/10/06.

Gahier, A.K., and S.K. Gujral. 2021. Cross Domain Recommendation Systems using Deep Learning: A Systematic Literature Review. In *Proceedings of the International Conference on Innovative Computing & Communication (ICICC) 2021*. Delhi: Springer. https://doi.org/10.2139/ssrn.3884919.

Gerdemann, S., and G. Spindler. 2023a. Das Gesetz über digitale Dienste (Digital Services Act) (Teil 1). Grundlegende Strukturen und Regelungen für Vermittlungsdienste und Host-Provider. *GRUR – Gewerblicher Rechtsschutz und Urheberrecht* 2023: 3–11.

———. 2023b. Das Gesetz über digitale Dienste (Digital Services Act) (Teil 2). Die Regelungen für Online-Plattformen sowie sehr große Online-Plattformen und -Suchmaschinen. *GRUR – Gewerblicher Rechtsschutz und Urheberrecht* 2023: 115–125.

Gesmann-Nuissl, D. 2020. Zivil- und Gewerberecht als wechselseitige Auffangordnungen. In *150 Jahre Gewerbeordnung*, ed. W. Kluth and S. Korte, 64–82. Göttingen: Cuviller.

Gielen, N., and S. Uphues. 2021. Digital Markets Act und Digitals Services Act – Regulierung von Markt- und Meinungsmacht durch die Europäische Union. *Europäische Zeitschrift für Wirtschaftsrecht* 2021 (14): 627–636.

Goanta, C., and G. Spanakis. 2020. Influencers and Social Media Recommender Systems: Unfair Commercial Practices in EU and US Law. TTLF Working Paper no. 54. https://doi.org/10.2139/ssrn.3592000.

Hamilton, W.A., O. Garretson, and A. Kerne. 2014. Streaming on Twitch: Fostering Participatory Communities of Play within Live Mixed Media. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 1315–1324. Toronto: Association for Computing Machinery. https://doi.org/10.1145/2556288.2557048.

Hohman, F.M., M. Kahng, R. Pienta, and D.H. Chau. 2018. Visual Analytics in Deep Learning: An Interrogative Survey for the Next Frontiers. *IEEE Transactions on Visualization and Computer Graphics* 25 (8): 2674–2693. https://doi.org/10.1109/TVCG.2018.2843369.

Holzinger, A. 2018. Explainable AI (ex-AI). *Informatik Spektrum* 41: 138–143. https://doi.org/10.1007/s00287-018-1102-5.

Humphreys, P. 2008. The Philosophical Novelty of Computer Simulation Methods. *Synthese – An International Journal for Epistemology, Methodology and Philosophy of Science.* 169: 615–626. https://doi.org/10.1007/s11229-008-9435-2.

Isaias, P., C., Casaca, and S. Pifano. 2010. Recommender Systems for Human Resources Task Assignment. 2010 24th IEEE International Conference on Advanced Information Networking and Applications, 214–221. https://doi.org/10.1109/AINA.2010.168.

Knijnenburg, B.P., M.C., Willemsen, Z., Gantner et al. 2012. Explaining the user experience of recommender systems. User Modeling and User-Adapted Interaction 22: 441–504. https://doi.org/10.1007/s11257-011-9118-4.

Körner, S.J. 2020. Nachvollziehbarkeit von KI-basierten Entscheidungen. In *Rechtshandbuch Artificial Intelligence und Machine Learning*, ed. M. Kaulartz and T. Braegelmann, 15–58. München: Beck/Vahlen.

Leerssen, P. 2020. The Soap Box as a Black Box: Regulating Transparency in Social Media Recommender Systems. *European Journal of Law and Technology* 11 (2). https://doi.org/10.2139/ssrn.3544009.

Linardatos, D. 2020. § 1 Technische und rechtliche Grundlagen. In *Rechtshandbuch Robo Advice. Automatisierte Finanz- und Versicherungsdienste*, ed. D. Linardatos, 1–28. München: Vahlen, Beck.

Lindstrom, L., and R. Jeffries. 2003. *Information Security Management Handbook*. 8th ed. Auerbach Publications.

Lipton, Z. 2016. The Mythos of Model Interpretability. In *Communications of the ACM*, vol. 61, 36–43. New York: Association for Computing Machinery. https://doi.org/10.1145/3233231.

Lutzi, T. 2020. *Private International Law Online*. Oxford: Oxford University Press.

Mahesh, T. R., and V. Vivek. 2021. *Recommendation Systems: The Different Filtering Techniques, Challenges and Review Ways to Measure the Recommender System.* https://doi.org/10.2139/ssrn.3826124.

Marr, D. 1982. *Vision: A Computational Investigation into the Human Representation and Processing of Visual Information*. Cambridge: MIT Press.

Maume, P. 2021. Robo-advisors. How do they fit in the existing EU regulatory framework, in particular with regard to investor protection? Study, requested by the ECON committee. Available online: https://www.europarl.europa.eu/RegData/etudes/STUD/2021/662928/IPOL_STU(2021)662928_EN.pdf. (Accessed 31 May 2023).

Mohanty, S.N., J.M. Chatterjee, S. Jain, A.A. Elngar, and P. Gupta. 2020. *Recommender Systems with Machine Learning and Artificial Intelligence*. Hoboken: Wiley. https://doi.org/10.1002/9781119711582.

Niederée, C., and W. Nejdl. 2020. Technische Grundlagen der KI. In *Künstliche Intelligenz und Robotik*, ed. M. Ebers, C. Heinze, T. Krügel, and B. Steinrötter, 42–81. München: Beck.

Nielson, C., and D. Killeen. 2022. Artificial Intelligence: The Impending Change of Work. https://cjnielson.com/wp-content/uploads/2022/04/ENC-AI-and-the-Future-of-Work-Christopher-Nielson.pdf

Pasquale, F. 2015. *The Black Box Society: The Secret Algorithms that Control Money and Information*. Cambridge: Harvard University Press.

Raj, A. 2020. Explainable AI: From Black Box to Glass Box. *Journal of the Academy of Marketing Science* 48 (1): 137–141. https://doi.org/10.1007/s11747-019-00710-5.

Rapaz, J., J. McAuley, and K. Aberer. 2021. Recommendation on Live-Streaming Platforms: Dynamic Availability and Repeat Consumption. In *Proceedings of the Fifteenth ACM Conference on Recommender Systems*, 390–399. Amsterdam: Association for Computing Machinery. https://doi.org/10.1145/3460231.3474267.

Ribeiro, M. T., S. Singh, and C. Guestrin. 2016. Why Should I Trust You?: Explaining the Predictions of Any Classifier. arXiv:1602.04938. https://doi.org/10.48550/arXiv.1602.04938

Rieder, G., and J. Simon. 2017. Big Data: A New Empiricism and its Epistemic and Socio- political Consequences. In *Berechenbarkeit der Welt? Philosophie und Wissenschaft im Zeitalter von Big Data*, ed. W. Pietsch, J. Wernecke, and M. Ott, 85–105. Wiesbaden: Springer.

Rieder, B., A. Matamoros-Fernández, and Ò. Coromina. 2018. From Ranking Algorithms to 'ranking cultures': Investigating the Modulation of Visibility in YouTube Search Results. *Convergence: The International Journal of Research into New Media Technologies* 24 (1): 50–68. https://doi.org/10.1177/1354856517736982.

Robbins, S. 2019. A Misdirected Principle with a Catch: Explicability for AI. *Minds and Machines* 29 (4): 495–514. https://doi.org/10.1007/s11023-019-09509-3.

Sarker, I. 2021. Machine Learning: Algorithms, Real-World Applications and Research Directions. *Preprint*. https://doi.org/10.20944/preprints202103.0216.v1.

Schmidt, J.-H., J. Sørensen, S. Dreyer, and U. Hasebrink. 2018. Wie können Empfehlungssysteme zur Vielfalt von Medieninhalten beitragen? *Media Perspektiven* 2018 (11): 522–531. https://www.ard-media.de/fileadmin/user_upload/media-perspektiven/pdf/2018/1118_Schmidt_Soerensen_Dreyer_Hasebrink.pdf.

Schwemer, S.F. 2021. Recommender Systems in the EU: from Responsibility to Regulation? In *FAccTRec Workshop '21*. Amsterdam: Association for Computing Machinery. https://ssrn.com/abstract=3923003

Schwemer, S.F., L. Tomada, and T. Pasini. 2021. Legal AI Systems in the EU's Proposed Artificial Intelligence Act. In *Proceedings of the Second International Workshop on AI and Intelligent Assistance for Legal Professionals in the Digital Workplace (LegalAIIA 2021)*. Sao Paulo: Ceur WS. https://ssrn.com/abstract=3871099

Silva, D.V. 2019. Information retrieval models for recommender systems. https://www.dc.fi.udc.es/~dvalcarce/thesis.pdf

Sousa, J., and J. Barata. 2021. Tracking the Wings of Covid-19 by Modeling Adaptability with Open Mobility Data. *Applied Artificial Intelligence* 35 (1): 41–62. https://doi.org/10.1080/08839514.2020.1840196.

Spindler, G. 2021. Der Vorschlag für ein neues Haftungsregime für Internetprovider – der EU-Digital Services Act (Teil 1). *Gewerblicher Rechtsschutz und Urheberrecht* 4: 545–553.

Statista. 2022. Number of visits to steampowered.com from October 2019 to December 2021. https://de.statista.com/statistik/daten/studie/1112237/umfrage/anzahl-der-visits-pro-monat-von-steampoweredcom/

Steam (© 2022 Valve Corporation). 2020. https://store.steampowered.com/news/app/593110/view/1716373422378712840.

Tomsett, R., D. Braines, D. Harborne, A. Preece and S. Chakraborty. 2018. Interpretable to Whom? A Role-based Model for Analyzing Interpretable Machine Learning Systems. ArXiv1806.07552. https://doi.org/10.48550/arXiv.1806.07552.

Uzair, M., and N. Jamil. 2020. Effects of Hidden Layers of the Efficiency of Neural Networks. In *Proceedings of the IEEE 23rd International Multitopic Conference (INMIC)*, 1–6. Bahawalpur: IEEE. https://doi.org/10.1109/INMIC50486.2020.9318195.

Van Drunen, M., N. Helberger, and M. Bastian. 2019. Know Your Algorithm: What Media Organizations Need to Explain to Their Users about New Personalization. *International Data Privacy Law* 9 (4): 220–235. https://doi.org/10.1093/idpl/ipz011.

Veale, M., and F.Z. Borgesius. 2021. Demystifying the Draft EU Artificial Intelligence Act. *Computer Law Review International* 22 (4): 97–112. https://doi.org/10.31235/osf.io/38p5f.

Waltl, B. 2019. Erklärbarkeit und Transparenz im Machine Learning. In *Philosophisches Handbuch Künstliche Intelligenz*, ed. K. Mainzer, 1–23. Wiesbaden: Springer. https://doi.org/10.1007/978-3-658-23715-8_31-1.

Wang, Z., W. Zhu, P. Cui, L. Sun, and S. Yang. 2013. Social Media Recommendation. In *Social Media Retrieval*, ed. N. Ramzan, R. van Zwol, J. Lee, K. Clüver, and X. Hua, 23–42. London: Springer. https://doi.org/10.1007/978-1-4471-4555-4_2.

Zafar, M.B., I. Valera, M.G. Rodriguez, K.P. Gummadi, and A. Weller. 2017. From Parity to Preference-Based Notions of Fairness in Classification. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, 228–238. Long Beach: Curran Associates. https://dl.acm.org/doi/proceedings/10.5555/3294996.

Zech, H. 2019. Künstliche Intelligenz und Haftungsfragen. *Zeitschrift für die gesamte Privatrechtswissenschaft* 5 (2): 198–219.

Zednik, C. 2021. Solving the Black Box Problem: A Normative Framework for Explainable Artificial Intelligence. *Philosophy & Technology* 34: 265–288. https://doi.org/10.1007/s13347-019-00382-7.

Zhou, R., S. Khemmarat, and L. Gao. 2010. The Impact of YouTube Recommendation System on Video Views. In *Proceedings of the 10th ACM SIGCOMM Conference on Internet Measurement*, 404–410. Melbourne: Association for Computing Machinery. https://doi.org/10.1145/1879141.1879193.

Ziegler, J., and B. Loepp. 2019. Empfehlungssysteme. In *Handbuch Digitale Wirtschaft*, ed. T. Kollmann, 717–741. Wiesbaden: Springer.

Chapter 5
Digital Labor as a Structural Fairness Issue in Recommender Systems

Sergio Genovesi

Abstract This contribution moves from the assumption that algorithmic outcomes disadvantaging one or more stakeholder groups is not the only way a recommender system can be unfair since additional forms of structural injustice should be considered as well. After describing different ways of supplying digital labor as waged labor or consumer labor, it is shown that the current design of recommender systems necessarily requires digital labor for training and tuning, making it a structural issue. The chapter then presents several fairness concerns raised by the exploitation of digital labor. These regard, among other things, the unequal distribution of produced value, the poor work conditions of digital laborers, and the unawareness of many individuals of their laborer's condition. To address this structural fairness issue, compensatory measures are not adequate, and a structural change of the ways training data are collected is necessary.

Keywords Digital labor · Consumer labor · Ghost work · Algorithmic fairness

5.1 Introduction: Multisided (Un)Fairness in Recommender Systems

Current research on AI fairness extensively focuses on ways to detect, measure and prevent discrimination and unjustified unequal treatment of individuals or group of individuals being affected by automated decisions (Barocas et al. 2019). Examples of unfair outcomes of automated decisions include discrimination of credit applicants based on gender (Verma and Rubin 2018) or race (Lee and Floridi 2021), unfair distribution of access to medical treatment among patients (Giovanola and Tiribelli 2022), disadvantaging women when selecting job applications (Dastin 2018), etc. These examples have a common denominator: fairness issues are

S. Genovesi (✉)
University of Bonn, Bonn, Germany
e-mail: genovesi@uni-bonn.de

S. Genovesi et al. (eds.), *Recommender Systems: Legal and Ethical Issues*, The International Library of Ethics, Law and Technology 40,
https://doi.org/10.1007/978-3-031-34804-4_5

investigated in one specific stakeholder's group; namely, the group of users applying for access to a service or a position and being subject of algorithmic classification. While this kind of fairness evaluation is certainly essential to detect discrimination of groups or individuals, in some cases a multi-sided consideration of the different actors involved in the design, commercialization and use of an AI system is necessary to complete the picture. The ethic audit of recommender systems is one of these cases.

Evaluating fairness in recommender systems is a complex task and requires a multi-stakeholder analysis (Burke 2017; Milano et al. 2021) as well as a precise definition of the fairness aspects and indicators being audited (Deldjoo et al. 2021). Since recommender systems are socio-technical systems tailoring content recommendations for different users, Milano et al. identify four categories of stakeholders involved in a recommendation: users, content providers, system viz. platform providers and developers, and society at large (Milano et al. 2021). When asking whether a recommender system is a fair system, it is therefore necessary to ask to whom the system is being fair. Indeed, unfair treatment might affect either a group of members of one (or more) stakeholder group(s), or one (or more) stakeholder group(s) at large. To analyze a cross-stakeholder group scenario, Burke introduced the notions of C-fairness (consumer fairness), P-fairness (provider fairness) and CP-fairness (consumer and provider fairness) (Burke 2017) highlighting that, when it comes to recommending content or services, not only consumers can be discriminated (C-fairness issues), e.g. by not being shown offers that are classified as being out of their league even though they could be interesting for them, but also service providers (P-fairness issues), e.g. by being given reduced visibility to their product compared to other products of the same type. Depending on the risk for consumers, service providers, or both, of being discriminated against, a system should meet C-fairness, P-fairness or CP-fairness conditions (Burke 2017). A concrete example of discrimination risk for both stakeholders can be found in real estate sharing economy platforms, where real estate owners offering their property for rent are matched with travelers looking for accommodation. In this scenario, it has been shown that the system's performance might vary across demographic groups based, among other things, on users' self-declared gender, sexual orientation, age, and main spoken language (Solans et al. 2021). Thus, while some travelers belonging to specific demographic groups might experience limited access to housing, some renters might enjoy less visibility on the platform.

After asking who is being treated unfairly, it is essential to ask how the affected stakeholders are being treated unfairly. This adds a complexity layer to the analysis since there may be different indicators of unequal treatment. When it comes to (mis) classification of individuals, much effort is being put into current research to identify suitable metrics to quantify the disparity of system performance for different demographic groups (Mehrabi et al. 2021; Verma and Rubin 2018). Since in some applications misclassification can lead to missing important life chances, such as job or education opportunities, or access to credit or housing, it is crucial to detect and address the fairness issues related to algorithmic discrimination based on demographic attributes. In addition to this, there are fairness issues that are not related to misclassification and are hardly expressible through statistical quantities. These issues concern, among other things, the application environment of an AI-system, the decision-making process (procedural fairness) (Grgić-Hlača et al. 2018), and structural injustice (Kasirzadeh 2022). Therefore, a stakeholder belonging to a disadvantaged group might experience discrimination even though a fairness metric does not detect an unfair outcome, e.g., by being penalized before interacting with or being scrutinized by an AI system or being in a disadvantaged position compared with other stakeholders interacting with the system, e.g., having less control on the system or less bargaining power than others.

In this chapter I describe a structural fairness issue of this last type concerning recommender systems. I argue that certain stakeholder groups are structurally in a disadvantaged position because of the design of most recommender systems and their commercial implementation in a techno-capitalist environment. In Sect. 5.2, I define the disadvantaged stakeholders as those supplying data used to train and fine-tune a recommender system through their "digital labor". I characterize digital labor as a form of exploitation and describe different forms it may take. In Sect. 5.3, I highlight why the exploitation of digital labor is unethical by pointing not only at the unfair distribution of produced value among stakeholders, but also at further related issues concerning transparency and human well-being. In Sect. 5.4, I propose ways to address this structural fairness issue through the acknowledgment of digital labor, its regulation, and the end of its exploitation.

5.2 Digital Labor as a Structural Issue in Recommender Systems

In the last two decades, many scholars attempted to rethink the Marxist concept of labor in light of the digital transformation of capitalism and of the rise of social media and platform economy, examining new sites of labor market and value production (Fuchs and Fisher 2015; Maxwell 2016; Scholz 2013). Digital labor is generally understood as a value-producing activity in online environments and can either be waged or unwaged (Scholz 2013). In both cases it is characterized by explicit exploitation dynamics. Digital labor can take many forms depending on who is producing value and how the laborer's activity is exploited. In this section I introduce some kinds of digital labor presenting very different settings, but sharing the feature of being value producing activities subject to exploitation.

Starting with waged labor, it is possible to consider the cases of independent gig workers and of employed workers. Gig work based waged digital labor such as click-working to train AI-systems on Amazon Mechanical Turk is usually poorly paid, and until now did not entitle workers to a minimum wage (Aytes 2013). The almost complete absence of laws regulating work on emerging platforms at the beginning of the last decade, as well as the absence of a syndicate for gig workers, paved the way for workers' exploitation in digital environments. In addition to offering low pay for completing tasks, gig work platforms did not provide social or health insurance for workers, making their life and work conditions even more precarious. Contrary to independent gig workers, employed workers receive a fixed salary. However, this is not a guarantee of good work conditions. We can take content moderation on Facebook as an example. Content moderators might be part-time or full-time employees hired by the social media or a third-party company and not gig workers. However, also in this case unacceptable work conditions and low pay were reported (Newton 2019). Facebook content moderators revealed that employees cope with seeing traumatic images and videos by telling dark jokes about committing suicide, then smoking marijuana during breaks to numb their emotions. Moreover, employees are constantly monitored at the workplace and are allowed to take very few short breaks. After they leave the company, they often develop PTSD-like symptoms and are not eligible for any support by Facebook (Newton 2019). All these forms of waged digital labor have been recently referred to as "ghost work" (Gray and Suri 2019) to highlight the fact that those people executing (micro)tasks essential for the functioning of many apps and platforms we use in our everyday life are invisible to the end users.

Concerning unwaged digital labor, also referred as a kind of consumer labor (Jarrett 2015), users are not always aware they are engaging in value-producing activities, e.g. by training an AI-system or providing businesses with essential data. Using captchas and re-captchas as authentication methods is an example of unpaid cognitive labor since most users don't know that, at the same time, they are producing valuable data for image recognition systems (Aytes 2013). The same goes for tagging images and using hashtags on social media and other online platforms since this contributes to improved content labeling by machine learning systems (Bouquin 2020; Casilli and Posada 2019). Some scholars highlighted that even general online time on social networks, engagement with social media posts or searches on search engines should be seen as a form of consumer labor since they produce navigation data that are very valuable for the service provider (Fisher 2020; Vercellone 2020). Even though nowadays an increasing number of users are aware that navigation data are recorded and processed for algorithm optimization, user classification, targeted advertising and many other purposes, for many years ignorance about these facts were prevailing and many users did not know they were exchanging digital labor for search results or social media content. On subscription platforms users might even pay while supplying service providers with digital labor. It should be reminded that consumer labor does not only concern digital services and can be found in many other industries (Jarrett 2015) – for example when people advertise for clothing companies by having their logos printed over the shirt, or when user behavior is tracked by stores. The specificity of our case, that is, of recommender systems on digital platforms, resides in the fact that users, through their online activity, supply navigation data that are indispensable for the system to run.

Moreover, there are forms of unwaged digital labor that are not disguised by system providers and of which users are aware. One example is the unpaid creation of content that will be spread by recommender systems, thereby fueling their very functioning and constituting their reason of existence. While in some cases, posters might improve their reputation and visibility, or enjoy commercial revenue from the creation of content, private posting aiming at reaching a network of friends or followers does not generally have such a return. Particularly remarkable is the case of cultural and/or creative content produced by semi-professionals or amateurs on blogs, vlogs and other online platforms (Terranova 2000), in which individuals invest a great amount of time in work that is hardly acknowledged or rewarded by the cultural industry or by system providers.

It is important to remark that digital labor does not only concern the users' group but is rather a cross-stakeholder issue. Indeed, click-workers and gig workers constitute an independent stakeholder group since they are not necessarily platform users and are not covering the users' role in any case while completing system-training tasks. The same goes for content creators, who might be digital laborers working from a service provider's side using a content recommending platform to promote the product they represent.

To conclude this section, it is possible to show that the above-mentioned forms of digital labor are structurally necessary for the existence, functioning and profitability of recommender systems in different application environments. First, the paid or unpaid production of training data for an AI-system allows one to optimize functions such as natural language processing and image recognition, which are essential, for example, to classify social media posts and recommend them to the right users (Bouquin 2020). Furthermore, content moderation by human operators – irrespective of their employment status – is vital both to avoid the spread of harmful or illegal content that AI-content moderation systems fail to recognize, and to allow content that was wrongly flagged as offensive, whose removal would constitute a limitation on free expression. A functioning moderation loop can prevent digital platforms from losing a consistent share of their profits. For instance, in the European legal framework, as prescribed by the Digital Service Act, failure to remove notified illegal content leads to the platform's civil liability. Finally, when it comes to consumer labor, users in the role of either content viewers or creators, provide both the raw material and the final product of recommender systems and of the platforms they run on. Even though some news aggregators, as well as many music and film streaming platforms, also recommend content created by professionals or at least not in digital labor conditions, user generated content (UGC) is indispensable for all social networks, which could not exist without it. On dating apps recommending possible users to match based on their proximity, online activity and shared interests and photos (Tinder Newsroom 2022), users' profiles are the very UGC being consumed. In addition, users' online behavior and navigation data are necessary to improve the expected relevance of recommendations and, as a consequence, users' engagement on the platform. Moreover, users' content views are the very product of those content recommendation platforms (including social networks) whose main revenue depend on advertising. Without UGC and/or the monetization of users' attention, many content recommendation platforms would neither have any source of revenue nor a reason to exist.

5.3 Fairness Issues from Value Distribution to Work Conditions and Laborers' Awareness

The reliance of recommender systems on one or more forms of digital labor is structural and design driven. This fact is part of the reason why digital labor is so valuable for system providers. It is difficult to estimate how much value digital labor produces, but for many activities it is possible to roughly quantify the worth at stake. A user spending online time on a social network and whose attention is being monetized through targeted advertising is producing value worth the cost of the single advertisement's views and clicks,[1] plus a value corresponding to the worth of keeping the system updated to their current interests to optimize future content recommendation. Content creators might generate content for free, which then contributes to keeping users online that are either targeted with advertising or paying a subscription (or even both). A group of click-workers receive cents to train software that will be sold for thousands of dollars. A group of moderators paid minimum wage saves millions of dollars that would otherwise be spent on fines, lawyers and statutory damages, allowing tech companies to have higher profits. These examples highlight two main recurrent and not mutually exclusive ways digital labor produces value. The first is the direct monetization of the individual activity without considering its synergy with other laborer's activity and is more intuitive to quantify. Examples include a user clicking on advertising, a gig worker filling out data individually in a database that will be sold at a high price, a content creator creating content that will be put behind a paywall by the system provider, etc. The second derives from the synergy between the activity of many laborers and requires a holistic consideration of the value producing process. The activity of a single click-worker classifying images to train an image recognition system, like the activity of an individual worker in an assembly line, do not produce any value considered alone. Value is produced only if the outcome of this activity is integrated with the outcome of many other worker's activity, e.g., contributing to form a database big enough to train a system that will be sold for a high price or that will successfully recommend ads. The same goes for users' navigation data. These, considered together, are actually extremely important for recommender systems, since automated predictions based, among other things, on collaborative filtering techniques, that is, in filtering patterns of rating or usage produced by users' interaction with the systems (Koren and Bell 2011).

[1] Also people reading magazines or watching TV are exposed to advertising. However, the specificity of recommended advertising on online platforms is microtargeting. On media that do not allow microtargeted advertising such as cable TV or magazines, it is still possible to target a specific group of customers – e.g., young readers of an indie music magazine, people interested in new furniture for their home reading an interior design magazine, or children watching cartoons on TV on Sunday morning – but all the readers and viewers, including those who are not potential customers, will see the same advertising. This fact affects how the price of advertising space is determined since in the micro-targeting scenario this can be related to the number of potential customers reached and to the actual clicks on the ad, while in the other case it is related, among other things, to the overall visibility of the medium and its prominence for the target audience.

In order to understand how value is distributed among stakeholders, it is also necessary to make an estimate of the value that is given to digital laborers in exchange for their value producing activity. However, this is not an easy task either since the value given to digital laborers in exchange for work is not always a monetary one. Following the above-mentioned examples, three categories of laborer can be distinguished based on the quality and quantity of value corresponding to them. The first is constituted by the waged laborer paid money to complete tasks. The other two derive from a sub-classification of the group of unwaged laborers, that is, those supplying "consumer labor" (Jarrett 2015). On the one hand, there are those using a service for free – who are very likely to be targeted with ads, producing value both individually and in synergy with other users. On the other hand, there are those paying a subscription fee to access a service – who usually have access to more exclusive content and whose navigation might be ads-free, making them produce less or no value individually, but still being involved in synergic value production. Indeed, even though they are corresponding the monetary value for the service they get to the system provider, they still supply digital labor since their ratings and navigation data are necessary for the functioning of the recommender system, especially if running on collaborative filtering. I'll call the first group "subscription-free laborers" and the second group "subscriber laborers". I will use the label "consumer laborers" to refer to both.

Contrary to the waged laborer, consumer laborers do not receive any money. They are given something else in exchange for their value-producing activity and, in the case of subscriber laborers, for their subscription fee. What they get is namely what is keeping them hooked to their online habits (Eyal and Hoover 2019): interesting information, entertaining content, exciting networking opportunities in work and private life, etc. Moreover, in the case of social media, hosting and spreading UGC is part of the service. Content recommendations are valuable if they are relevant, that is, if a user finds the recommended content interesting, entertaining, etc. While it is possible to measure the relevance in terms of accuracy from a statistical point of view, for example comparing predicted ratings of an item with the actual user ratings (Doshi 2018), measuring relevance from the subjective point of view of a user and quantifying its value for them poses a challenge. In the case of a subscriber, the willingness to keep paying a subscription fee to receive content recommendations from a certain provider instead of looking for other options or cancelling the service, might be taken as an indicator that the service is worth at least the amount of the subscription fee to the user.[2] Users' subscriptions is a successful business model for many content recommending platforms such as Netflix and Spotify,

[2] In general, and not limited to the recommender system scenario, this consideration does not apply to services that are strictly necessary to the user and/or to which there are no alternatives, such as paid software or network subscription necessary to carry out a work task, and/or whose fee payment is compulsory, such as the broadcasting contribution for public radio and television in some countries. In these cases, users have to pay the fee even if they don't value the service. However, this is unlikely to be the case for many application fields of recommender systems and, at least nowadays, definitely not the case for the field of entertainment.

and ads-free, premium versions of dating apps like Tinder and Bumble, of social networks like LinkedIn, and of video broadcasting platforms such as Youtube, which crossed the 50 M subscribers threshold in 2021 (blog.youtube 2022). Moreover, social networks such as Instagram and TikTok are exploring subscription models to individual influencers' accounts (Dutta 2022; TikTok 2022).

Also, in the case of subscription-free laborers, the willingness to start paying a fee could be taken as an indicator of the value recommended content has for them. How much should the fee be, and should it be the same for all users? Let's consider the case of Meta (formerly Facebook Inc.) social networks. If Meta wanted to provide ads-free navigation, they should earn their revenue by collecting subscription fees. In 2020 their average revenue per user (ARPU) was 32.02 USD (Dixon 2022b). Considering the geographical distribution of the revenue adds key information: in 2020 Facebook Inc. ARPU was 163.86 $ in the US and Canada, 50.95 $ in Europe, 13.77 $ in Asia and Pacific, and 8.76 $ in the rest of the world (Dixon 2022a). Requiring a 32 $/year subscription fee would therefore be inadequate for two reasons: first, it would not reflect the regional revenue of the company; second, it would not consider the different average income around the world. Accordingly, the subscription fee should be adapted to the regional average ARPU and national average income. Considering that the Meta group profits are growing – 40.96 $ global ARPU in 2021 (Dixon 2022b) and 98.54 $ in the first half of 2022 just in US and Canada (Dixon 2022a) –, and that these averages include inactive users, making it reasonable to suppose that excluding inactive users from the count would increase the sum significantly, would an average US active user be ready to pay around 200$/year to be on the Meta Group social networks?

Since the answer to this question depends on the individual degree of appreciation of the service and on the impact of the fee on the individual income, it is not possible to provide a general answer. Some surveys show that the majority of participants would still prefer an ads-based business model to a subscription-based one (Hutchinson 2020; Sindermann et al. 2020). However, this might depend not only on satisfaction with the service, but also on the affordability of the fee. Imagine a low-earning person having to pay a monthly fee for every digital service they use: one for the Meta group, one for the Google-Alphabet group, one for Spotify, one for Netflix, one for Amazon, and so on. The digital services bill would easily go over 100$/month. If every digital platform – including search engines – added a subscription fee, a consequence would be that low-income people would have to quit using some services and experience digital exclusion. This would also not be in the economic interest of tech companies since they would lose users and revenue in this way. Since just a smaller part of the users would be interested in and could afford to pay a subscription fee to use services that are free right now, for tech companies to keep increasing their profits – which is the companies' goal in the capitalist economic framework – exploiting consumer labor by monetizing their online time must belong structurally to their business model. Unless it was against the law, of course.

The considerations on the different worth of produced value and corresponded value of different kinds of digital labor presented in this and in the previous section can be summed up as follows:

1. Waged laborers are likely to be paid poorly, often below minimum wage, even though their work is indispensable to train and develop software commercialized for thousands of dollars to many customers and essential to run systems at the base of million dollars' worth businesses.
2. Consumer laborers produce essential data for collaborative filtering, contributing to make predictions more accurate. Without processing their navigation data, recommender systems could not work at all. This happens without any economic return and irrespective of whether the laborer is paying a subscription or not.
3. Subscription-free laborers' online time is monetized, among other things, through targeted advertising. It is arguable whether the value of the content recommendation received in exchange is worthy for the laborers as the monetization of their time is for tech companies and depends on the individual case. While some people might find this exchange fair or just don't mind the fact that their data are being further processed, all those not valuing the received content recommendations enough are involved in an unbalanced exchange.

It is now possible to highlight several structural fairness issues related to the value production of different kinds of digital labor and to their redistribution. Starting from the mere consideration of the exchanged value, an ethical issue concerning the unfair distribution of the generated economic value immediately stands out. On the one hand, the very low pay of the waged laborer allows system providers to increase their profits, which they partially redistribute to a reduced number of high-earning executives, software developers, marketing and communication managers, etc., without rewarding those laborers whose work fuels ML-systems in the first place.[3] On the other hand, consumer laborers are supplying essential navigation data for free, which, processed together, allow the functioning of recommender systems. Some subscription-free laborers might believe that they are getting valuable content in exchange for that, but they are already being targeted with ads in exchange for content recommendation, which makes the additional, synergic way of producing value come on top. Only some subscription-free laborers might find recommended content so valuable to think it fair to supply the system provider with so much digital labor.

Focusing specifically on waged laborers, another issue is related with their unfair work conditions. On the one hand, workers are poorly paid, which in the case of platforms crowdsourcing work from independent contractors such as Amazon Mechanical Turk is usually below minimum wage (Irani 2015) and could amount to one or two dollars an hour (Milland 2019; Newman 2019). This pushes gig workers to extend their working days and be always available for new gigs to make a living wage. Moreover, they usually don't have job benefits such as health or social

[3] Of course, these considerations on the unfair redistribution of value, as well as those regarding poor working conditions, also apply to many other industries and were at the center of Marx' critique of labor after the first industrial revolution (Fuchs and Fisher 2015). As mentioned above concerning consumer labor, the specificity of our case resides in the fact that the unfairly distributed surplus value comes from data production and processing.

insurance (Sawafta 2019). On the other hand, work conditions are physically and mentally extenuating. Contract-workers have very few short breaks and are constantly monitored (Newton 2019). Content moderators and content taggers are exposed to sensitive content that is explicit and/or offensive and left without help to alleviate eventual trauma cause by viewing this content (Newton 2019; Sawafta 2019). These work conditions put workers' health at risk and are far from guaranteeing any financial stability or work-life balance. Finally, since most tasks can be executed remotely, digital labor represents a case of work outsourcing and offshoring, and can be seen as part of the larger phenomenon of "algorithmic coloniality" or "data colonialism" (Mohamed et al. 2020), meaning that inhabitants of former colonies are still affected by certain oppression and exploitation patterns they used to be subjected to during the colonial time. In concrete terms, users in western countries usually benefit from the result of underpaid digital labor in Africa, Latin America and Southeast Asia (Anwar and Graham 2019; Rani and Furrer 2021).

Focusing on consumer labor, systems providers' limited transparency and/or users' unawareness concerning the use of their navigation data puts many users in a disadvantaged position in the stakeholders' group since they lack the resources to control what data they provide and how this is monetized, and therefore to defend their privacy. The fact that many digital services are controlled *de facto* by monopolies or oligopolies of big players – owning the most used platforms and being able to develop more performing software because of the larger amount of training data they have access to – further diminished the bargaining power of single users involved in an unfair exchange loop. On top of this, the discrimination issues mentioned at the beginning (Barocas et al. 2019; Burke 2017; Milano et al. 2020) also apply to the consumer laborer since the content recommendation they get in exchange for their digital labor and/or their subscription fees might be inaccurate and biased against some user groups.

Considering the unfair treatment and the poor working conditions of digital laborers, the label "digital proletariat" was used by several authors to highlight the analogy with the factory working class during the industrial revolution (Gabriel 2020; Jiménez González 2022; The Economist 2018). As well as at the beginning of the industrial age, the absence of laws and regulations protecting laborers' rights led to their exploitation. The acknowledgment of the fairness issues concerning digital labor calls for more laborers' rights and for tailored solutions to tackle the problem in many application fields – including recommender systems.

5.4 Addressing the Problem

As shown in Sect. 5.2, digital labor is a structural issue in recommender systems. That means that those fairness issues related to it, like the case of discrimination issues rooting in structural injustice, cannot be simply addressed through measures aiming at solving the problem by correcting code or datasets – in other words, cannot be solved by computer scientists and software developers alone since the problem is not just a computational issue (Balayn and Gürses 2021; Kasirzadeh 2022). To address structural injustice, the whole economic and socio-political frameworks surrounding the examined unjust interactions should be considered in order to understand how responsibility for the generated disadvantages is distributed (Young 2011). In the specific case of digital labor in recommender systems, we have seen that the absence of a specific legal framework to regulate new forms of work, combined with the launch of new digital platforms whose ability to succeed in a capitalist market depends on the collection of cheap data in a large amount and short times, and with the great demand of recommendations for digital content, facilitated the establishment of labor exploitation practices. Acknowledging this fact and acknowledging the existence of digital labor is the necessary first step towards a fairer treatment of those stakeholders being now disadvantaged.

Focusing on the unfair distribution of produced value, some observers suggest the introduction of a "data dividend", that is, a share of the worth generated by data processing to be paid by tech companies to the users (Feygin et al. 2019). A similar proposal is the introduction of a digital basic income – to be funded through higher taxation of tech companies – to compensate users for their digital labor and for the negative consequences that the platform and gig economies are having on the job market (Ferraris 2018, 2021). Even though the will to fairly redistribute value shared by these approaches is well intended, it can be objected that these solutions do not address fairness issues at the root and might even raise additional concerns. Indeed, getting financial compensation for digital labor – whatever form this compensation takes – can be seen as an incentive for users to sell off their privacy and other basic rights. Moreover, this would raise further questions concerning who will determine the amount of the data dividend. Given the large number of data dividend beneficiaries (potentially the whole world population), even the multi-billion-dollar revenues of the biggest tech industries would end up being split in single-digit shares (Tsukayama 2020). This cannot represent a form of compensation for giving up privacy as a basic right or fair work conditions – which are rather priceless – nor can it smooth out the social inequalities accentuated by digital labor. Therefore, cash flow from tech-companies to users will not solve fairness issues.

Instead of compensating digital laborers for generating useful data, laborer exploitation should not occur in the first place. Concerning waged workers, workers' rights – e.g., as formulated in Art. 23 and 24 of the Universal Declaration of Human Rights (United Nations 1948) or in Art. 6 and 7 of the International Covenant on Economic, Social and Cultural Rights (OHCHR 1976) – should be respected. The effective hourly pay should allow a decent standard of living and in no case be below minimum wage; work conditions should not put workers' physical or mental health at risk; rest, leisure and limitation of working hours, as well as periodic holiday with pay, should be guaranteed; workers should have the right to form and join trade unions and the right to strike. This can be achieved by acknowledging waged digital labor as work and regulating this as such. Also, the exploitation of gig workers masked as "independent contractors" should be subject to stricter regulation aimed at guaranteeing respect of human and workers' rights.[4] In general, supply chain laws like the one entered in to force in Germany of January 1st 2023 (BAFA 2023) or the one being currently drafted by the European Commission (European Commission 2022), requiring large companies to investigate their supply chains, to identify corporate social responsibility risks, and to take appropriate action when risks for the environment or for human rights are discovered, represent a powerful tool to fight workers' rights violations and should be applied to the digital labor market as well. Furthermore, for workers to stand united for their rights and gain bargaining power when it comes to negotiating work conditions with big corporations, gig workers and digital laborers' trade unions are needed.

Concerning consumer labor, transparency about data collection and processing should be guaranteed.[5] Supplying system providers with data that are not strictly necessary for the functioning of the system should be the result of a visible opt-in choice and not a default setting that can be opted out of. Users should be able to decide in an informed way whether they want to supply additional data in exchange for a service[6] and the actual form of the terms of service of many digital services do not allow this kind of decision since they are too long and hardly comprehensible (Patar 2019), and users are nudged to accept them without reading them (Berreby 2017). It should be made clear in a brief and understandable way for the average user which data are necessary for the functioning of the recommender system and whether the system providers intend to further process this data for training (or other) purposes for users to make an informed decision on accepting the service conditions. Indeed, while system providers must always comply to applicable law concerning transparency and data protection and regulations should address the application scenarios that put user rights and societal wellbeing at risk, when a system is released, it is ultimately up to individual users to decide whether the service is worth the data (eventually coming on top of a service fee) that are lawfully asked in exchange.

5.5 Conclusion

This chapter has shown that the fast development of work in digital environments met a regulatory gap that allowed different forms of labor exploitation. This constitutes a structural fairness issue for recommender systems since their functioning and their success in the market depends on digital labor and on the unjust practices connected with it. To tackle this problem, digital labor must be acknowledged for

[4] For challenges to the existing law due to virtual work see Haupt and Wollenschläger (2001).

[5] In the European legal framework, the General Data Protection Law (GDPR) addresses these issues. See also the contribution by Levina and Mattern in Section III of this volume for a discussion of transparency and data protection issues.

[6] Concerning the transfer of personal data in exchange for digital content and digital services in current European directives, cf. Kaesling (2021); Latte (2021).

what it is, and future regulations must counter unfair treatment of digital laborers. This will require a structural change of the value production and redistribution processes, remunerating waged workers with adequate pay, granting proper work conditions, and giving to users a real choice when it comes to the exchange of data for services.

References

Anwar, Mohammad Amir, and Mark Graham. 2019. Digital Labour at Economic Margins: African Workers and the Global Information Economy. *Review of African Political Economy*.

Aytes, Ayhan. 2013. Return of the Crowds. Mechanical Turk and Neoliberal States of Exception. In *Digital Labor: The Internet as Playground and Factory*, ed. Trebor Scholz. New York: Routledge.

BAFA. 2023. Information on the Supply Chain Act. https://www.bafa.de/DE/Lieferketten/Multilinguales_Angebot/multilinguales_angebot_node.html. Accessed 17 Mar 2023.

Balayn, Agathe, and Gürses, Seda. 2021. *Beyond Debiasing: Regulating AI and its inequalities*. https://edri.org/wp-content/uploads/2021/09/EDRi_Beyond-Debiasing-Report_Online.pdf.

Barocas, Solon, Moritz Hardt, and Arvind Narayanan. 2019. *Fairness and Machine Learning*. fairmlbook.org.

Berreby, David. 2017. Click to agree with what? No one reads terms of service, studies confirm. https://www.theguardian.com/technology/2017/mar/03/terms-of-service-online-contracts-fine-print. Accessed 24 Sept 2022.

blog.youtube. 2022. 50 million. https://blog.youtube/news-and-events/50-million/. Accessed 4 Sept 2022.

Bouquin, Stephen. 2020. "Il n'y a pas d'automatisation sans micro-travail humain" – Grand entretien avec Antonio A. Casilli. *Les Mondes du travail* 24–25: 3–21.

Burke, Robin. 2017. Multisided Fairness for Recommendation. https://arxiv.org/pdf/1707.00093.

Casilli, Antonio A., and Julian Posada. 2019. The Platformization of Labor and Society. In *Society and the Internet: How Networks of Information and Communication are Changing our Lives*, ed. Mark Graham and William H. Dutton, 293–306. Oxford: Oxford University Press.

Dastin, Jeffrey. 2018. Amazon scraps secret AI recruiting tool that showed bias against women. https://www.reuters.com/article/us-amazon-com-jobs-automation-insight-idUSKCN1MK08G. Accessed 29 Aug 2022.

Deldjoo, Yashar, Vito Walter Anelli, Hamed Zamani, Alejandro Bellogín, and Tommaso Di Noia. 2021. A Flexible Framework for Evaluating User and Item Fairness in Recommender Systems. *User Modeling and User-Adapted Interaction* 31 (3): 457–511. https://doi.org/10.1007/s11257-020-09285-1.

Dixon, S. 2022a. Facebook: Average Revenue per User Region 2022. Statista. https://www.statista.com/statistics/251328/facebooks-average-revenue-per-user-by-region/. Accessed 6 Sept 2022.

———. 2022b. Meta ARPU 2021. Statista. https://www.statista.com/statistics/234056/facebooks-average-advertising-revenue-per-user/. Accessed 6 Sept 2022.

Doshi, Neerja. 2018. Recommendation systems – Models and evaluation – Towards data science. https://towardsdatascience.com/recommendation-systems-models-and-evaluation-84944a84fb8e. Accessed 4 Sept 2022.

Dutta, Soumyadip. 2022. Instagram paid subscription: Price, Who is eligible, how to enable, how to subscribe. https://www.techbloat.com/instagram-paid-subscription-price-who-is-eligible-how-to-enable-how-to-subscribe.html. Accessed 4 Sept 2022.

European Commission. 2022. EUR-Lex – 52022PC0071 – EN – EUR-Lex. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A52022PC0071. Accessed 17 Mar 2023.

Eyal, Nir, and Ryan Hoover. 2019. *Hooked: How to Build Habit-Forming Products*. New York: Portfolio/Penguin.

Ferraris, Maurizio. 2018. Salario di mobilitazione: un'idea per i nuovi poveri dell'era digitale. https://www.agendadigitale.eu/cultura-digitale/ferraris-salario-di-mobilitazione-unidea-per-i-nuovi-poveri-dellera-digitale/. Accessed 9 Sept 2022.

———. 2021. Webfare e libertà, se il "consumo" produce valore: ecco come e perché ricompensarlo. https://www.agendadigitale.eu/cultura-digitale/webfare-e-liberta-se-il-consumo-produce-valore-ecco-come-e-perche-ricompensarlo/. Accessed 9 Sept 2022.

Feygin, Yakov, Hecht, Brent, Prewitt, Matthew, Li, Hanlin, Vincent, Nicholas, Lala, Chirag, and Scarcella, Luisa. 2019. *A Data Dividend that Works: Steps Toward Building an Equitable Data Economy*. https://www.berggruen.org/ideas/articles/a-data-dividend-that-works-steps-toward-building-an-equitable-data-economy/.

Fisher, Eran. 2020. Audience Labour on Social Media: Learning from Sponsored Stories.

Fuchs, Christian, and Eran Fisher, eds. 2015. *Reconsidering Value and Labour in the Digital Age*. London/Imprint: Palgrave Macmillan UK/Palgrave Macmillan.

Gabriel, Markus. 2020. *Fiktionen*. Berlin: Suhrkamp Verlag.

Giovanola, Benedetta, and Simona Tiribelli. 2022. Beyond Bias and Discrimination: Redefining the AI Ethics Principle of Fairness in Healthcare Machine-Learning Algorithms. *AI & Society*: 1–15. https://doi.org/10.1007/s00146-022-01455-6.

Gray, Mary, and Siddharth Suri. 2019. *Ghost Work: How Amazon, Google, and Uber are Creating a New Global Underclass*. Boston: Houghton Mifflin Harcourt Publishing.

Grgić-Hlača, Nina, Muhammad Bilal Zafar, Krishna P. Gummadi, and Adrian Weller. 2018. Beyond Distributive Fairness in Algorithmic Decision Making: Feature Selection for Procedurally Fair Learning. *Proceedings of the AAAI Conference on Artificial Intelligence* 32 (1). https://doi.org/10.1609/aaai.v32i1.11296.

Haupt, Susanne, and Michael Wollenschläger. 2001. Virtueller Arbeitsplatz – Scheinselbständigkeit bei einer modernen Arbeitsorganisationsform. *NZA* 6: 289–296.

Hutchinson, Andrew. 2020. Would people pay to use social media platforms to avoid data-sharing? [Infographic]. https://www.socialmediatoday.com/news/would-people-pay-to-use-social-media-platforms-to-avoid-data-sharing-info/575956/. Accessed 6 Sept 2022.

Irani, Lilly. 2015. Justice for "data janitors" – Public books. https://www.publicbooks.org/justice-for-data-janitors/. Accessed 7 Sept 2022.

Jarrett, Kylie. 2015. *Feminism, Labour and Digital Media: The Digital Housewife*. London: Routledge.

Jiménez González, Aitor. 2022. Law, Code and Exploitation: How Corporations Regulate the Working Conditions of the Digital Proletariat. *Critical Sociology* 48 (2): 361–373. https://doi.org/10.1177/08969205211028964.

Kaesling, Katharina. 2021. § 327 BGB Anwendungsbereich. In *jurisPK-BGB 9. Aufl.*, ed. Herberger, Martinek, Rüßmann, Weth, and Würdinger.

Kasirzadeh, Atoosa. 2022. Algorithmic Fairness and Structural Injustice: Insights from Feminist Political Philosophy. In AIES'22: Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society: August 1–3, 2022, Oxford, 349–356. AIES'22: AAAI/ACM Conference on AI, Ethics, and Society, Oxford United Kingdom. 19.05.2021, 21.05.2021. New York: The Association for Computing Machinery. https://doi.org/10.1145/3514094.3534188.

Koren, Yehuda, and Robert Bell. 2011. Advances in Collaborative Filtering. In *Recommender Systems Handbook*, ed. Francesco Ricci, Lior Rokach, Bracha Shapira, and Paul B. Kantor, 145–186. Boston: Scholars Portal.

Latte, Simona. 2021. Personal Data in Exchange for Digital Content and Digital Services: Directive 2019/770/EU. *European Journal of Privacy Law & Technologies*.

Lee, Michelle Seng Ah., and Luciano Floridi. 2021. Algorithmic Fairness in Mortgage Lending: From Absolute Conditions to Relational Trade-offs. *Minds and Machines* 31 (1): 165–191. https://doi.org/10.1007/s11023-020-09529-4.

Maxwell, Richard, ed. 2016. *The Routledge Companion to Labor and Media*. New York/London: Routledge.

Mehrabi, Ninareh, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. 2021. A Survey on Bias and Fairness in Machine Learning. *ACM Computing Surveys* 54 (6): 1–35. https://doi.org/10.1145/3457607.

Milano, Silvia, Mariarosaria Taddeo, and Luciano Floridi. 2020. Recommender Systems and Their Ethical Challenges. *AI & Society* 35 (4): 957–967. https://doi.org/10.1007/s00146-020-00950-y.

———. 2021. Ethical Aspects of Multi-stakeholder Recommendation Systems. *The Information Society* 37 (1): 35–45. https://doi.org/10.1080/01972243.2020.1832636.

Milland, Kristy. 2019. From bottom to top: How Amazon mechanical Turk disrupts employment as a whole – Brookfield Institute for Innovation + Entrepreneurship. https://brookfieldinstitute.ca/from-bottom-to-top-how-amazon-mechanical-turk-disrupts-employment-as-a-whole/. Accessed 7 Sept 2022.

Mohamed, Shakir, Marie-Therese Pong, and William Isaac. 2020. Decolonial AI: Decolonial Theory as Sociotechnical Foresight in Artificial Intelligence. *Philosophy & Technology* 33 (4): 659–684. https://doi.org/10.1007/s13347-020-00405-8.

Newman, Andy. 2019. I found work on an Amazon website. I made 97 cents an hour. https://www.nytimes.com/interactive/2019/11/15/nyregion/amazon-mechanical-turk.html?mtrref=undefined&gwh=27F128B69E3C4E8334EA5840428F3472&gwt=pay&assetType=PAYWALL. Accessed 7 Sept 2022.

Newton, Casey. 2019. The secret lives of Facebook moderators in America. https://www.theverge.com/2019/2/25/18229714/cognizant-facebook-content-moderator-interviews-trauma-working-conditions-arizona. Accessed 1 Sept 2022.

OHCHR. 1976. International Covenant on economic, social and cultural rights. https://www.ohchr.org/en/instruments-mechanisms/instruments/international-covenant-economic-social-and-cultural-rights. Accessed 9 Sept 2022.

Patar, Dustin. 2019. Most online ‘terms of service’ are incomprehensible to adults, study finds. https://www.vice.com/en/article/xwbg7j/online-contract-terms-of-service-are-incomprehensible-to-adults-study-finds. Accessed 24 Sept 2022.

Rani, Uma, and Marianne Furrer. 2021. Digital Labour Platforms and New Forms of Flexible Work in Developing Countries: Algorithmic Management of Work and Workers. *Competition & Change* 25 (2): 212–236. https://doi.org/10.1177/1024529420905187.

Sawafta, Sara. 2019. The working conditions of digital workers in Amazon mechanical Turk. https://medium.com/@ssara1977/the-working-conditions-of-digital-workers-in-amazon-mechanical-turk-dedb2022539a. Accessed 7 Sept 2022.

Scholz, Trebor, ed. 2013. *Digital Labor: The Internet as Playground and Factory*. New York: Routledge.

Sindermann, Cornelia, Daria J. Kuss, Melina A. Throuvala, Mark D. Griffiths, and Christian Montag. 2020. Should We Pay for Our Social Media/Messenger Applications? Preliminary Data on the Acceptance of an Alternative to the Current Prevailing Data Business Model. *Frontiers in Psychology* 11: 1415. https://doi.org/10.3389/fpsyg.2020.01415.

Solans, David, Francesco Fabbri, Caterina Calsamiglia, Carlos Castillo, and Francesco Bonchi. 2021. Comparing Equity and Effectiveness of Different Algorithms in an Application for the Room Rental Market. In Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society, 978–988. AIES’21: AAAI/ACM Conference on AI, Ethics, and Society, Virtual Event USA. 19.05.2021, 21.05.2021. New York: Association for Computing Machinery. https://doi.org/10.1145/3461702.3462600.

Terranova, Tiziana. 2000. Free Labor: Producing Culture for the Digital Economy. *Social Text* 18 (2): 33–58.

The Economist. 2018. Should Internet Firms Pay for the Data Users Currently Give Away? *The Economist,* January 11.

TikTok. 2022. Exploring New Ways for Creators to Build Their Community and be Rewarded with LIVE Subscription. *TikTok,* May 23.

Tinder Newsroom. 2022. Powering Tinder® – The method behind our matching. https://www.tinderpressroom.com/powering-tinder-r-the-method-behind-our-matching/. Accessed 24 Sept 2022.

Tsukayama, Hayley. 2020. Why getting paid for your data is a bad deal. https://www.eff.org/it/deeplinks/2020/10/why-getting-paid-your-data-bad-deal. Accessed 9 Sept 2022.

United Nations. 1948. Universal Declaration of Human Rights. United Nations. https://www.un.org/en/about-us/universal-declaration-of-human-rights. Accessed 9 Sept 2022.

Vercellone, Carlo. 2020. Les plateformes de la gratuité marchande et la controverse autour du Free Digital Labor: une nouvelle forme d'exploitation? *Revue ouverte d'ingénierie des systèmes d'information* 1 (2). https://doi.org/10.21494/ISTE.OP.2020.0502.

Verma, Sahil, and Julia Rubin. 2018. Fairness Definitions Explained. In FairWare 2018: 2018 ACM/IEEE International Workshop on Software Fairness: proceedings: 29 May 2018, Gothenburg, Sweden, 1–7. ICSE'18: 40th International Conference on Software Engineering, Gothenburg Sweden. 29.05.2018, 29.05.2018. Los Alamitos: IEEE Computer Society. https://doi.org/10.1145/3194770.3194776.

Young, Iris Marion. 2011. *Responsibility for Justice*. New York/Oxford: Oxford University Press.

Part II
Manipulation and Personal Autonomy

Chapter 6
Recommender Systems, Manipulation and Private Autonomy: How European Civil Law Regulates and Should Regulate Recommender Systems for the Benefit of Private Autonomy

Karina Grisse

Abstract Recommender systems determine the content that users see and the offers they receive in digital environments. They are necessary tools to structure and master large amounts of information and to provide users with information that is (potentially) relevant to them. In doing so, they influence decision-making. The chapter examines under which circumstances these influences cross a line and can be perceived as manipulative. This is the case if they operate in opaque ways and aim at certain decision-making vulnerabilities that can comprise the autonomous formation of the will. Used in that way, they pose a danger to private autonomy that needs to be met by law. This chapter elaborates where the law of the European Union already adequately addresses these threats and where further regulation is needed.

Keywords Manipulation · Private autonomy · Regulation · Digital Services Act · Unfair competition law

6.1 Introduction

Recommender systems select the content that is being displayed to platform users and thereby shape user's perception of available content, information, choice and – in a way – of the world. Some platforms consist almost exclusively of recommendations (Seaver 2019). There is a vivid debate about how recommender systems influence equality, societal discourse, polarization and democracy (Milano et al. 2020; Beam 2014; Susser et al. 2019a). Clearly, recommender systems do influence

K. Grisse (✉)
University of Cologne, Cologne, Germany
e-mail: kgrisse@uni-koeln.de

S. Genovesi et al. (eds.), *Recommender Systems: Legal and Ethical Issues*, The International Library of Ethics, Law and Technology 40,
https://doi.org/10.1007/978-3-031-34804-4_6

behavior (Calvo et al. 2020). Some concerns have been raised as to the manipulative potential of recommender systems and their negative effects for human autonomy.[1]

This chapter looks at recommender systems from a civil law perspective, more precisely, from the perspective of European Private Law and private autonomy. It explores the question of whether and when recommender systems and their recommendations manipulate recipients' decision-making in a commercial context. It further examines where the law of the European Union already prohibits such manipulative influences, where relevant regulations are emerging and where there is still a need for regulation.[2]

After giving a brief introduction to the role of private autonomy and private law's stance towards influence (6.2), the paper maps different recommendation settings relevant to the question asked above (6.3). It goes on to look at different philosophical concepts of manipulation (6.4). The non-legal concepts of manipulation serve to assess different features of recommender systems and under which circumstances recommender systems' influences should be considered manipulative (6.5). Finally, the paper examines which of the manipulative settings identified in Sect. 6.5 are already subject to regulation and which issues should be further regulated to safeguard the autonomous decision of recommendation recipients (6.6).

6.2 Autonomy and Influence in Private Law

The term private autonomy describes the right of individuals to shape their legal relationships according to their own will (Flume 1979; BVerfG NJW 1994, 36). The idea of it derives from the image of mankind as naturally free that is the basis for all human rights and freedoms.[3] Private autonomy is a fundamental principle of

[1] E.g. quite boldly: "Evidently, recommender systems deprive human users of liberty due to their controlling influences, and also often agency since human users do not usually provide informed consent when using recommender systems (users often lack the choice and are given a 'take it or leave it' option when accessing online services" (Varshney 2020); Calvo et al. have examined the potential influences of a recommender systems on human autonomy in spheres (levels) of life (Calvo et al. 2020; Ebers 2018; Mik 2016; Susser et al. 2019b).

[2] There is a more general debate on whether recommender systems decrease (e.g. because they cause humans to make less variant and diverse choices) or enhance overall human autonomy (e.g. by facilitating quick decision-making and saving time that can be used in a self-determined way), see for example Calvo et al. 2020. Albeit central, this question goes beyond the scope of this paper. The paper also does not examine legal problems concerning the relationship between platforms deploying recommender systems and those being recommended (either themselves, e.g. as employers on recruitment platforms or as potential partners on dating sites, or their products and services). These issues are being addressed by Regulation (EU) 2019/1150 of the European Parliament and of the Council of 20 June 2019 on promoting fairness and transparency for business users of online intermediation services (P2B Regulation). The paper is also not concerned with data protection issues regarding the collection of personal data.

[3] For an example of the natural law view on human beings, cf. Hobbes 1794; the idea of self-determination has its origin in the essence of man and his need for self-realisation: Busche 1999. On Kant's influences on the understanding of freedom, especially in private law: Schapp 1992.

European private law.[4] Within private law, this principle offers the basis for freedom of contract, including freedom of choice and the principle of will or intent (Study Group on a European Civil Code and Research Group on EC Private Law (Arquis Group) 2009: II. – 4:101 DCFR), i.e., that contracts are being formed because of a declared will. Private autonomy requires, on the one hand, that the state leaves citizens in principle free to shape their legal relationships (Busche 1999) and, on the other hand, that the state creates and secures conditions that enable them to exercise their rights (Study Group on a European Civil Code and Research Group on EC Private Law (Arquis Group) 2009). For constellations of obvious power imbalances, in which the stronger party can impose their will on the weaker, leaving no or little room for the weaker to exercise their autonomy, the state must limit one party's autonomy to protect at least a minimum of autonomy for the other (Möllers 2018; Busche 1999).

Because humans live together in societies and form legal relationships with each other, private autonomy can never exist in absolute terms. It must necessarily be limited in order to guarantee the rights and freedoms of others as well as certain (public) values and principles (e.g., personal responsibility, fairness, legal certainty and others. See Riesenhuber 2003). Social coexistence must therefore be regulated to a certain extent. Being a principle, private autonomy can only be realized to a certain degree (Riesenhuber 2018). The enjoyment of autonomy by one person in comparison to the autonomy of another or with regard to other principles is the result of a balancing exercise and largely a value judgement. The results of these balancing exercises are by no means set in stone.[5] Each generation must determine which values, policies and principles should be given priority to and, in situations of colliding individual autonomy, whose autonomy to strengthen and whose to limit.[6] Even though private autonomy is a legal concept, the decisions about its scope and level of legal protection (in the light of other values and legal principles) is mainly a political decision, hence one that may change over time or in the face of technical developments, and is open to extra-legal influences.[7]

An autonomous decision requires also that the decision-making process, the formation of the will, was sufficiently autonomous. European private law assumes that a decision is autonomous when it is informed[8] and free from certain types of

[4] This is true even if the conceptions and dogmatic justifications may be disputed in detail or vary in the different countries. In the DCFR, it is conceptualised as an underlying principle of European Private law, "Party autonomy" (Study Group on a European Civil Code and Research Group on EC Private Law (Acquis Group) 2009).

[5] Constitutions and fundamental rights do set certain limits but leave a wide margin for discretion.

[6] Cf. Bumke 2017: Autonomy in law must be rethought again and again.

[7] Cf. Röthel 2017: points to the importance of extra-legal concepts of autonomy for the positive understanding of private autonomy. Cf. also Hacker 2017; Also, Specht 2019.

[8] Real autonomous decisions require that the decision-maker knows what she is doing and that she can (to a certain degree) foresee the consequences of a wilful action or declaration: cf. e.g. Annex I Directive 2005/29/EC of the European Parliament and of the Council of 11 May 2005 concerning unfair business-to-consumer commercial practices in the internal market, No. 7; see also CJEU joined cases C-54/17 and C-55/17: Judgment of the Court (Second Chamber) of 13 September

influences (Cf. Study Group on a European Civil Code and Research Group on EC Private Law (Acquis Group) 2009: II. – 7:205 ff. DCFR).

When legal subjects communicate and establish legal relationships with each other, they (in principle legitimately) follow their own interests. They necessarily (try to) influence each other. Whoever wants to sell something needs to present the sales item in a good light. Whoever wants to conclude whatever kind of contract with someone else needs to convince the other to do so (Cf. also Köhler 2021: UWG § 1 Rn. 17). Obviously, not any form of influence on someone else's decision can count as an interference with the other's autonomy that is to be prevented or otherwise sanctioned by law. Usually, the following types of influences are considered undue influences: coercion, unlawful threat and deception (or fraud).[9] Where the conclusion of a contract was induced by such means, the contract can be voided.[10] European Unfair competition law prohibits commercial conduct that is contrary to the requirements of professional diligence, and is likely to materially distort consumers' economic behavior,[11] especially when it is misleading (deceiving)[12] or aggressive (harassing, coercing, or unduly influencing in a way that is at least likely to significantly impair the average consumer's freedom of choice or conduct).[13]

6.3 Recommender Systems and Their Influence

The EU Digital Services Act (DSA)[14] defines recommender systems as a "fully or partially automated system used by an online platform[15] to suggest in its online interface specific information to recipients of the service or prioritise that

2018, AGCM v Wind Tre SpA and Vodafone Italia SpA, para 45. There is a broad consensus today that knowledge is a prerequisite for autonomous decision-making (Bumke 2017). To enable informed decision making, European Private Law establishes numerous information obligations for situations in which one party typically has superior knowledge than a potential contractual partner.

[9] Including fraudulent non-disclosure of relevant information, II – 7:205 DCFR.

[10] E.g. § 123 BGB (D) and articles 1130, 1131 Code civil (FR); cf. also II. – 7:205 DCFR. II. – 7:206 DCFR.

[11] Article 5 (1) + (2) Directive 2005/29/EC of the European Parliament and of the Council of 11 May 2005 concerning unfair business-to-consumer commercial practices in the internal market and amending Council Directive 84/450/EEC, Directives 97/7/EC, 98/27/EC, and 2002/65/EC of the European Parliament and of the Council and Regulation (EC) No 2006/2004 of the European Parliament and of the Council (Unfair Commercial Practices Directive, here: UCP-D).

[12] Article 6 and 7 UCP-D.

[13] Article 8 and 8 UCP-D.

[14] 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), OJ L 277/1.

[15] The use by an online platform, however, is not really a defining criterion for a recommender system. This becomes obvious by article 38 DSA, which applies also to very large search engines.

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."[16] While recommender systems normally are programmed to display information or offers that are likely to be of interest for the recipient (a platform user) (Ricci et al. 2012), that does not mean that the predicted interest is necessarily the only reason behind a recommendation (Seaver 2019). Recommender systems exist in different forms and contexts. One can roughly[17] distinguish two categories of recommendations: related and unrelated recommendations.

'Related recommendations' are recommendations that are related to what the user is currently looking at or looking for. In this category fall recommendations that are being generated in reply to a specific user search query, e.g., the result list on Amazon when a user searches for a smartphone. The recommender system decides which offers are being displayed and in which order (ranking). Then there are recommendations that are related to a current user choice (maybe following a search query). For example, when a user clicks on one of the offered smartphones, the Amazon recommender displays similar items and/or accessories like display protections, earphones or cases. The newsfeeds in a social network made up of posts and shared content by "friends" or dynamic news websites are other examples of related recommendations. Recommender systems decide what content (which news or whose posts) to display, and these decisions sum up to lists of related recommendations: they are based on what the user is looking for when she visits a news page or social network.

'Unrelated recommendations' are recommendations that are not related to what the user is currently looking for. Many websites are financed through advertisement and recommender systems can be used to decide which advertisement is displayed to which user (Calo 2014). Those advertisements are usually unrelated recommendations, because the user is not looking for them but for other content on the website. That is the case e.g., of advertisements displayed in a social network's newsfeed or news websites that users consult to see what content "friends" have posted and shared or to read news. Another form of unrelated recommendations are those being displayed in more or less fixed categories on many homepages of sales platforms. The recommendations may either address every user[18] or may be personalized for a

The mentioning of platforms in the definition of article 3 lit. s DSA can be explained with the fact, that the DSA originally should only be applied to platforms. The application was later extended to very large search engines. The fact that platforms continue to be mentioned in the definition of recommender systems is likely due to an omitted editorial correction.

[16] Article 3 lit. s DSA.

[17] In some cases, the line might be blurry.

[18] E.g. the homepage of booking.com displays photos of some elected cities that the user can click on to get to accommodation offers in that city or under the headline "Homes guest love" displays a choice of accommodation options in different places.

specific user.[19] They might be triggered by earlier search queries of the user but are in any case not related to what the user is currently looking for. Unrelated recommendations are often a form of personalized or targeted advertising.

Recommendations can be based on or influenced by many different factors and filtering techniques (Ricci et al. 2012): Content-based recommender systems recommend items that are similar to those that a user has preferred in the past. Other recommender systems base their recommendations on demographic characteristics of the users. In "community-based"-filtering, recommendations are based on the preferences of the user's "friends". Collaborative filtering recommends items to the user that other users with similar preferences have liked in the past. These are just examples of common filtering techniques. Often, the different techniques are combined in hybrid models to overcome the weaknesses of some techniques.[20] Many recommendation techniques, e.g., basing recommendations on prior user behavior or demographic characteristic require some level of profiling.[21]

Recommender systems decide, in any case, which options are being brought to a user's attention. Recommender systems do not threaten or coerce users into any decisions. They are filtering tools that pick, usually from large pools of contents or offers, what to present to users. They influence the perception of available choices and thereby the users' choices. Depending on how recommender systems are integrated into platforms, they might steer user attention in a certain direction. Is this manipulative, as some have claimed?

6.4 Manipulation

To understand what is behind the claims that recommender systems are manipulative and why manipulation is undesirable, it is helpful to take a brief look at the philosophical literature on manipulation. In philosophical discussions, the term "manipulation" is usually[22] used to describe influences that do not respect the

[19] E.g. on the Amazon homepage, one finds a number recommendations sorted in categories. When a user is logged in, some recommendation categories are based on prior user behaviour (e.g. "Keep shopping for" or "Buy again"), while others are not and seem to address everyone (e.g. "Top Deal" or "Amazon devices").

[20] For example, collaborative filtering and community-based recommending cannot make statements about new products.

[21] Article 4(4) GDPR: "'profiling' means any form of automated processing of personal data consisting of the use of personal data to evaluate certain personal aspects relating to a natural person, in particular to analyse or predict aspects concerning that natural person's performance at work, economic situation, health, personal preferences, interests, reliability, behaviour, location or movements."

[22] Sometimes, however, the term is used in a broader sense. For example, Faden and Beauchamp use the term to describe any influence that is neither coercive nor persuasive and distinguish between manipulative influences that are controlling (incompatible with the autonomy of the influenced) and those manipulations that are non-controlling (compatible with the other's autonomy) (Faden and Beauchamp 1986).

autonomy of the influenced person (Susser et al. 2019b; Raz 1988). They are perceived as immoral or unfair precisely because they are thought to disrespect and harm the other's autonomy (Sunstein 2016; Raz 1988; Susser et al. 2019b).

The value of autonomy is mostly undisputed in the western world (Rössler 2017; Raz 1988). Kant attributed the unconditional value of human dignity to human autonomy.[23] According to self-determination theory in psychology, autonomy is one of three basic psychological human needs (Ryan and Deci 2017). While the lack of it negatively influences health and wellbeing, autonomy improves human energy and motivation (Ryan and Deci 2017, 2000; Deci and Ryan 2008). Philosophers have understood autonomy as a necessary (though not sufficient) condition for a felicitous life (Rössler 2017; Dworkin 1988; Raz 1988). Individual autonomy is also an essential presupposition of democracy. "It is only because we believe individuals can make meaningfully independent decisions that we value institutions designed to register and reflect them" (Susser et al. 2019a). Autonomy is also recognized as a value from a more utilitarian perspective, and was for instance supported with the argument that economic systems based on self-determination have so far been the most successful systems for increasing general welfare (Lobinger 2007 arguing that because the improvement of one's own living conditions regularly sought through autonomous action usually only succeeds if the needs of others are also satisfied). Whether it is thought in an instrumental way or not, autonomy is a value and manipulation is problematic because it is incompatible with this value.

From the perspective of autonomy, many acts can count as manipulative (Sunstein 2016). The term "manipulation" describes a targeted influence to control the behavior of others,[24] the "handling or managing of persons (Harper 2022)." In that sense, coercing someone to do something, for instance, would be an act of manipulation. However, in the philosophical debate, and often in ordinary language, the word manipulation is usually used to describe a more distinctive type of influence that is neither coercive nor persuasive (Noggle 2020; Faden and Beauchamp 1986). There is a considerable number of attempts in the literature to find a unitary concept of manipulation.[25] As I am not concerned with regulating manipulation in general, but only with assessing whether certain features of recommender systems can count as manipulative, I don't need a conclusive definition of manipulation. It is sufficient to identify typical features of manipulative influences and criteria to distinguish unwelcome manipulative from benign i.e., non-manipulative influences.

[23] Cf. Kant 2016: For autonomy, the human will, which permits a moral self-legislation that (according to the principle of the categorical imperative) is at the same time suitable as a universally valid legislation, is the condition for man to be able to regard himself and others as ends in themselves and not merely as means.

[24] See "manipulation" in Cambridge Dictionary; cf. also "Manipulation" in *Digitales Wörterbuch der deutschen Sprache*, (accessed 13 January 2022).

[25] Sunstein expresses some doubt as to whether manipulation is a unitary concept after all and admits that his own account might not be exhaustive (Sunstein 2016); Ackerman describes manipulation as a term of combinatorial vagueness (with reference to Alston 1967) to which no enlisted condition is sufficient or necessary (Ackerman 1995).

Manipulative influences are often described as attempts of subverting rational deliberation or the capacity for conscious decision making, bypassing deliberation altogether or introducing non-rational influences in the deliberation process (Susser et al. 2019b; Noggle 2020; Faden and Beauchamp 1986). *Raz* claims that manipulation "perverts the way [a] person reaches decisions, forms preferences, or adopts goals" (Raz 1988). Drawing on the heuristic concept of the two systems of the mind (Kahneman 2011), in which system 1 is considered to be an "automatic, intuitive system, prone to biases and to the use of heuristics, while System 2 is more deliberative, calculative, and reflective", *Sunstein* says that manipulators usually address system 1 and try to bypass system 2 (Sunstein 2016). *Susser/Rössler/Nissenbaum* speak of targeting and exploiting decision-making vulnerabilities (Susser et al. 2019a; cf. also Spencer 2020).

Only the introduction of non-rational influence cannot be sufficient to label an influence as manipulative. Many non-rational influences seem to be perfectly acceptable with a view to autonomy and are commonly considered benign (e.g., using perfume and dressing up for a date) (Noggle 2020; Sunstein 2016). *Sunstein* therefore suggests that an effort to influence someone's choice should only count as manipulative when "it does not *sufficiently* engage or appeal to their capacity for reflection and deliberation" (emphasis added) (Sunstein 2016). This sufficiency criterion is context sensitive and allows one to take a number of other factors into account that seem to be relevant when judging whether an act or conduct is manipulative, e.g., the particularities of the context, the role of the influencer and the relationship with the influenced (cf. Sunstein 2016).[26]

Another attempt to distinguish between manipulative and non-manipulative influences builds on the idea that manipulators intend to make others fall short of ideal behavior. According to *Noggle*, the common feature of manipulative acts is that they try to lead the victim astray (Noggle 1996). He claims, there are certain ideals to which we strive for when we form our beliefs, develop emotions and desires – such as the ideal that one should believe what is true, or focus on what is relevant and have emotions that are appropriate to given situations (Noggle 1996). *Noggle* observes that "[m]anipulative action is the attempt to get someone's beliefs, desires, or emotions to violate these norms, to fall short of these ideals" (Noggle 1996). In his view, the ideal setting is not to be determined by "what the influencer thinks are the ideal settings for the person being influenced" (Noggle 1996). Drawing on *Noggle's* concept, *Barnhill* suggests that manipulation is "directly influencing someone's beliefs, desires, or emotions such that she falls short of ideals for belief, desire, or emotion in ways typically not in her self-interest or likely not in her self-interest in the present context" (Barnhill 2014). In this account, what is crucial is not what the influencer thinks about the ideal settings for the influenced

[26] The sufficiency criterion has been criticised as normative and not helpful (cf. Susser et al. 2019b) and Noggle, who says that by saying that some forms of emotional appeals are not manipulative brings the problem of defining manipulation back to the start (Noggle 2020).

person, but whether his act or conduct would typically benefit or not the self-interest of the influenced (Susser et al. 2019b). This is a more objective and verifiable criterion.

Faden/Beauchamp suggest that an influence is manipulative when the influence is either difficult to resist or when it attempts to cause the influenced to fail to substantially understand his action, the circumstances or the consequences (Faden and Beauchamp 1986). They analyze, amongst others, cases of manipulation through offering. This is particularly interesting in our context because recommender systems display offers. *Faden/Beauchamp* suggest that, as a rule, a welcome offer made while the influencee "is not simultaneously under some different and controlling influence causing acceptance or rejection of the offer" is compatible with the autonomy of the influenced (Faden and Beauchamp 1986). An unwelcome offer is compatible with autonomy "if it can be reasonably easily resisted" (Faden and Beauchamp 1986). They judge offered "mere goods" to be usually easy to resist and "generally more compatible with autonomous action than […] harm-alleviating goods" (Faden and Beauchamp 1986).

Manipulation is typically (but not necessarily (Sunstein 2016; Barnhill 2014; Klenk 2021)) covert,[27] in the sense that the target of manipulation "is not conscious of the manipulator's strategy[28] while they are deploying it" and "couldn't easily become aware of were they to try and understand what was impacting their decision-making process" (Susser et al. 2019a). In many accounts, deception – i.e., covertly influencing the decision-making process by causing false beliefs in the victim – is one case of manipulation (Susser et al. 2019a, b; Faden and Beauchamp 1986; Noggle 1996, 2020).

Despite the (smaller or bigger) differences in the various conceptualizations of manipulation, there is general agreement in some regards: Manipulation can have many different forms[29] and can address different levers: someone's beliefs, desires, emotions, habits, or behaviors (Susser et al. 2019a, b).[30] Manipulative influence

[27] For Susser et al. covertness is even the defining feature that makes manipulation distinctive (Susser et al. 2019b): "Strictly speaking, the only necessary condition of manipulation is that the influence is hidden; targeting and exploiting vulnerabilities are the means through which a hidden influence is imposed." They argue that attempted covertness is crucial because when the decision-maker is aware of the influencers strategy that knowledge becomes part of the decision-making-process.

[28] What is typically hidden is not the "manipulative stimulus" but the "manipulative mechanism", Spencer 2020: "For example, an actor trying to exploit the anchoring bias must make the anchor visible to the subject. This anchor is the stimulus, and it cannot be hidden. What is hidden from the subject, however, is the manipulative mechanism – the cognitive process that drives her estimate toward the anchor."

[29] See for example Faden and Beauchamp 1986: Reducing or increasing options, making them more or less attractive, altering the understanding of a situation to modify the perception of options, and influencing "belief or behavior by causing changes in mental processes other than those involved in understanding," can be ways to manipulate.

[30] Noggle identifies three levers that a manipulator can "adjust": beliefs, emotions and desires (Noggle 1996).

(i.e., influence that does not respect the other's autonomy) is a controlling kind of influence (Faden and Beauchamp 1986; Noggle 1996);[31] manipulators treat people as "puppets on a string" (Wilkinson 2013; Sunstein 2016; Susser et al. 2019b).[32] *Sunstein* says: "[t]he problem of manipulation arises when choosers justly complain that because of the action of a manipulator, they have not, in a sense, had a fair chance to make a decision on their own" – that is, if due to the influence the decision was made without sufficiently assessing costs and benefits on the choosers' own terms (Sunstein 2016). For an act to count as manipulative it requires the intention to manipulate (Susser et al. 2019a; Spencer 2020; Noggle 1996; Faden and Beauchamp 1986).

Where the influence fails and the influencer does not achieve his goal one cannot say that someone has been manipulated, but the influencing act itself can still count as manipulative.[33] Even an unsuccessful act of attempted manipulative influence disrespects the other's autonomy. Considering this, it is convincing to say that an influence can be manipulative even when the influenced person would not have acted differently without the influence (Susser et al. 2019b; Calo 2014).

There remains one last question that needs to be answered here: Is manipulation always wrong? What if the manipulated person turns out to be very happy with the result of her manipulated choice? In *Barnhill's* account, an act would not count as manipulative when the influence aims at the best interest of the influenced (Barnhill 2014). This is not convincing from the perspective of autonomy (Sunstein 2016). Autonomy includes the freedom to make unreasonable decisions and to act in ways that are not in one's best interest (Raz 2009; Scanlon 1986).

But to be straightforward: Manipulation is not always wrong; it can be justified (Susser et al. 2019b). From a welfarist point of view, manipulation could be accepted if it makes the life of the manipulated or others better (for instance, Sunstein (2016) examines the welfarist perspective, pro and cons towards manipulation). There may be a convincing argument for this if the subject of the discussion is manipulation by the state's democratically legitimized organs acting in the public's interest (Thaler and Sunstein 2008) or maybe by parents who have a special duty to care for their children and to act in their best interest. However, if the question is about manipulation in impersonal[34] civil relations, this argument is questionable. In this realm, no

[31] Noggle 1996: "It's as though the manipulator controls his victim by 'adjusting her psychological levers'."

[32] Noggle writes: "The term "manipulation" suggests that the victim is treated as though she were some sort of object or machine" (Noggle 1996). Scanlon says: "An autonomous person cannot accept without independent consideration the judgment of others as to what he should believe or what he should do. He may rely on the judgment of others, but when he does so he must be prepared to advance independent reasons for thinking their judgment likely to be correct, and to weigh the evidential value of their opinion against contrary evidence" (Scanlon 1972).

[33] E.g. Susser et al. point to the fact that detecting real (successful manipulation) often would require "far-flung empirical findings that are difficult, if not impossible, to access" (Susser et al. 2019b). Wood 2014: manipulation is a "success concept".

[34] To leave out for example family relationships with duties of care. In such personal or intimate relationships, a different assessment may be in order in the individual cases.

one has a right to manipulate the other's decision, though one can try to influence it in many ways. It is also hard to imagine that in this context manipulation happens for the mere good of the manipulated person,[35] even though the person might be happy with her decision (either because the result is good for her or because she never learned about better options). Private autonomy is limited by the rights of others but not by other's opinions on what is good or in someone's interest.

6.5 Recommender Systems and Manipulation

6.5.1 Recommendations in General

Recommending something is always an attempt to take influence, but not necessarily an attempt to manipulate and to disrespect the other's autonomy. Filtering and ranking information is inevitable. Platform algorithms need to make a choice on what to display and in which order. Even if the filtering was not automatized, someone would need to fulfil this task in one way or another. Almost everything that we see in the world around us is based on a choice (hence a filtering) someone made. The shop owner needs to decide which goods to sell and where to place them, whether to highlight them etc. These decisions determine our options and how we perceive them. The mere fact that platforms use algorithms to fulfil the necessary task of deciding which options to present and which to highlight does not render the influence manipulative, even when some knowledge on the user plays a role in determining the recommender's choice.

As a principle and result of private autonomy, platform owners are free in what they offer and what they highlight. If it is a commercial platform, it is obvious to users that commercial interests determine the platforms' choices. The commercial interests do not need to clash with the user's interests. When, for example, a platform's recommender system manages to recommend relevant items to the user who consults the platform with the intention to conclude a contract of whatever kind, that benefits everyone. Recommendations can hardly be considered per se manipulative from the perspective of private autonomy. There are, however, certain aspects to consider.

A recommendation is in many situations understood as "a suggestion or proposal as to the best course of action."[36] It follows that the addressee usually has certain expectations about the recommended something (action, good, service …), e.g., that it is of special quality, has a good price or is of interest (Peifer 2021). One asks for

[35] "If they are genuinely concerned about the welfare of the chooser, why not try to persuade them?" asks Sunstein 2016. Cf. also Susser et al. 2019b; Mik 2016.

[36] See the entry "recommendation" in the Lexico *English Dictionary:* https://www.lexico.com/definition/recommendation

recommendations when one is looking for something that fulfils one's needs or that one is likely to enjoy. One expects the recommender to base the recommendation on the knowledge of the recommended item and its specific features. In some cases, one might also expect the recommender to consider the specific needs of the addressee or his or her taste.

A search query on a platform is in a sense asking for recommendations. When users search for something on a platform and receive a list of results, they would usually have the expectation that the ranking is determined by some kind of quality criteria and that the results displayed at the top of the list are there because they are particularly relevant to the search and/or of a particular good quality (e.g., price-performance ratio). However, the platform providers may have all kinds of reasons to display and recommend certain products and program their recommender systems accordingly. They might for example have a special interest in selling some products more than others, might have received money in return for a better ranking or might display items that are overpriced to make the user accept the average price easily, or the like.

If recommendations are opaquely dominated by other factors than those that a user would reasonably expect in a given situation, it can be considered misleading or manipulative. Ranking an item high up in the search results induces the belief that it would be good for the user to choose this item because it was ranked well. That applies even though users are generally aware that the platform operator is pursuing commercial purposes, because that usually does not foreclose that an offeror acts also in the interest of her customers. If the decisive motive to rank the item high is a different one, the user is being induced to form a wrong belief and is led to make a decision that may not be in her best interest.[37] The user is being deceived and manipulated, because they are led to base their decision on a wrong assumption. They are being led astray, to speak with *Noggle's* words, from the ideal to base a decision on true facts. If the unexpected recommendation-criteria are made transparent to the user, there is usually[38] no problem.

Pointing a user to other items that are somehow related to the one they were just looking at, does not raise concerns. Drawing a user's attention to other similar items so they can compare and make the best choice for them or to look at other items that could be additionally useful cannot be considered as leading them astray, because first, the additionally given information is not irrelevant nor is it typically against their interest to look at further items. It may even be autonomy-enhancing. Apart from this, additional offers are usually easy to resist, and there is nothing irrational or covert about the influence.

[37] Which would usually be buying a good and relevant product for a fair price.

[38] For exceptions, see below 6.5.3.2.2. *Exploring emotions*, and 6.5.3.2.3. *Addressing fears through (allegedly) harm-alleviating offers*.

6.5.2 Labelled Recommendations

In many cases, transparency regarding recommendation criteria will foreclose manipulation. But can the explicit reference to a recommendation criterion not be itself manipulative?

Often, recommendations (related or not) are based on experiences with other customers (Ricci et al. 2012). Some platforms label popular items as "bestseller" or headline a recommendation category as "popular on [platform name]", "other customers also bought" or alike. The labelling draws attention to what other people do. One can suspect that this draws on the bandwagon effect, the human tendency to follow the crowd (Thaler and Sunstein 2008). Can this practice be considered manipulative?

Even though the impulse to follow the crowd may not be purely rational, it is also not fully irrational. Of course, the bandwagon effect may reinforce the success of already successful products. However, the fact that others have opted for something is at least an indication that a product may be good or useful (in some way), if some of the buyers have informed themselves about the product, were happy with it and recommended it privately.[39] Bestseller lists are long established, especially in the music and book trade. If the information is true (i.e., that an item is a bestseller or that other people, who bought the item the user is currently looking at, also bought a certain accessory), providing it is neither deceiving nor manipulative in some other way. Even though the information is also addressing the less reflective system of thinking (system 1), the capacity for reflection and deliberation (system 2) is still sufficiently engaged, especially considering that it is a longstanding and accepted practice. Merely informing (in a non-deceptive way) about a fact can usually not be considered manipulative, even if people tend to react to the information given in a certain way (expected and desired by the influencer) (cf. Sunstein 2016).[40]

6.5.3 Unrelated Recommendations

6.5.3.1 In General

Unrelated recommendations address the user at a time when she is not looking for something, or rather, when she is looking for something else. They can divert the user's attention away from what she was originally interested in. Are they manipulative?

[39] Sunstein finds that information on what other people do can be part of reflective deliberation (Sunstein 2016).

[40] Also Susser et al. do not consider merely informational nudges as manipulative (Susser et al. 2019b).

Noggle and *Barnhill* would probably consider unrelated recommendations (and maybe most advertisement) as manipulative, because when a user is reading a news page or scrolling through his social media feed, the ideal for these situations would be to focus on the primary information sought and (ad-)recommendations that are placed in other contexts are intended to capture the attention and draw it away from the primary information.

However, unrelated recommendations are usually non-controlling. An advertisement that is not related to the context in which it is placed is socially accepted and, apart from its direct ends, serves the objective of financing services. This is not a new or special phenomenon in the online environment. Though most advertisements do not only address rational capacities for deliberation, it usually still sufficiently engages reflection. If one accepts *Sunstein's* sufficiency criterion, one must be open to acknowledge that the addressee's awareness of the role and purpose of the influencer matters. Seeing an advertisement that is recognizable as such, the viewer usually knows its motives and purposes. Unrelated recommendations are – welcome or not – offers that are generally easy to resist.[41] Internet users are accustomed to this practice and the brain is usually capable of filtering out information that it does not consider relevant in a given situation (Mik 2016) and that does not capture the attention due to special circumstances.[42] Of course, these kinds of recommendations may still leave a subconscious trace. Though this is a calculated effect, it is socially accepted and unrelated recommendations are not per se manipulative.

6.5.3.2 Targeted Recommendations

6.5.3.2.1 In General

Targeted unrelated recommendations are based on an analysis of the user's data. The aim is to target a user with advertisements or offers that are likely to be of interest to her in order to increase the chance that she is going to react to it.

Targeted recommendations are likely to be harder to resist than a random advertisement. If the recommender system manages to display something to the user that really is of interest to her, she will feel more tempted to click on the add.[43] However, as I have said above, as a matter of principle the platform owner is free to display and offer what she chooses to, and the fact that the user is indeed interested in what is displayed to her does not render the recommendation's influence controlling. Targeted advertisements do not generally engage rational deliberation less than other forms of advertisements and commercial speech.

[41] See for exceptions below: 6.5.3.2.3. *Addressing fears through (allegedly) harm-alleviating offers*.

[42] See below: 6.5.3.2.3. *Addressing fears through (allegedly) harm-alleviating offers*.

[43] E.g. Aguirre et al. 2015 show that personalization leads to a greater click through, though it can also have negative effects like loss of trust, if the personalization is too pronounced. C.f. also Tucker 2013; see also Matz et al. 2017.

The question under which conditions personal data may be used for targeted recommendations is a matter of data protection law. As a rule, data processing for this purpose is allowed if the user has consented to it.[44] In this case there is usually also no problem with a view to private autonomy, because the user has agreed to this kind of influence.

Targeted unrelated recommendations cannot in general be considered manipulative. They can, however, be manipulative under special circumstances. I suggest that this is the case either when they exploit vulnerable moments or when they are likely to play on certain fears of the influenced subject and present allegedly harm-alleviating offers.

6.5.3.2.2 Exploiting Emotions

It has been reported that platform algorithms can detect users' feelings and emotions (e.g. insecurity, anxiety, stress, inattentiveness) in real time, based on the analysis of users' posts, behavior, tone of voice or by measuring mouse or eye movement (Miotto Lopes and Chen 2021; Susser et al. 2019b; Calo 2014).[45] The information about a user's current emotional state could be used to exploit vulnerabilities caused by it, to recommend items that the user might be more receptive to in her emotional state. A platform could for example show anti-stress items (meditation apps, anti-stress balls, books on time management etc.) to a stressed user or (overpriced) cosmetic items to insecure teenage girls.

Such practices would specifically target vulnerable situations in which the user is more likely to make less rational and more emotional decisions.[46] One can argue that, under these circumstances, the rational capacities for decision-making are not sufficiently engaged but rather attempted to be bypassed. Users are more likely to be lead astray from the ideal to first reflect before deciding. Of course, the success of such a strategy depends on the individual users. Some might impulsively buy books on stress management that they will never find the time to read (or they will read it and find it quite helpful) while others will still ignore the advertisement or notice it and decide to calmly research the possibilities of stress management later. Using emotion as a decisive criterion to target advertisement hints to the intention to manipulate, i.e., to exploit a vulnerable situation to provoke a decision in the advertiser's favor. Exploiting emotions to target recommendations is manipulative especially when the situation is used to incentivize a decision that is likely not in

[44] Article 6 (1) lit. a GDPR.

[45] Of course, some of this information can only be used if the used devices record voce or eye-movement which usually requires that the user allows it or actively uses these functions (e.g., to speak with a "digital assistant" like Alexa, Siri etc.).

[46] Calo 2014: "[T]he concern is that hyper-rational actors armed with the ability to design most elements of the transaction will approach the consumer of the future at the precise time and in the exact way that tends to Guarantee a Moment of (profitable) Irrationality".

the objective interest of the influenced, because she does not really need it or because she is offered something at a higher price.

Would the influence be of a different nature if the user was informed that the recommendation was based on her feelings? That depends on how much of the influence strategy would be revealed. Telling the user “we recommend this mascara to you because you feel insecure” or “we recommend the meditation app because you are stressed” would no doubt make her aware of the connection between the recommendation and her feeling, but would it help to overcome the decision-making vulnerability? Does that knowledge make her less vulnerable?[47] In my view, this would not be enough to foreclose manipulation in the case of emotion-based recommendations, because it does not uncover the manipulative mechanism. That would be different if the user was told e.g., “we show this mascara to you because we know that you are insecure about your appearance and we assume that this emotional state makes you more likely to spontaneously buy the mascara and accept the price we set”; or “we suggest the meditation app to you because you are stressed and in this moment probably so desperate for relief that you are willing to conclude a contract for 25 € per month.” It seems unlikely that a platform provider would want to fully disclose this underlying mechanism. Indeed, it is more likely that they would want to keep this strategy secret. This is a further hint of the manipulative nature on such strategies.

6.5.3.2.3 Addressing Fears Through (Allegedly) Harm-Alleviating Offers

Where the platform holds detailed profiles on their users, the data probably shows when a user suffers from or is under certain conditions (e.g., health issues, financial problems (Susser et al. 2019a)). When this knowledge is used to target recommendations with commercial offers on health products or credit-schemes in a context and at a time where the user is looking for something else, for example reading news, this seems problematic. By seeing recommendations on such products while looking for something else, the user is reminded of her condition in a situation where she intended to focus on something else (cf. Noggle 1996). Because the recommendation touches upon a sensitive issue, it is less likely that she can ignore it or that the brain subconsciously filters it out. To be confronted with one’s problems in an unexpected situation is likely to cause anxiety or other negative feelings, which may cause decision-making vulnerabilities.

When someone decides to inform herself on products related to her problems or health conditions, then she arms herself beforehand, and is prepared to deal with the issue. If confronted with the issue by surprise, that is probably less the case, which seems to make her more vulnerable for less rational decisions. Advertisements for

[47] Doubting as well, Calo 2014.

credits and drugs – to stay with my examples – can be perceived as harm-alleviating offers, and it is generally harder to resist to them (Faden and Beauchamp 1986). Targeted harm-alleviating unrelated recommendations should be considered manipulative for the same reasons I have put forward regarding emotion-based recommendations.

6.5.4 Interim Conclusion: Recommender Systems, Manipulation and Private Autonomy

Recommender systems are not per se manipulative, but they can be used and programmed in a manipulative way. This is the case when recommendations are opaquely based on unexpected criteria or contradict the nature of a recommendation. Recommender systems also work in a manipulative way when they are programmed to use profiling to address fears and target commercial[48] (allegedly) harm-alleviating offers to users when they are not looking for it. It is further the case when recommender systems are programmed to base recommendations on real time emotion recognition to exploit decision-making vulnerabilities evoked by the emotional state.

This last statement, however, must be qualified. A problem with private autonomy does not arise when the user's decision does not include any commitment and does not have any legal consequences. For example, the decisions within abo models such as Netflix or Spotify about which song to hear or which movie to watch is in no way binding and unconditionally revocable without any effort. If recommender systems on these platforms would base recommendations on users' current emotions, this would not impact the formation of a legal relationship.[49] This is different when it comes to the formation of a binding legal contract. Though contracts concluded online can in many cases be revoked, this comes with further obligations and is conditioned.[50]

It is difficult – if not impossible – to assess which impact recommender systems have, how many decisions they influence that otherwise would not have been taken in that way and what is the economic harm for users that results from such decisions. However, when they are programmed to exert manipulative influences,

[48] Including service for data offers.

[49] In this context, it is already questionable whether basing recommendations on emotions could count as manipulative, because taking a decision e.g. on music to hear according to one's emotions is rather reasonable and in the users' interest. There is no hidden strategy involved, when a platform recommends music, that the user is probably going to like in her emotional state.

[50] The right to withdrawal is limited in time (article 9 CR-D) and consumer need to return the received good on his own expense (article 14 CR-D).

they pose a danger to the recipients' autonomy that should be controlled (Calo 2014). The named cases in which recommender systems could interfere with the autonomous decision of users to form legal relationships should be regulated.

6.6 Regulation Regarding Recommender Systems

For a long time, recommender systems were not subject to explicit regulation. However, this does not mean that no rules applied to them. The deployment of recommender systems and the presentation of the recommendations by online platforms is part of their commercial communication to which all relevant civil law rules concerning commercial communication apply. The processing of personal data for the purpose of making recommendations must of course comply with data protection laws.

However, in recent years, recommender systems have increasingly come to the attention of European legislation. Several legislative instruments have been adopted or are on their way to be adopted that either directly or indirectly regulate recommender systems and targeted advertising. This part examines whether the existing and upcoming legal instruments in EU law are suited to foreclose the manipulative potential of recommender systems identified above.

6.6.1 Unexpected Recommendation Criteria

The EU Unfair Commercial Practices Directive[51] (UCP-D) prohibits unfair commercial practices, in particular misleading and aggressive commercial practices (article 5 UCP-D). From this general prohibition, it follows that statements accompanying recommendations must be true.[52] The information that a recommendation is based on other criteria than those to be expected for a recommendation is material for an informed decision and must be made transparent. Not making unexpected recommendation criteria transparent constitutes a misleading omission in the sense of article 7 (1) UCP-D (Peifer 2021).

[51] Directive 2005/29/EC of the European Parliament and of the Council of 11 May 2005 concerning unfair business-to-consumer commercial practices in the internal market and amending Council Directive 84/450/EEC, Directives 97/7/EC, 98/27/EC, and 2002/65/EC of the European Parliament and of the Council and Regulation (EC) No 2006/2004 of the European Parliament and of the Council (Unfair Commercial Practices Directive, here: UCP-D).

[52] E.g. if recommendations are headed with "customers who bought this item, also bought…", this statement must be true; or when a recommended item is labelled as a "bestseller" it must be a bestseller.

A recently adopted[53] new paragraph 4a of article 7 UCP-D makes this explicit: when platforms[54] provide consumers with the possibility to search for products and services offered by someone other than the platform itself, information "on the main parameters determining the ranking of products presented to the consumer as a result of the search query and the relative importance of those parameters, as opposed to other parameters shall be regarded as material".[55] A new article 6a (1) lit. a in the Consumer Rights Directive[56] (CR-D) contains a correlating precontractual information obligation.[57] Furthermore, the provision of "search results in response to a consumer's online search query without clearly disclosing any paid advertisement or payment specifically for achieving higher ranking of products within the search results" has been added to the UCP-D's blacklist of unfair commercial practises (Annex No. 11a UCP-D).

These rules solve the problem of unexpected recommendation criteria for search result lists on platforms,[58] however, only in relation to consumers.[59] This gap is filled

[53] As a part of the "New Deal for Consumers" the European Parliament and the Council have delivered a new Directive (EU) 2019/2161 of 27 November 2019 amending the Unfair Commercial Practices Directive (Directive 2005/29/EC) to achieve a better enforcement and modernisation of Union consumer protection rules. A consolidated version of the Unfair Commercial Practices Directive (2005/29/EC) as amended by Directive 2019/2161/EU is available under https://eur-lex.europa.eu/eli/dir/2005/29/2022-05-28. The new rules must be implemented into member states' law and be applied from 28 May 2022 onwards.

[54] According to article 7 (4a) s. 2 UCP-D, "[t]his paragraph does not apply to providers of online search engines as defined in point (6) of article 2 of Regulation (EU) 2019/1150 of the European Parliament and of the Council".

[55] Platforms are not required to lay open their algorithms but rather to provide a general description on the default settings that determine the ranking, Directive (EU) 2019/2161 of the European Parliament and of the Council of 27 November 2019, Rec. 22 ff. Article 85 German MStV contains a similar provision regarding media content.

[56] Directive 2011/83/EU of the European Parliament and of the Council of 25 October 2011 on consumer rights, amending Council Directive 93/13/EEC and Directive 1999/44/EC of the European Parliament and of the Council and repealing Council Directive 85/577/EEC and Directive 97/7/EC of the European Parliament and of the Council.

[57] Article 5 of the Regulation (EU) 2019/1150 (P2B Regulation) obliges platforms to make ranking criteria transparent towards businesses using the platform to market their products and services, so they know what they can do to achieve better rankings.

[58] In a way, they even go beyond what was necessary to protect consumers' autonomous decisions when they also require platforms to lay open expected criteria. However, requiring transparency on all ranking criteria allows consumers to make more informed decisions without putting a heavy burden on the operators of the recommender system or the platforms. From this perspective, the transparency obligation is surely welcome. Search engines are regulated by Regulation (EU) 2019/1150 of the European Parliament and of the Council of 20 June 2019 on promoting fairness and transparency for business users of online intermediation services (P2B Regulation).

[59] In the online context, it becomes less and less obvious that only consumers need certain protection and not also agents acting in a commercial interest, e.g. as proxy for a small or medium-sized company. But this is a matter that goes beyond the scope of this paper. Apart from that, in practice it is likely that also commercially acting agents will sufficiently benefit from these transparency rules, at least, as long as they use the same platforms used also by consumers.

by the Digital Services Act (DSA)[60] that contains explicit rules for recommender systems and the application is not limited to consumer-to-business-relationships.

Article 27 (1) DSA requires providers of online platforms to explain in their terms and conditions 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." Where several recommendation options are available,[61] platforms shall make it easy for recipients to choose and modify at any time the relevant recommendation settings.[62] Art. 38 DSA extends these obligations to very large search engines. It is unfortunate that the DSA only enforces the disclosure of the criteria in the terms and conditions and not more directly connected with the user interface visible with the recommendations.

The DSA further contains a transparency obligation for search unrelated advertisement[63] recommendations: article 26 (1) lit. d provides that platforms that display advertisement shall provide real-time information for each specific advertisement "about the main parameters used to determine the recipient to whom the advertisement presented and, where applicable, about how to change those parameters". The information shall be "directly and easily accessible from the advertisement".

6.6.2 *Targeted Recommendations Exploiting Emotions or Addressing Fears*

EU law does not prohibit personalized and targeted advertising. In the political discussion surrounding the DSA, it was debated whether the DSA should generally prohibit targeted advertising, however a full ban was not adopted.

Some limitation to targeted recommendations that exploit emotions or address personal fears or problems of the user derives from the General Data Protection Regulation[64] (GDPR). Article 9 (1) GDPR prohibits the processing of certain

[60] 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), OJ L 277/1.

[61] That means either, where the user can choose criteria for the relative order of the recommendations (on shopping sites such criteria are typically e.g. price, popularity, sustainability of the products, on travel booking sites criteria are typically price, location of hotels, in the case of travel by public transport, the number of intermediate stops or the duration of the trip) or where she can choose the method used, see footnote 20 ff.

[62] Article 27 (3) DSA.

[63] Art. 3 lit. r DSA defines advertisement as "information designed to promote the message of a legal or natural person, irrespective of whether to achieve commercial or non-commercial purposes, and presented by an online platform on its online interface against remuneration specifically for promoting that information".

[64] 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).

sensitive personal data, amongst others health data and “biometric data for the purpose of uniquely identifying a natural person”. Real time emotion recognition to be used for recommendations would either require the analysis of content the user has just posted or an analysis of specific movements or bodily functions. When detecting the latter (e.g. eye or mouse movement, tone of voice), biometric data[65] can be collected in the process. (The voice for example is considered biometric data, Schild 2021). If this is not the case, because for example mouse movement does not allow to identify a person and the movement data is not set in connection with other personal data but used in an anonymized way, the GDPR does not apply.[66] It is conceivable that for displaying a recommendation based on emotions, the algorithm does not need to know who the user is and does not necessarily need to connect the finding on an emotion (e.g. that a user is stressed) with other data. The mere information about the emotional state could be sufficient information to target advertisement for stress remedies to that user.

The prohibition to process sensitive data does not apply when the person concerned explicitly consented to the processing for the specific purpose. The consent can be requested for several purposes at the same time (art. 9 (2) lit. a GDPR). As a result, the information about the processing purposes often results in long texts that have led to the well-known phenomenon that most internet users (almost necessarily, due to information overload)[67] click a consent button without reading and reflecting about what they are consenting to (cf. e.g., Ben-Shahar and Schneider 2014). The prohibition to process sensitive data further[68] does not apply to data which the data subject has made public (art. 9 (2) lit. e GDPR). When, for instance, a user publicly posts information about her disease or disability, a platform could use this information to target advertising for medication or other health products.

Article 9 (1) GDPR contains a conclusive list of what is considered sensitive data. Financial information or information on payment behavior is not included. Though the processing of non-sensitive personal data is not unconditionally lawful, it is less restricted. Apart from consent, art. 6 (1) lit. b GDRP offers a broad legal basis for the processing, that is, if it “is necessary for the performance of a contract to which the data subject is party or in order to take steps at the request of the data subject prior to entering into a contract”.[69]

[65] Defined in article 4 (4) GDPR as “‘biometric data’ means personal data resulting from specific technical processing relating to the physical, physiological or behavioural characteristics of a natural person, which allow or confirm the unique identification of that natural person, such as facial images or dactyloscopic data”.

[66] The GDPR does not apply to anonymous data, rec. 26s. 6 GDPR.

[67] Considering that this decision must usually be taken several times a day on different platforms, websites and in apps.

[68] Article 9 (2) GDPR contains more exceptions for the prohibition to process sensitive date, but none of them would be applicable in the cases this paper is concerned with.

[69] The practice of personalized pricing shows that financial information is used by platforms (c.f. OECD 2018).

It can be concluded that the GDPR gives users the possibility to mostly prevent personalized and targeted recommendations by denying consent. In practice, however, this appears to be of limited effectiveness.[70] Whether, and to what extent, emotion-based recommendations can be prevented by data protection law is questionable (Miotto Lopes and Chen 2022).

One could consider whether targeted unrelated recommendations that exploit emotions or unfortunate personal circumstances for harm-alleviating offers could count as prohibited aggressive commercial practices (article 5 (1), (4) lit. b, 8 UCP-D). A commercial practice is aggressive *inter alia* if it takes undue influence that is at least likely "to significantly impair the average consumer's freedom of choice or conduct with regard to the product and thereby causes him or is likely to cause him to take a transactional decision that he would not have taken otherwise" (article 8 UCP-D). In determining whether an influence is undue, several factors shall be considered, amongst others the timing (article 9 lit a UCP-D) and the exploitation "of any specific misfortune or circumstance of such gravity as to impair the consumer's judgement, of which the trader is aware" (article 9 lit. c UCP-D). At first glance, this could stand in the way of exploiting emotions or personal hardships to target commercial offers. However, moods or emotions themselves are not specific circumstances[71] of severe gravity, comparable to a misfortune and the platform is unlikely to know what circumstances caused the mood. Recommendations based on real time emotion recognition therefore cannot be considered an aggressive commercial practice. But articles 5 (1), (4) lit. b, 8 UCP-D also does not foreclose recommendations that address fears for harm-alleviating offers. Article 2 lit. j UCP-D defines undue influence as "exploiting a position of power in relation to the consumer *so as to apply pressure* […], in a way which significantly limits the consumer's ability to make an informed decision"[72] (emphasis added). Simply displaying a certain product depending on the user's mood or problems does not create pressure (Ebers 2018; Miotto Lopes and Chen 2022).[73]

The DSA does not fully foreclose targeted advertising and recommendations based on emotion recognition or personal hardship. However, it establishes some restrictions beyond the information requirements on the targeting criteria as provided for by articles 27 DSA that alone do not eliminate the manipulative potential

[70] Giving consent is usually just a click that users often undertake without real awareness of what exactly they are consenting to (c.f. also footnote 68) and also consent is not needed under all circumstances.

[71] Something that deviates from the normal course of events (Köhler 2021: UWG § 4a Rn. 1.89).

[72] Article 2 lit. j UCP-D.

[73] CJEU case C-628/17: Judgment of the Court (Fifth Chamber) of 12 June 2019, Presez Urzędu Ochrony Konkurencji i Konsumentów v Orange Polska S. A., paras 32, 46. The pressure must be of a kind that an average consumer is likely to not withstand it. That is the case if the average consumer assumes that he/she cannot escape the pressure and therefore considers behaving in the way the company wants, in order to avoid a threatened disadvantage (Köhler 2021: UWG § 4a Rn. 1.60).

of recommendations based on emotions and addressing personal hardships or fears (Miotto Lopes and Chen 2022).

Minors may no longer be targeted with advertisement based on profiling, article 28 (2) DSA. Article 26 (3) DSA prohibits online platforms the targeting of advertisement based on profiling as defined in Article 4, point (4) GDPR using sensitive data (article 9 (1) GDPR). While the parliament's proposal of the DSA foresaw a general prohibition for targeting and amplification techniques using sensitive data,[74] the final DSA version contains the insertion that sensitive data must not be used as input data for profiling for targeted advertisement. With this limitation Article 26 (3) DSA is hardly effective against recommendations exploiting emotions or addressing fears, because it does not prohibit targeting information based on sensitive data that was inferred from other non-sensitive data.[75] Large platforms often have enough user data to be able to draw conclusions, for example, about a user's state of health or sexual orientation. They can draw conclusions from search queries and purchased products. The DSA's rules hence do not put a full stop to targeted commercial recommendations based on health data. Emotion-based advertisement targeting is only prohibited as far as it requires the processing of biometric data for profiling.

Article 38 DSA requires at least providers of very large platforms and very large search engines as defined in article 33 DSA to give users the opportunity to choose a recommendation option that is not at all based on profiling in the sense of article 4 (4) GDPR. Platforms that have not been designated as very large do not have to meet this requirement.

The use of emotion recognition systems[76] is addressed in the proposed Artificial Intelligence Act (D-AI Act).[77] A prohibition of emotion recognition for commercial or advertising purposes is not envisioned in the draft.[78] Emotion recognition operating on commercial platforms is also not considered a high-risk system.[79] Insofar, article 52 (2) s. 1 D-AI Act only provides for an information obligation: "[u]sers of an emotion recognition system or a biometric categorisation system shall inform of the operation of the system the natural persons exposed thereto." This provision is

[74] Article 24 (1b) Amendments adopted by the European Parliament on 20 January 2022 on the proposal for a regulation of the European Parliament and of the Council on a Single Market For Digital Services (Digital Services Act) and amending Directive 2000/31/EC (2020/0361(COD)), P9_TA(2022)0014.

[75] For the distinction between input data for profiling and inferred data see Lorentz 2020.

[76] Article 3 (34) draft AI-A: "'emotion recognition system' means an AI system for the purpose of identifying or inferring emotions or intentions of natural persons on the basis of their biometric data".

[77] Proposal COM/2021/206 final for a Regulation of the European Parliament and of the Council laying down harmonised rules on Artificial Intelligence (Artificial Intelligence Act) and amending certain Union legislative acts.

[78] Article 5 (1) li. a + b D-AI Act prohibits certain Ai deployments. But though recommender systems may fall under the AI definition and recommendations based on emotion recognition can be taken to be subliminal techniques, recommendations are unlikely to cause physical or psychological harm (Miotto Lopes and Chen 2022).

[79] According to article 6 D-AI-Act.

intended to protect a person's informed choice and recognizes the manipulative potential of emotion recognition systems.[80] But as said above, the mere information is not sufficient to foreclose manipulation in the here assessed cases.

6.6.3 Regulative Measures to Take Regarding Recommender Systems

The existing European legal instruments do not sufficiently foreclose manipulation through recommender systems to protect platform users' private autonomy. European legislators should close the identified protection gaps. Not only users of very large platforms should mandatorily be given a choice against profiling-based recommendations. The principal information on recommendation criteria should not just be accessible through links to the terms and conditions but should be given in short right next to the recommendations or in catchwords with the recommendation, as it is required by article 26 DSA only for search unrelated advertisement recommendations. Further and more detailed information can and should be made accessible through links.

There should be clear rules that prohibit targeted recommendations based on real time emotion recognition that apply for all platforms. The exploitation of emotions with the intention to influence legally relevant decision making in a commercial context is highly manipulative. Even if a user would consent to the use of emotion recognition when registering on the platform this does not seem to be sufficient to maintain autonomy in the situation of influence.

Profiling-based harm-alleviating commercial recommendations also in the form of targeted advertising should be prohibited at least as a default setting, even if a user consented to the use of his personal data for advertisement purposes when registering on a platform. Only when a user explicitly opts-in to receiving harm-alleviating recommendations based on profiling the user's autonomy is sufficiently preserved.

These regulatory measures would not unduly restrict the autonomy of the platforms. In principle, they would remain free to decide on their offerings, but they would not be allowed to influence user decisions in the ways described.

The existing contract law is not suitable to sufficiently compensate for the identified autonomy risks for users. A closed contract can only be avoided if a user can prove deception.[81] Deception could only be alleged against opaque unexpected ranking criteria, not against the other potential manipulative influences revealed in

[80] See Proposal COM/2021/206 final for a Regulation of the European Parliament and of the Council laying down harmonised rules on Artificial Intelligence (Artificial Intelligence Act) and amending certain Union legislative acts, Explanatory Memorandum, 5.2.4.

[81] The other cases of avoidance do not play a role in connection with recommendation systems.

this paper, and only when the contractor is responsible for the deception. It is questionable whether the platform that uses the recommender system to recommend third party offers is acting as a representative of the contractual partner and whether a user could show that he would have acted differently without the given influence (Mik 2016). Apart from this, the user is likely to never learn that she has been manipulated.

For this reason, also the withdrawal rights provided by EU law for consumer contracts concluded on the Internet are not sufficient compensation. The withdrawal rights are limited in time and the consumer may incur return shipping costs.[82]

6.7 Conclusion

Recommender systems are useful tools without which it would be impossible to cope with the flood of information available online. They influence decision-making because they decide about what comes to a user's attention. In many settings, they facilitate choice. They can, however, be used in manipulative ways and be programmed to manipulate. This paper has shown that this is the case when unexpected recommendation criteria are not made transparent and where recommendations aim at certain decision-making vulnerabilities.[83] This should be prohibited by law. Especially in the amended Unfair Commercial Practices directive, the amended Consumer Rights directive and the Digital Services Act, the European legislator has already taken steps to regulate recommender systems, but there is a need for further legislative action.

Private platforms enjoy private autonomy, just like private platform users. It follows that platforms are in principle free to design their offers and services. If there are no additional circumstances (such as a dominant or significant market position),[84] they are basically free to decide what they present and offer to whom and on what terms. However, if platform operators try to manipulate legally relevant decisions of their users, it is appropriate to set limits to their freedom to protect the private autonomy of users. Even if the harm to autonomy in the individual case or decision might be small, the systematization and potential scale of this kind of manipulation are worrisome (Calo 2014). Principiis obsta (Susser et al. 2019b)!

[82] See above footnote 50.

[83] One can put forward different reasons why personalised advertising should generally be prohibited, in particular one can have something against the underlying profiling. But from the perspective of private autonomy, targeted and personalized advertisement is only problematic if used in a manipulative way and is hence a danger to autonomy.

[84] Regulation (EU) 2022/1925 of the European Parliament and of the Council of 14 September 2022 on contestable and fair markets in the digital sector and amending Directives (EU) 2019/1937 and (EU) 2020/1828 (Digital Markets Act) contains a number of restrictions for "gatekeepers", see especially Art. 6 (5).

References

Ackerman, F. 1995. The Concept of Manipulativeness. *Philosophical Perspectives* 9: 335–340. https://www.jstor.org/stable/2214225.

Aguirre, E., D. Mahr, D. Grewal, K. De Ruyter, and M. Wetzels. 2015. Unraveling the Personalization Paradox: The Effect of Information Collection and Trust-Building Strategies on Online Advertisement Effectiveness. *Journal of Retail* 91 (1): 34–49.

Alston, W.P. 1967. Vagueness. In *The Encyclopedia of Philosophy*, ed. P. Edwards, vol. 8, 218–220. New York: Collier-Macmillan.

Barnhill, A. 2014. What is Manipulation? In *Manipulation: Theory and Practice*, ed. C. Coons and M. Weber. Oxford: Oxford University Press. https://oxford.universitypressscholarship.com/view/10.1093/acprof:oso/9780199338207.001.0001/acprof-9780199338207-chapter-3.

Beam, M.A. 2014. Automating the News: How Personalized News Recommender System Design Choices Impact News Reception. *Communication Research* 41 (8): 1019–1041. https://doi.org/10.1177/2F0093650213497979.

Ben-Shahar, O., and C.E. Schneider. 2014. *More Than You Wanted To Know: The Failure of Mandated Disclosure*. Princeton: Princeton University Press. https://www.jstor.org/stable/j.ctt5hhrqj.

Bumke, C. 2017. Privatautonomie. In *Autonomie im Recht*, ed. Bumke, C. and A. Röthel. Tübingen: Mohr Siebeck.

Busche, J. 1999. *Privatautonomie und Kontrahierungszwang*. Tübingen: Mohr Siebeck.

Calo, R. 2014. Digital Market Manipulation. *The George Washington Law Review* 82 (4): 995–1051. https://digitalcommons.law.uw.edu/cgi/viewcontent.cgi?article=1024&context=faculty-articles.

Calvo, R.A., D. Peters, K. Vold, and R.M. Ryan. 2020. Supporting Human Autonomy in AI Systems: A Framework for Ethical Enquiry. In *Ethics of Digital Well-Being. A Multidisciplinary Approach*, Philosophical Studies Series, ed. C. Burr and L. Floridi, vol. 140. Cham: Springer. https://doi.org/10.1007/978-3-030-50585-1_2.

Deci, E.L., and R.M. Ryan. 2008. Self-Determination Theory: A Macrotheory of Human Motivation, Development, and Health. *Canadian Psychology/Psychologie canadienne* 49 (3): 182–185. https://psycnet.apa.org/doi/10.1037/a0012801.

Dworkin, G. 1988. *The Theory and Practice of Autonomy*. Cambridge: Cambridge University Press.

Ebers, M. 2018. Beeinflussung und Manipulation von Kunden durch Behavioral Microtargeting. *Multimedia und Recht* 7: 423–428.

Faden, R.R., and T.L. Beauchamp. 1986. *A History and Theory of Informed Consent*. Oxford: Oxford University Press.

Flume, W. 1979. *Allgemeiner Teil des Bürgerlichen Rechts: Das Rechtsgeschäft*. Berlin/Heidelberg: Springer.

Hacker, P. 2017. *Verhaltensökonomik und Normativität*. Tübingen: Mohr Siebeck.

Harper, D. 2022. Etymology of manipulation. *Online Etymology Dictionary*. Accessed 13 Jan 2022.

Hobbes, T. 1794. *Leviathan*, Erster Band. Halle: Joh. Christ. Hendel Verlag.

Kahneman, D. 2011. *Thinking, Fast and Slow*. New York: Farrar, Straus and Giroux.

Kant, I. 2016. *Grundlegung zur Metaphysik der Sitten*. Hamburg: Meiner Felix Verlag.

Klenk, M. 2021. (Online) Manipulation: Sometimes Hidden, Always Careless. *Review of Social Economy* 80 (1): 85–105. https://doi.org/10.1080/00346764.2021.1894350.

Köhler, H. 2021. Die drei Mittel der Beeinflussung. In *Gesetz gegen den unlauteren Wettbewerb – Kommentar*, ed. H. Köhler, J. Bornkamm, and J. Feddersen. München: C. H. Beck.

Lobinger, T. 2007. Vertragsfreiheit und Diskriminierungsverbote. Privatautonomie im modernen Zivil- und Arbeitsrecht. In *Vertragsfreiheit und Diskriminierung*, ed. J. Isensee. Berlin: Duncker & Humblot.

Lorentz, N. 2020. *Profiling*. Tübingen: Mohr Siebeck.

Matz, S.C., M. Kosinski, G. Nave, and D.J. Stillwell. 2017. Psychological Targeting as an Effective Approach to Digital Mass Persuasion. *The Proceedings of the National Academy of Sciences* 114 (48): 12714–12719. www.pnas.org/cgi/doi/10.1073/pnas.1710966114.

Mik, E. 2016. The Erosion of Autonomy in Online Consumer Transactions. *Law, Innovation and Technology* 8 (1): 1–38. https://doi.org/10.1080/17579961.2016.1161893.

Milano, S., M. Taddeo, and L. Floridi. 2020. Recommender Systems and Their Ethical Challenges. *AI & Society* 35: 957–967. https://doi.org/10.1007/s00146-020-00950-y.

Miotto Lopes, L., and J. Chen. 2021. Timed Influence: The Future of Modern (Family) Life and the Law. *Scripted – A Journal of Law, Technology & Society*, September 10, 2010. https://scipt-ed.org/blog/timed-influence-the-future-of-modern-family-life-and-the-law/

———. 2022. Manipulation, Real-time Profiling, and Their Wrongs. In *The Philosophy of Online Manipulation*, ed. Michael Klenk and Fleur Jongepier. New York/London: Routledge. https://doi.org/10.4324/9781003205425.

Möllers, T.M.J. 2018. Working with Legal Principles – Demonstrated Using Private Autonomy and Freedom of Contract as Examples. *European Review of Contract Law* 14 (2): 101–137. https://doi.org/10.1515/ercl-2018-1007.

Noggle, R. 1996. Manipulative Actions: A Conceptual and Moral Analysis. *American Philosophical Quarterly* 33 (1): 43–55. http://www.jstor.org/stable/20009846.

———. 2020. The Ethics of Manipulation. In *The Stanford Encyclopedia of Philosophy*. Summer 2020 edition. ed. Zalta, E.N. Stanford University: Metaphysics Research Lab.

Organisation for Economic Co-operation and Development (OECD). 2018. Background Note by the Secretariat: *Personalised Pricing in the Digital Era*. DAF/COMP(2018)13. https://www.oecd.org/competition/personalised-pricing-in-the-digital-era.htm

Peifer, K.-N. 2021. Die neuen Transparenzregeln im UWG (Bewertungen, Rankings und Influencer). *Gewerblicher Rechtsschutz und Urheberrecht* 123 (12): 1453–1461.

Raz, J. 1988. *The Morality of Freedom*. Oxford/New York: Oxford University Press. https://doi.org/10.1093/0198248075.001.0001.

———. 2009. *Between Authority and Interpretation: On the Theory of Law and Practical Reason*. Oxford/New York: Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199562688.001.0001.

Ricci, F., L. Rokach, and B. Shapira. 2012. *Recommender Systems Handbook*. New York: Springer. https://doi.org/10.1007/978-1-0716-2197-4.

Riesenhuber, K. 2003. *System und Prinzipien des Europäischen Vertragsrechts*. Berlin: De Gruyter.

———. 2018. Privatautonomie – Rechtsprinzip oder "mystifizierendes Leuchtfeuer". *Zeitschrift für die gesamte Privatrechtswissenschaft* 3: 352–368.

Rössler, B. 2017. *Autonomie. Ein Versuch über das gelungene Leben*. Berlin: Suhrkamp.

Röthel, A. 2017. Forschungsgespräche über Autonomie im Recht. In *Autonomie im Recht*, ed. C. Bumke and A. Röthel. Tübingen: Mohr Siebeck.

Ryan, R.M., and E.L. Deci. 2000. The "What" and "Why" of Goal Pursuits: Human Needs and the Self-Determination of Behavior. *Psychological Inquiry* 11 (4): 227–268. https://selfdeterminationtheory.org/SDT/documents/2000_DeciRyan_PIWhatWhy.pdf.

———. 2017. *Self-Determination Theory*. New York/London: Guilford Press.

Scanlon, T.M. 1972. A Theory of Freedom of Expression. *Philosophy & Public Affairs* 1 (2): 204–226.

———. 1986. The Significance of Choice. *The Tanner Lectures on Human Values* 7: 149–216.

Schapp, J. 1992. Über die Freiheit im Recht. *Archiv für die zivilistische Praxis* 192 (5): 355–389.

Schild, H.-H. 2021. Biometrische Daten. In *BeckOK Datenschutzrecht – Kommentar*, ed. S. Brink and H.A. Us Wolff. München: C. H. Beck.

Seaver, N. 2019. Captivating Algorithms: Recommender Systems as Traps. *Journal of Material Culture* 24 (4): 421–436. https://doi.org/10.1177/2F1359183518820366.

Specht, L. 2019. *Diktat der Technik*. Baden-Baden: Nomos.

Spencer, S.B. 2020. The Problem of Online Manipulation. *University of Illinois Law Review* 3: 959–1000.

Study Group on a European Civil Code and Research Group on EC Private Law (Acquis Group). 2009. *Principles, Definitions and Model Rules of European Private Law* – Draft Common Frame of References (DCFR), ed. von Bar, C., Clive, E., and Schulte-Nölke, H. Available online: https://www.ccbe.eu/fileadmin/speciality_distribution/public/documents/EUROPEAN_PRIVATE_LAW/EN_EPL_20100107_Principles__definitions_and_model_rules_of_European_private_law_-_Draft_Common_Frame_of_Reference__DCFR_.pdf. Accessed on 23.09.2022

Sunstein, C.R. 2016. *The Ethics of Influence: Government in the Age of Behavioral Science*. Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9781316493021.

Susser, D., B. Rössler, and H. Nissenbaum. 2019a. Technology, Autonomy, and Manipulation. *Internet Policy Review* 8 (2): 1–22. https://doi.org/10.14763/2019.2.1410.

———. 2019b. Online Manipulation: Hidden Influences in a Digital World. *Georgetown Law Technology Review* 4 (1). https://doi.org/10.2139/ssrn.3306006.

Thaler, R.H., and C.R. Sunstein. 2008. *Nudge*. London: Penguin Books.

Tucker, C. 2013. Social Networks, Personalized Advertising, and Privacy Controls. *Journal of Marketing Research* 51 (5): 546–562.

Varshney, L.R. 2020. Respect for human autonomy in recommender systems. https://arxiv.org/pdf/2009.02603.pdf

Wilkinson, T.M. 2013. Nudging and Manipulation. *Political Studies* 61 (2): 341–355. https://doi.org/10.1111/j.1467-9248.2012.00974.x.

Wood, A.W. 2014. Coercion, Manipulation, Exploitation. In *Manipulation: Theory and Practice*, ed. Coons, C. and Weber, M. https://oxford.universitypressscholarship.com/view/10.1093/acprof:oso/9780199338207.001.0001/acprof-9780199338207-chapter-2

Chapter 7
Reasoning with Recommender Systems? Practical Reasoning, Digital Nudging, and Autonomy

Marius Bartmann

Abstract One of the core tasks of recommender systems is often defined as follows: Find good items. Recommender systems are thus designed to support our decision-making by helping us find our way around the online world and guiding us to the things we want. However, relying on recommender systems has a profound effect on our decision-making because they structure the environment in which we make choices. In this contribution, I examine how recommender systems affect our practical reasoning and whether they pose a threat to autonomy, i.e., what influence recommender systems have on our capacity for making our own choices. I argue that a basic requirement for integrating automated recommendations in autonomous decision-making consists in being able to identify the rationale behind recommendations: only if we understand why we are being presented with certain recommendations is it possible for them to be integrated into decision-making in a way that preserves autonomy.

Keywords Human-computer interaction · Digital nudging · Digital ethics · Decision-making · Autonomy

7.1 Introduction

In a classic paper, Herlocker et al. define one of the core tasks of recommender systems as follows: *Find good items* (Herlocker et al. 2004, 9). This definition, however, is deceptively simple. What is a good item? And good for whom? For the user or for the provider of the recommender system? Or even for a third party? These questions are much more complicated than they might seem, and they are not as frequently dealt with as one might expect. As Jannach and Adomavicius note, a "question that is rarely asked explicitly in recommender systems research, is: *What*

M. Bartmann (✉)
University of Bonn, Bonn, Germany
e-mail: bartmann@uni-bonn.de

S. Genovesi et al. (eds.), *Recommender Systems: Legal and Ethical Issues*, The International Library of Ethics, Law and Technology 40,
https://doi.org/10.1007/978-3-031-34804-4_7

is a good recommender system? (Or: *What is a good recommendation?*)" (Jannach and Adomavicius 2016, 8).

If we take, for example, e-commerce as one of the most widespread applications of recommender systems, a simple answer that first comes to mind is this: A good recommendation – the recommendation of a good item – would simply be the one resulting in a purchase. After all, if we did not think the recommendation was a good one, we would not buy the recommended item. However, even though purchases may in some cases serve as an indicator for good recommendations, purchases and good recommendations are not necessarily equivalent. If they were, cases in which consumers are somehow manipulated by a recommender system into buying things would have to count as good recommendations as well. The same goes for click-through rates (CTR) as an indicator for good recommendations. Clicking through links can be the result of a user having been presented with user-relevant content, and therefore possibly constitute an example of a good recommendation. But the all too familiar cases of clickbait clearly represent examples of misleading and deceptive recommendations (Burr et al. 2018, 743). Clickbait may seem harmless when a user is "merely" presented with irrelevant content. Some critics argue, however, that it can become a serious problem when it contributes to creating so-called "filter bubbles" and "echo chambers" (Bozdag and van den Hoven 2015; Flaxman et al. 2016), although their actual significance and precise impact is controversial (Dubois and Blank 2018).

These are only some of the ethical challenges raised by recommender systems. There are many others (Milano et al. 2020). In this contribution, I examine how recommender systems affect our practical reasoning and whether they pose a threat to autonomy, i.e., what influence recommender systems have on our capacity for making our own choices. I will argue that a basic requirement for integrating recommendations in autonomous decision-making consists in being able to identify the rationale behind recommendations: only if we understand why we are being presented with certain recommendations is it possible for them to be integrated into decision-making in a way that preserves autonomy.

The aim of this paper is thus to answer two questions: (1.) What role do recommender systems play in our decision-making? (2.) How can automated recommendations be integrated in practical reasoning such that it yields autonomous decisions? To answer the first question, I explore in a first step what is generally involved in making recommendations and integrating them in decision-making (Sect. 7.2). In a second step, I examine the influence of automated recommendations on our decision-making by critically discussing a popular proposal in the literature, which conceives of recommender systems as a form of digital nudging. In doing so, I highlight what automated recommendations and digital nudges have in common but also point out significant differences (Sect. 7.3). To answer the second question, I use the conceptual tools developed in the previous sections and argue that if our decision-making based on automated recommendations is to be autonomous, then the features used in principles of selection and filtering must be transparent because otherwise the recommendation's rationale cannot be identified and hence the recommendation cannot be integrated in practical reasoning (Sect. 7.4).

7.2 Practical Reasoning, Choices, and Recommendations

Recommendations and choices are elements of practical reasoning. Practical reasoning is the capacity for weighing reasons to decide on a particular course of action among various alternatives (Wallace 2020). Choices are the result of particular pieces of practical reasoning. Recommendations guide other people's choices by supporting their practical reasoning. They are issued, accepted, and rejected in pretty much every area of life, for example, if we ask a salesperson what product to buy or if we ask a friend what to do about a certain moral conundrum. Recommendations can thus be understood as offering decision-making support in answering questions of the general form "What should I do?" That makes recommendations a species of *normative* or *value-judgments*, judgments whose subject matter is not only concerned with matters of fact but are also closely connected to action (Rosati 2016).

Richard Hare, in his *The Language of Morals*, argued that the ultimate purpose of moral judgments – in fact, all value-judgments – is to guide choices rather than making truth-apt statements (Hare 1952, 29, 127). Non-cognitivist approaches such as Hare's dominated metaethics until the 1980s but became increasingly controversial since (Schroeder 2010). Nowadays cognitivism represents the mainstream (Bourget and Chalmers 2014, 476).

Fortunately, Hare's main thesis, his wider (meta-)ethical views and the controversies surrounding them need not concern us here. What makes Hare's reflections particularly suitable for my purposes is that in arguing for his thesis he provides a general analysis of what is involved in guiding choices in both moral and non-moral contexts, an analysis that is independent of his main thesis and the rest of his theory. So, it does not matter whether his main thesis is correct. Since the point of recommendations is to guide choices, Hare's analysis will prove useful to shed light on what is involved in making recommendations and integrating them in decision-making.

Hare starts with the observation that choosing – understood as an instance of practical reasoning as opposed to picking out something at random – is intrinsically linked to standards. Standards provide norms, rules, or principles of selecting items in order to decide one way or another:

> We only have standards for a class of objects, we only talk of the virtues of one specimen as against another, we only use value-words about them, when occasions are known to exist, or are conceivable, in which we, or someone else, would have to choose between specimens. (Hare 1952, 128)

The classes of objects – also called "class[es] of comparison" (Hare 1952, 133) – that are the targets of deliberation can be diverse: Hare's examples comprise cars, pictures, billiard-cues, and fish bait. It does not matter much which classes of comparison we consider, or whether the context of choice is actual or counterfactual. In theory, at least, a context of choice in which we can or need to decide between different specimens within a particular class of comparison can be thought of for virtually every class of objects. And calling a certain specimen of a class of comparison a good one is tantamount to suggesting it should be chosen (Hare 1952, 127).

According to Hare, such contexts of choice involve standards, which specify characteristics allowing us to compare items with one another and thus providing us with reasons to choose one specimen rather than another. For example, telling someone or being told by someone “This is a good car” implies that the car possesses certain characteristics on the grounds of which it is called a good car. These so-called “good-making characteristics” (Hare 1952, 133) are descriptive properties forming the basis for value-judgments to guide choices, whether they are our own or those of others. Depending on various factors such as (consumer) ends, values, and preferences, good-making characteristics may include, for example, facts about the car’s safety, speed, stability, or sustainability. Whatever the characteristics may be in the particular case, these characteristics form the standard according to which the cars under consideration are judged, and they can therefore figure in the reasons for either choosing or recommending a particular car. That is why questions of the form, “Why are you calling this X a good one?”, “Why should I choose X?”, and “Why are you recommending X to me?” can be answered by referring to the good-making characteristics as a particular standard of judgment.

Now, standards of judgment introduce another logical element in contexts of choice:

> As we shall see, all value-judgements are covertly universal in character, which is the same as to say that they refer to, and express acceptance of, a standard which has an application to other similar instances. (Hare 1952, 129)

Value-judgments such as “That is a good X” are “covertly universal” for the simple reason that the standards entailed by this kind of judgments are in principle applicable to other members of the same class of comparison – in fact, the same standard must be applied to other members of the class of comparison on pain of inconsistency. If I tell someone “This is a good car”, and my reason for this judgment is that the car in question is stable on the road, then it would be inconsistent to say of another car with the exact same characteristic “This is a bad car” – other things being equal. Since the good-making characteristics form the basis of my value-judgment, it would be inconsistent to make a contradictory judgment about another object possessing the same good-making characteristics. Of course, usually we consider several potentially good-making characteristics and weigh them against one another to arrive at an “all things considered” judgment. Thus, one specific property serving as a good-making characteristic in one case – e.g., stability on the road – may be outweighed by other negative characteristics – e.g., poor energy efficiency. Still, unless the good-making characteristics of a particular object are somehow affected by other relevant factors, logical consistency requires that value-judgments about other relevantly similar objects cannot differ unless the good-making characteristics also differ. In other words, a difference regarding the value-judgment implies a difference regarding the good-making characteristics (Hare 1952, 131).

In sum, choosing involves standards of judgments providing us with the resources to compare and choose among specimens of a particular class of objects based on reasons. Standards are universal, not in the sense of being applicable always and everywhere, but rather in the sense that, once they are employed to judge the merits of a particular object, they automatically apply to other objects that are similar in the relevant respects simply in virtue of the objects' sharing the same characteristics on which we base our corresponding value-judgments. The upshot of this analysis of value-judgments is that judgments of the form "This is a good X" are never just about a single object but also, implicitly, about other objects sharing relevantly similar characteristics due to the involvement of universal standards.

With these considerations in mind, we can draw certain conclusions with respect to the role recommendations play in decision-making. As we have seen, choices and recommendations involve universal standards of judgment. These standards of judgment are essential for understanding the reasons on which choices and recommendations are based. If a salesperson at a car dealership recommended a certain car to me by saying "Take this car (it is a good one)", then I would have to know the standard – the good-making characteristics – informing this recommendation in order understand its rationale. After all, the good-making characteristics of it may vary significantly depending on different customer ends, values, and preferences. Maybe it is a good car in terms of speed but not in terms of sustainability; maybe it is good for commuting but not for long-distance drives; and maybe it is a good car in virtue of the size of the commission the salesperson will receive – whatever the good-making characteristics, and hence the standard of judgment, the point is that unless I know the standard and thus can understand the rationale behind the recommendation, I am in no position to integrate it in my decision-making. For, I simply cannot assign a role to something I do not understand. In short, standards of judgment provide the identity conditions for the rationale behind a recommendation; I cannot understand the recommendation's rationale if I do not know the standard (and hence the reasons) informing it.

The importance of this point lies in the fact that even a *good* reason for a particular recommendation may not necessarily constitute a good recommendation for *me* if the rationale behind the recommendation is inconsistent with my ends, values, and preferences (but possibly with those of someone else). Only if I can adopt the reasons behind a recommendation as my reasons for accepting the recommended choice can we truly speak of a good recommendation capable of being integrated into my decision-making.

In general, the good-making characteristics of an item and the ends, values, and preferences on which they depend constitute the rationale behind a particular recommendation. Basing a decision on a recommendation, or at least including it as a relevant factor in decision-making, presupposes knowing the rationale behind the recommendation in order to understand its potential role in decision-making. And only if I understand its role can I determine what it contributes to my decision. One important consequence of this, which will be relevant in Sect. 7.4, is that decisions based on recommendations whose rationale we do not understand carry the danger of making our decisions opaque to us, posing a severe threat to autonomy – and this is so regardless of whether the recommendations in question are well-intentioned or not.

In this section I explored the role of recommendations in practical reasoning in the analogue world. In the following section, I will explore the specifics of recommendations under the conditions of a digital environment, in particular with respect to recommender systems.

7.3 Recommender Systems and Digital Nudging

Navigating the online world we encounter recommendations at every turn. On streaming portals, shopping sites, social media platforms, and online newspapers we are confronted with recommendations regarding what video to watch, what products to buy, what posts to like, and what articles to read. The automated systems behind these recommendations promise to prune back the digital jungle of content and guide us to the things we are actually looking for.

As noted, the primary purpose of recommendations is to guide people's choices, and this is no different in the digital world. Whatever the motives and reasons behind the recommender systems' operators, whether they have their users' best interests at heart or not, it seems rather uncontroversial that recommender systems are designed to guide people's choices in one way or other by affecting their decision-making. That is why the employment of recommender systems has been linked to *digital nudging*, the broader online practice of influencing people's behavior with digital means (Burr et al. 2018; Jesse and Jannach 2021; Milano et al. 2020). However, the proposal to conceive of recommender systems as tools of digital nudging is rarely fleshed out and the ethical questions recommender systems raise are rarely addressed. In this section, I will discuss this proposal in more detail.

Weinmann et al. define digital nudging as follows:

> Digital nudging is the use of user-interface design elements to guide people's behavior in digital choice environments. (Weinmann et al. 2016, 433)

Digital nudging is a practice that structures people's digital environment in ways that affects their decision-making, and thus the resulting decisions. The design elements responsible for influencing choices can assume a variety of forms and usually draw on diverse psychological mechanisms known to be involved in decision-making. For example, one comparatively simple psychological phenomenon is the *middle-option bias*: given the choice between three or more options, people display a propensity towards the option situated in the middle (Schneider et al. 2018, 70). A more complex phenomenon is the *decoy effect*: people are more likely to choose an option if the option next to it is highly unattractive (Schneider et al. 2018, 69). Another powerful nudging mechanism involves setting default options. Default options exploit the *status-quo bias*, i.e., the disposition to stick with preset options out of inertia (Caraban et al. 2019, 4). Square, an app for making online payments, for example, includes "tipping" as the default option, and users have to opt-out if they do not want to tip, which increases the probability of tips (Carr 2013). The list of such psychological phenomena and the nudging mechanisms employing them is enormous. In their survey paper, Jesse and Jannach have identified and categorized 58 psychological phenomena and 87 nudging mechanisms (Jesse and Jannach 2021).

There are two important things to note about the concept of digital nudging. First, the concept of nudging was originally developed by Thaler and Sunstein and applied first and foremost to offline contexts. Second, Thaler and Sunstein embedded nudging as the central tool in their policymaking approach they dubbed *libertarian paternalism*, which is designed to help people make decisions that are better for themselves but do not restrict their freedom of choice (Thaler and Sunstein 2008). Although the background of libertarian paternalism is frequently mentioned in the literature on digital nudging, more often than not it plays only a subordinate role. This should be kept in mind when considering the ethical implications of digital nudging because the kind of goals pursued by digital nudges is highly relevant for assessing them, yet not all digital nudges are implemented with the aim to benefit users.

To bring digital nudging, particularly as exemplified by recommender systems, into sharper relief, let us compare it with the original concept from Thaler and Sunstein. Thaler and Sunstein build their libertarian paternalism on insights provided by the behavioral sciences, according to which people are less than perfect decision makers (to put it mildly). For one thing, they do not always make decisions that are in their best interest. For another, particular types of decision-making are frequently prone to a whole array of fallacies (Thaler and Sunstein 2008, 8). Practical reasoning as depicted in the previous section – in which a well-informed individual carefully balances reasons for and against the purchase of a certain kind of car, comparing its strengths and weaknesses with other cars in the light of considered ends, values, and preferences, reflecting on and integrating the expertise of a seasoned salesperson – represented only one type of decision-making. The other type consisted in decision-making that works much more quickly and without extensive reflection. This process relies more on intuition, habits, and rules of thumb, which enable us to act faster and nearly automatically in comparison to explicitly reason-based decision-making. But it is also much more error-prone (for example, the susceptibility to the middle-option bias and the decoy effect). In the established terminology introduced by Kahneman, decisions are made employing either the fast thinking of System 1 or the slow thinking of System 2, depending on what the circumstances require (Kahneman 2012, 20–22). For example, thinking through carefully the purchase of a car is a job for System 2, whereas driving it home from the car dealership on an empty and familiar road is a job for System 1.

The key idea behind nudging is that the deficiencies of System 1 can be utilized to help people make better decisions and thus increase their quality of life. Unlike conventional paternalism, libertarian paternalism tries to achieve this goal without bans and incentives, but simply by intervening in what Thaler and Sunstein call "choice architecture", i.e., people's decision environment. Whereas conventional paternalists typically restrict people's choices by prohibiting certain options, libertarian paternalists present the options in ways that make it more likely for people to choose what is supposedly better for them anyways – they nudge (Thaler and Sunstein 2008, 6).

Most importantly, then, nudges target the presentation and not the content of options with which people are confronted – nudges modify the how, not the what, which is essential for libertarian paternalists because they adamantly insist on preserving freedom of choice. The standard example is a cafeteria where fruit is put at eye level of customers to increase the probability of their choosing this healthy option rather than a less healthy one. Crucially, the less healthy options remain available. It is just that they are assigned a less attractive position in the choice architecture of the cafeteria. With freedom of choice thus ensured, libertarian paternalists promise to respect people's autonomy because they can still choose according to their preferences and thus pursue their ends unimpeded. And libertarian paternalists intend to help those whose psychological vulnerabilities – for example, weakness of the will or inertia – would otherwise prevent them from choosing what is in their own best interest.

How does this analogue form of nudging in offline context compare to digital nudging in online contexts? Generally speaking, most recommender systems employ either content-based recommendations or collaborative recommendations (Jannach et al. 2011). Content-based recommendation systems filter options by searching for items sharing specific features. This involves past user behavior because in order for the recommender system to suggest similar items it needs some point of reference, for example, items the user liked in the past. The rationale behind this recommendation technique is closely related to the role of standards in practical reasoning explored in the previous section. If I judge a certain product to be good, this judgment is usually based on certain features the product possesses (the good-making characteristics). Since these features on which I base my value-judgment form a general standard that applies to other items exhibiting the same features, it is therefore reasonable to assume that similar products will receive the same assessment. Content-based filtering thus appears to ensure that if I am presented with items similar to the ones I previously liked, then I will probably like these items as well.

The other type of recommender system frequently employed is based on collaborative recommendations. Here users are simply presented with items other users with similar user preferences have liked in the past. This "crowdsourcing" method for generating recommendations has the advantage of not needing data on the content of items as well as not needing much data on a particular user to make predictions of, and subsequently recommendations for users.

Now, in light of the definition of digital nudging there is an obvious sense in which the employment of recommender systems qualifies as a form of digital nudging: recommender systems are designed to guide people's choices by modifying their digital choice architecture. They select and order options, customize information, and suggest alternatives (Jesse and Jannach 2021, 7). Automated recommendations can thus be understood as digital nudging insofar as they influence our decision environment.

There is, however, at least one important difference between recommendations and nudges that makes it somewhat misleading to categorize recommender systems as digital nudging tools. Recall that one of libertarian paternalism's central doctrines puts the utmost emphasis on freedom of choice. Nudges are supposed to modify only how options are presented, and not to alter the range of options itself. But many recommender systems do precisely that. And given their central task – finding good items – this is rather unsurprising. The whole point of a recommender system is to filter out options that are deemed irrelevant for the user, just as a salesperson at the car dealership also does not show you all the cars they sell but only recommends to you a subset of them based on your preferences. This constitutes a significant difference between nudges and recommendations. Both are used to guide choices – but nudges *arrange* a given space of options in a way that makes it more likely for people to pick the option choice architects think is in their best interest, whereas recommendations *create* a space of options by presenting only those options purportedly tailored to people's preferences.

Even if a recommender system were to display all available options, in most cases there would be so many options that it would be practically impossible for users to review them all. Consider, for example, Google's widely used search engine. Millions of search results are displayed and ranked on countless pages for every query, yet most people only browse through the first few search result pages (Pasquale 2006). This is an example of the so-called *positioning nudge*, according to which different visual arrangements of options significantly impact our choices (Caraban et al. 2019, 5). Although freedom of choice is theoretically preserved because a user could browse through all search results, in practice this is unfeasible. This is distinctive of many recommender systems given the huge amount of data they process. Thus, in effect, a recommender system presents users only with a subset of possible options, and many of them use personalized information to shape and limit the set of possible options even further. For example, simply the location of a user – whether the user logs on from Europe or from the U.S. – will make a difference for the search results on Google (Bozdag 2013, 212).

In sum, recommender systems can be understood as a form of digital nudging in the broad sense expressed in the definition by Weinmann et al., according to which any practice counts as digital nudging that aims at guiding people's choices by modifying their digital choice architecture. But since libertarian paternalism is an integral part of the original concept of nudging, the term "digital nudging" may be misleading if it is taken to be in the service of the very same agenda. To avoid confusion, it should be kept in mind that the stated goals of libertarian paternalism – helping people make better decisions but preserving their freedom of choice – are not necessarily elements of digital nudging.

Does this mean recommender systems violate freedom of choice, and thus autonomous decision-making? This question somewhat distorts the issue since recommendations, unlike nudges, are expressly designed to reduce our options for our own benefit. Given the huge amount of online information, we want recommender systems to present us only with a relevant subset of possible options, just as we want the salesperson to show us only products of potential interest. Narrowing down possible options is not a bug, it is a feature of recommendations. How this affects our decision-making and which conditions have to be fulfilled for automated recommendations to be integrated in practical reasoning in a way that preserves autonomy, then, is the topic of the next section.

7.4 Autonomy in Practical Reasoning with Recommender Systems

Heinrichs and Knell have recently described nicely a feeling of alienation that may emerge from engaging with recommender systems, a feeling that does not normally occur during conversations with salespeople:

> Think about the bookseller again: If she gives you a recommendation, maybe you would ask for more details. Or after a reading, perhaps you would tell her your impressions and discuss the book with her. This is not possible with an AI recommendation system – and this feels weird. When someone tells you something, you expect to be able to ask questions and make comments. If this is not possible, it is a profound deviation from our common discursive practice. (Heinrichs and Knell 2021, 1575)

The difference between the recommendation of a salesperson and a recommender system is that engaging in an exchange of reasons regarding the recommendation's rationale is impossible with a recommender system. That is why, Heinrichs and Knell argue, we must not (at least not yet) consider recommender systems bona fide participants in our discursive practices, but rather only as complex tools (Heinrichs and Knell 2021, 1578). I agree with this assessment. Reasoning with a recommender system in the same way as with a bookseller may be impossible (so far). But could it be possible to reason with it in the sense of using it as a supporting tool in decision-making, i.e., to integrate the output of recommender systems in one's decision-making in a non-alienating way? In the following, I will argue that being able to identify the rationale behind recommendations is a basic requirement to assign them a meaningful role in our decision-making and thus for the resulting choices to be autonomous.

As I argued in the previous section, recommender systems can be understood as tools of digital nudging insofar as they guide people's decisions by modifying their choice architecture. Arguably, the most profound ethical problem that has been raised about nudging practices concerns threats to autonomy (Engelen and Nys 2020). The term "autonomy" is highly contested and many different interpretations have been proposed (Dworkin 2007; Jennings 2007). This is also true of debates on the ethics of nudging, but most participants are concerned with what Vugts et al., in their literature review, have summarized under the umbrella term "agency":

> Apart from a context that allows choice, autonomy also requires a *capacity* to choose and decide, and this refers to *agency*. Agency involves being able to lead one's life and act on the basis of reasons and intentions […]. This presupposes that the person has relatively stable ultimate goals, can reason about what options are preferable given those goals, and can reflect on the choices he or she makes and has made. Practical reasoning is a necessary capability in agency. (Vugts et al. 2018, 116)

In short, autonomy consists in the capacity to set your own ends and to achieve them by exercising practical reasoning – by considering possible choices and weighing reasons to make decisions. Note that this notion of autonomy first and foremost concerns personal autonomy and not moral autonomy, i.e., it concerns the ends of people considered as individuals, not the relation between people pursuing their possibly different and conflicting ends (Waldron 2005). I do think it makes a difference for the ethical assessment of a recommender system whether the type of recommended item touches on societal matters, and thus that issues of moral autonomy need to be considered in such cases. For example, it seems to make a difference whether a recommender system is designed to help you find an exciting crime novel or a means of transportation with low carbon emissions to protect the environment. The societal dimension is beyond my focus here, but I have dealt with it elsewhere (Bartmann 2022).

Among the biggest threats to autonomy is manipulation (Noggle 2018; Schmidt and Engelen 2020; Vugts et al. 2018). Just as the term "autonomy" the term "manipulation" is also contested, and there are different accounts of what constitutes manipulating someone (Noggle 2022). For present purposes, I will draw on the following definition:

> [M]anipulation is *hidden influence*. Or more fully, manipulating someone means *intentionally and covertly influencing their decision-making, by targeting and exploiting their decision-making vulnerabilities*. (Susser et al. 2019, 4)

Against the backdrop of the definitions of autonomy and manipulation, threats to autonomy can arise roughly at two different levels: at the level of the ends people pursue (i.e. at the level of *what* goals to pursue), and at the level of the means they employ to achieve these ends (i.e. at the level of *how* they pursue them – decision-making). I will review both levels with respect to recommender systems in turn.

Let us start with the level of ends. Consider once again, for example, an e-commerce recommender system and consider the viewpoint of the provider and the viewpoint of the consumer. A recommendation resulting in a purchase may be a good one from the perspective of the provider – because the purchase increases profits – but maybe not necessarily a good one from the perspective of the consumer – because the purchase may not really reflect the consumer's preferences. Or consider a certain type of business model popular with many online services employing recommender systems. Most social media platforms, for example, make money not primarily with their users but rather with other corporations paying for targeted advertising based on the users' data (Koene et al. 2015). In such cases, the potential conflict of interest between a service and its users is built into the business model because it is the corporations and not the users who are the actual (paying) customers of the service, which thus has a substantive incentive to align its interests with its customers rather than with its users.

The potential mismatch of ends, values, and preferences just described is a species of the so-called *value-alignment problem* and occurs when a recommender system is "competing", rather than "collaborating" with users (Burr et al. 2018, 742). As Burr et al. elaborate:

> Importantly, one can describe the goal of the ISA [intelligent software agent] as either "maximising the relevance for the user", or as "maximising the CTR". These two quantities are often conflated in the technical literature, but they are not necessarily aligned. (Burr et al. 2018, 743)

The misalignment of values between providers and users of recommender systems thus represents an obvious source of ethical problems. However, the misalignment does not necessarily have to be intentional but may be the result of the fact that different parties are involved in the engagement of a recommender system. In the simplest case, the parties involved represent opposite end points of a recommender system divided into providers and users, and it seems obvious that their respective ends, values, and preferences do not necessarily coincide. As I argued in Sect. 7.2, even if a recommendation is well-intentioned and based on reasonable standards, this does not ensure the recommendation is a good one because the standard used might not be relevant to the user. The task "Find good items" can simply be realized in many different ways depending on the users' respective ends, values, and preferences. For example, from the user's viewpoint, a good recommendation may consist in items matching long-term preferences, in items representing relevant alternatives to a reference item, or simply in providing satisfying user experiences; from the provider's viewpoint, a good recommendation may consist in changing user behavior in desired ways, in increasing demand and sales, or simply in learning more about customers (Jannach and Adomavicius 2016, 8). Thus, just as being well-disposed towards one another does not rule out misunderstandings, in the same way a user and a recommender system can work at cross-purposes even if it is not designed to "compete" with its users. It may just very well be that a recommender system misidentifies the ends, values, and preferences of users due to its limited data or because of a misinterpretation thereof.

Let us now turn to the level of means. Assuming the ends, values, and preferences of users are respected and not in conflict with the operational goals of a recommender system, is the influence on people's choice architecture to support them in achieving their ends unproblematic? Jesse and Jannach, for example, imagine (but do not discuss) a "nudging-enhanced" recommender system generating recommendations in which recipes for healthy meals are highlighted after having identified the tendency of its user to choose predominantly unhealthy recipes (Jesse and Jannach 2021). Let us flesh out this example further. Imagine the nudging mechanism works successfully. The user chooses and prepares more healthy meals and even experiences a significant increase in health and well-being after some time, all the while not knowing how this improvement in quality of life came about. Would we consider the employment of such recommender systems ethically unobjectionable? After all, the choice for more healthy meals is made by the user, and the user receives real benefits from the engagement with such a recommender system. What is wrong with that?

The idea behind these kinds of arguments is rooted in a particular understanding of decision-making. A "pristine kind of decision-making that is purely deliberate and reflective" was an "unattainable mirage" (Engelen and Nys 2020, 145). Decision-making was always subject to external influences not fully under our control anyways. Even a completely random decision-environment was still a choice architecture influencing us one way or the other. The point they make is that it is impossible to present options in a neutral way. What is true of the food items in the cafeteria was also true in general: options must be arranged and presented in some way or other. This means choice architecture is inevitable – and since people's ends are respected, why not design choice architecture in such a way so as to help decision-makers achieve their self-chosen ends (Sunstein 2015)?

As I argued elsewhere, I doubt that the inevitability of choice architecture gives one license to modify it (Bartmann 2022). Rather, the inevitability of choice architecture puts a particular responsibility on nudgers because the changes made to it may have an effect on people's decision-making. But regardless, what seems to me ethically problematic is neither that there is no neutral design of choice architecture, nor that we rely on automated recommendations in decision-making. The problem arises when we are not aware of factors playing a significant role in our decision-making. This is precisely the problem with the above example of recommendations of healthy recipes because they would exert hidden influence and thus be instances of manipulation. If users are covertly nudged into preparing healthier meals by tapping into their cognitive biases and psychological vulnerabilities, they can no longer make sense of the choices made because they are unaware of the recommender system's rationale for the recommendations. If, on the other hand, users were aware of the rationale – by receiving some indication from the recommender system or by being able to adjust its settings themselves – then the manipulative aspect would disappear. Making the automated recommendations' rationale transparent that way would enable genuine agency because people would not be manipulated but could rather nudge themselves in a self-determined way. Of course, it would not make users' vulnerabilities disappear; however, it would address them in a way that allows for integration in practical reasoning and subsequent action.

In general, being able to discern the standards behind automated recommendations to identify their rationale is essential to ensure they can be integrated into practical reasoning. The danger here is even more profound than the one already mentioned at the beginning of this section (the feeling of alienation one may experience because of the impossibility of engaging in a genuine exchange of reasons with a recommender system). And this more profound danger is present even if the influences on users' choice architecture through recommender systems are not intended to exploit decision-making vulnerabilities: it is the danger of a recommender system's opaque standards to induce *self-alienation* in users by disrupting the connection between users' decisions and their reasons for making them. If I make a decision based on a recommendation whose rationale I do not understand, then I do not really understand why I made the decision because I do not know the reason for it. Decision-making becomes a black box, and this makes the resulting decision alien to me because I cannot recognize the decision as my own.

How can the problem of self-alienation be prevented? As I argued in Sect. 7.2, only if I can identify the rationale behind a recommendation that is aligned with my ends, values, and preferences can I assign it a role in my decision-making and integrate it in practical reasoning. One way of making an automated recommendation's rationale identifiable is by providing an explanation for how the recommender system works, i.e., the principles of selection and filtering behind the generation of recommendations such as: "Customers who bought this item also bought…" (Tintarev and Masthoff 2011, 479). Do explanations like this give a satisfying answer to the question why I am being recommended some item?

Consider, for example, a streaming portal recommending movies to you. The presentation of recommended movies often includes captions such as "Because you watched movie X" or "Others who watched movie X also liked movie Y", examples of content-based filtering and collaborative filtering, respectively. Now, it seems as if these captions provide the recommendations' rationale by giving a reason as to why these movies, and not others, are presented to you. But do they? Even if a particular movie is cited as the reference item used as a basis for recommendations, that still leaves open the movie's particular features used in filtering. What precisely are these features? The specific genre, specific filming locations, specific actors, directors, or screenwriters? It makes a substantive difference if I recommend to you a certain movie because it is an action movie or because the very same movie features your favorite actor. That is why a satisfying answer to the question why I am being recommended a certain item requires the disclosure of an item's features relevant for the principles of selection and filtering involved – the good-making characteristics. Otherwise, an identification of the recommendation's rationale would not be possible. This applies even more to collaborative filtering because they generate recommendations based on other people's preferences I do not know, such that the degree to which the recommendation is opaque to me is even higher.

One may object against this high standard of transparency by pointing out that even if we do not know how a recommendation was generated we are still able to assess the value of the recommended item independently. After all, I could, for example, simply read the abstract or the first few pages of a recommended book to determine whether buying it would be a good choice. That may be true. But recall that, as I argued in the previous section, (automated) recommendations reduce rather than merely rearrange possible options. So, even if I can assess the set of books recommended to me independently, we must not forget the fact that different principles of selection and filtering would have yielded different recommendations as my starting point of assessment, a fact of which I can only make sense if the underlying principles are identifiable. Given the amount of data recommender systems process, this preselection does impact our starting point, and hence our possible choices significantly. Therefore, if opacity and the associated danger of (self-) alienation is to be dissolved and not simply pushed back a step in practical reasoning, the good-making characteristics composing the standard behind a recommendation must be discernible to ensure autonomous decision-making.

7.5 Conclusion

Recommender systems profoundly affect our practical reasoning by influencing our decision-making. Filtering processes employed by recommender systems shape our choice architecture, not just by selecting and reducing the possible options among which we can choose, but also by presenting the remaining options in specific ways. Given the enormous amount of information available online, recommender systems can support our decision-making by providing us with relevant options. But for decision-making to be autonomous and to prevent the danger of self-alienation, we must be able to recognize our choices as our own. I have argued that integrating automated recommendations into our practical reasoning in a non-alienating way requires that we can identify the rationale behind a recommendation to understand a decision based on it. This, in turn, makes it necessary that the good-making characteristics of a recommendation be made transparent, i.e., those features used in the principles of selection and filtering to generate recommendations. Only if a recommendation's rationale is identifiable and thus provides a satisfying answer as to why I am being presented with a particular recommendation can I integrate it in practical reasoning and make autonomous decisions.

References

Bartmann, M. 2022. The Ethics of AI-Powered Climate Nudging – How Much AI Should We Use to Save the Planet? *Sustainability* 14 (9): 5153. https://doi.org/10.3390/su14095153.

Bourget, D., and D.J. Chalmers. 2014. What Do Philosophers Believe? *Philosophical Studies* 170 (3): 465–500. https://doi.org/10.1007/s11098-013-0259-7.

Bozdag, E. 2013. Bias in Algorithmic Filtering and Personalization. *Ethics and Information Technology* 15 (3): 209–227. https://doi.org/10.1007/s10676-013-9321-6.

Bozdag, E., and J. van den Hoven. 2015. Breaking the Filter Bubble: Democracy and Design. *Ethics and Information Technology* 17 (4): 249–265. https://doi.org/10.1007/s10676-015-9380-y.

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

Caraban, A., E. Karapanos, D. Gonçalves, and P. Campos. 2019. 23 Ways to Nudge: A Review of Technology-Mediated Nudging in Human-Computer Interaction. In *Proceedings of the 2019 CHI conference on human factors in computing systems*, 1–15. Glasgow: Association for Computing Machinery.

Carr, A. 2013. *How square register's UI guilts you into leaving tips*. https://www.fastcompany.com/3022182/how-square-registers-ui-guilts-you-into-leaving-tips.

Dubois, E., and G. Blank. 2018. The Echo Chamber is Overstated: The Moderating Effect of Political Interest and Diverse Media. *Information, Communication & Society* 21 (5): 729–745. https://doi.org/10.1080/1369118X.2018.1428656.

Dworkin, G. 2007. Autonomy. In *A Companion to Contemporary Political Philosophy*, ed. R.E. Goodin, P. Pettit, and T. Pogge, 443–451. Oxford: Blackwell Publishing.

Engelen, B., and T. Nys. 2020. Nudging and Autonomy: Analyzing and Alleviating the Worries. *Review of Philosophy and Psychology* 11 (1): 137–156. https://doi.org/10.1007/s13164-019-00450-z.

Flaxman, S., S. Goel, and J.M. Rao. 2016. Filter Bubbles, Echo Chambers, and Online News Consumption. *Public Opinion Quarterly* 80: 298–320.

Hare, R. 1952. *The Language of Morals*. Oxford: Clarendon Press.

Heinrichs, B., and S. Knell. 2021. Aliens in the Space of Reasons? On the Interaction Between Humans and Artificial Intelligent Agents. *Philosophy & Technology* 34 (4): 1569–1580. https://doi.org/10.1007/s13347-021-00475-2.

Herlocker, J.L., J.A. Konstan, L.G. Terveen, and J.T. Riedl. 2004. Evaluating Collaborative Filtering Recommender Systems. *ACM Transaction of Information Systems* 22 (1): 5–53. https://doi.org/10.1145/963770.963772.

Jannach, D., and G. Adomavicius. 2016. Recommendations with a Purpose. In *Proceedings of the 10th ACM conference on recommender systems*, 7–10. Boston: Association for Computing Machinery.

Jannach, D., Z. Markus, F. Alexander, and F. Gerhard. 2011. *Recommender systems. An introduction*. Cambridge: Cambridge University Press.

Jennings, B. 2007. Autonomy. In *The Oxford Handbook of Bioethics*, ed. B. Steinbock, 72–89. Oxford: Oxford University Press.

Jesse, M., and D. Jannach. 2021. Digital Nudging with Recommender Systems: Survey and Future Directions. *Computers in Human Behavior Reports* 3: 100052. https://doi.org/10.1016/j.chbr.2020.100052.

Kahneman, D. 2012. *Thinking, Fast and Slow*. London: Penguin.

Koene, A., E.P. Vallejos, C.J. Carter, R. Statache, S. Adolphs, C. O'Malley, T. Rodden, and D. McAuley. 2015. Ethics of Personalized Information Filtering. In *Proceedings of the 2015 international conference on internet science*, 123–132. Brussels: Springer.

Milano, S., M. Taddeo, and L. Floridi. 2020. Recommender Systems and Their Ethical Challenges. *AI & SOCIETY* 35 (4): 957–967. https://doi.org/10.1007/s00146-020-00950-y.

Noggle, R. 2018. Manipulation, Salience, and Nudges. *Bioethics* 32 (3): 164–170. https://doi.org/10.1111/bioe.12421.

———. 2022. The Ethics of Manipulation. In *The Stanford Encyclopedia of Philosophy*, ed. E.N. Zalta. Stanford: Stanford University.

Pasquale, F. 2006. Rankings, Reductionism, and Responsibility. *Cleveland State Law Review* 54 (1): 115–138.

Rosati, C.S. 2016. Moral Motivation. In *The Stanford Encyclopedia of Philosophy*, ed. E.N. Zalta. Stanford: Stanford University.

Schmidt, A.T., and B. Engelen. 2020. The Ethics of Nudging: An Overview. *Philosophy Compass* 15 (4): e12658. https://doi.org/10.1111/phc3.12658.

Schneider, C., N. Weinmann, and J. vom Brocke. 2018. Digital Nudging: Guiding Online User Choices Through Interface Design. *Communications of the ACM* 61 (7): 67–73. https://doi.org/10.1145/3213765.

Schroeder, M. 2010. *Noncognitivism in Ethics*. London: Routledge.

Sunstein, C.R. 2015. The Ethics of Nudging. *Yale Journal on Regulation* 32: 415–450. https://doi.org/10.2139/ssrn.2526341.

Susser, D., R. Beate, and N. Helen. 2019. Technology, Autonomy, and Manipulation. *Internet Policy Review* 8 (2). https://doi.org/10.14763/2019.2.1410.

Thaler, R.H., and C.R. Sunstein. 2008. *Nudge. Improving Decisions about Health, Wealth, and Happiness*. New Haven: Yale University Press.

Tintarev, N., and J. Masthoff. 2011. Designing and Evaluating Explanations for Recommender Systems. In *Recommender Systems Handbook*, ed. F. Ricci, L. Rokach, B. Shapira, and P.B. Kantor, 479–510. Dordrecht: Springer.

Vugts, A., M. Van Den Hoven, E. De Vet, and M. Verweij. 2018. How Autonomy is Understood in Discussions on the Ethics of Nudging. *Behavioural Public Policy* 4 (1): 108–123. https://doi.org/10.1017/bpp.2018.5.

Waldron, J. 2005. Moral Autonomy and Personal Autonomy. In *Autonomy and the Challenges to Liberalism: New Essays*, ed. J. Anderson and J. Christman, 307–329. Cambridge: Cambridge University Press.

Wallace, R.J. 2020. Practical Reason. In *The Stanford Encyclopedia of Philosophy*, ed. E.N. Zalta. Stanford University.

Weinmann, M., C. Schneider, and J. vom Brocke. 2016. Digital Nudging. *Business & Information Systems Engineering* 58 (6): 433–436. https://doi.org/10.1007/s12599-016-0453-1.

Chapter 8
Recommending Ourselves to Death: Values in the Age of Algorithms

Scott Robbins

Abstract Recommender systems are increasingly being used for many purposes. This is creating a deeply problematic situation. Recommender systems are likely to be wrong when used for these purposes because there are distorting forces working against them. RS's are based on past evaluative standards which will often not align with current evaluative standards. RS's algorithms must reduce everything to computable information – which will often, in these cases, be incorrect and will leave out information that we normally consider to be important for such evaluations. The algorithms powering these RSs also must use proxies for the evaluative 'good'. These proxies are not equal to the 'good' and therefore will often go off track. Finally, these algorithms are opaque. We do not have access to the considerations that lead to a particular recommendation. Without these considerations we are taking the machine's output on faith. These algorithms also have the potential to modify how we evaluate. YouTube has modified its algorithm explicitly to 'expand our tastes'. This is an extraordinary amount of power – and one that if my first argument goes through, is likely to take us away from the good. This influences our behavior which feeds back into the algorithms that make recommendations. It is important that we establish some meaningful human control over this process before we lose control over the evaluative.

Keywords Meaningful human control · AI ethics · Explainable AI · Artificial intelligence

S. Robbins (✉)
University of Bonn, Bonn, Germany
e-mail: srobbins@uni-bonn.de

S. Genovesi et al. (eds.), *Recommender Systems: Legal and Ethical Issues*, The International Library of Ethics, Law and Technology 40,
https://doi.org/10.1007/978-3-031-34804-4_8

8.1 Introduction

Recommender systems (RSs) have driven much of the digital services that we have today. Businesses use them for hiring and keeping your attention while consumers use them to find everything from flights to songs. It is difficult to find an online service that doesn't use an RS. While RSs have made our digital lives convenient, academics have raised ethical issues associated with them – including privacy violations, inappropriate content, fairness issues, and threats to autonomy and personal identity (Milano et al. 2020).

Indeed, with recommender systems, there have been clear consequences resulting from these ethical issues. The title of this chapter may be alarming; however, in some cases recommendations have been accused of causing death. There is the case of Molly Russel in the UK who killed herself after viewing graphic images of suicide and self-harm on Instagram (*BBC News* 2020). Then there is the case of genocide of the Rohingya in Myanmar. Facebook was sued for 150 billion pounds because it failed to prevent the amplification of hate speech and misinformation which lead to offline violence (Mozur 2018; Milmo 2021).

While these cases deserve attention, in this chapter I am not necessarily referring to literal death. With this chapter I show that many RSs are responsible for a vicious cycle that takes our most important judgments outside of our control. These judgements I am referring to are *evaluative* judgements that have ethical value associated with them – particularly when this judgement is about human beings. Recommending someone for a job or evaluating someone on their criminal risk level have important consequences. These judgments deserve the highest level of scrutiny.

In scrutinizing these types of recommendations in the context of using algorithms to arrive at them, at least four distorting forces exist that should make these evaluations suspect. First, because they depend upon machine learning algorithms, they are inherently based upon past evaluative standards.[1] This may not always be a bad thing, but rarely is the fact that "this is the way we have always done it" a good argument in favor of continuing to do it that way. Second, because we are using machines, we need all the input data to be machine readable. This necessarily forces us to reduce complex things like emotions and character traits to computable information[2] or to ignore them completely. Either option will distort the outputs in an undesirable way. Third, algorithms need some proxy to maximize that serves as its notion of the 'good' (Braganza 2022). Engagement, clicks, buys, etc. have all played a role. All have led to serious consequences including teen suicide, election

[1] Machine learning algorithms must be trained on data obtained in the past. National security agencies, for example, have trained algorithms to detect cell phone usage patterns that are associated with terrorists. However, this often fails as terrorists – and society – change their habits (Robbins 2022). Machine learning algorithms do well when there is a fixed target – something that is not true with evaluative standards (Robbins 2021).

[2] It is important to note that while there is much hype regarding emotion recognition in AI, those claims are rooted in the idea that inner emotional states can be detected by, for example, facial expression. This has since been shown to be false (Barrett et al. 2019).

interference, radicalization, etc. There is no doubt that revenue has increased due to these proxies; however, it is also clear that these proxies distort recommendations. Finally, the process by which these recommendations are generated is opaque. That is, the considerations that led to the recommendation are unknown. This leads us astray from the actual object of evaluation – which are the reasons for a recommendation – not necessarily the recommendation itself.

The concern of this chapter is not only that we are likely to have bad recommendations when it comes to the evaluative; it is also that these recommendations serve to change the way we evaluate. This has already had an effect on how, especially young women, evaluate their bodies (see e.g. Cohen et al. 2017). Just like algorithms learn from our behavior, we will 'learn' from the recommendations that the algorithms give to us. Large technology companies know this and use this to their advantage as they try to 'expand our tastes' to generate more engagement and revenue (Roose 2019). This, however, is taking humans out of the driver's seat when it comes to setting up evaluative standards. It is fundamentally a human enterprise to determine how the world *ought* to be. Algorithms can serve to help us achieve that world – but given the distorting forces I have outlined; they should never determine it.

8.2 Distorting Forces

The premise of this chapter is that recommender systems will inevitably play a role in the formation of our evaluative standards. Recommender systems telling us what news to read, movies to watch, people to hire, music to listen to, etc. will shape how we see the world. When Russia invaded the Ukraine in 2022 it was on the front page of every newspaper. The weeks after featured Ukraine on the front page every day and my news recommender system (Google News) also had it as the main headline. Today (26 July 2022) the New York Times still has stories about Ukraine on the front page. However, both the generic (non-signed in) and personalized Google News pages feature no stories specific to the war in Ukraine.[3] There is no way to verify if this is the case for everyone; however, the point is to show that how professional journalists rate the importance of news can greatly diverge from that of news recommender systems.

I am not going to argue whether stories about the Ukraine war *should* be featured as front-page news or not. Rather, what stories are shown as front-page news influences what stories we think are important. While the New York Times has an editor-in-chief and an editorial board deciding what is important, news recommender systems have an algorithm.

I cannot simply say that an editor is better than an algorithm. The New York Times has, for example, been criticized for coverage that distorted the truth. It has

[3] Though my personalized Google News page does feature a story about EU nations agreeing to use less gas in order to prepare for the winter as there are worries about Russia using gas as "a weapon" (Ainger and Nardelli 2022).

also been argued that there is a consistent bias in the New York Times towards corporations (Herman and Chomsky 2011) and that they propagated unchecked claims that Saddam Hussein was manufacturing weapons of mass destruction (*The New York Times* 2004). Algorithms should also receive this type of scrutiny. What biases do they have? How could this go wrong?

The claim that is being made in this chapter is that there are forces that will distort the outputs of algorithms when it comes to the evaluative. These forces are: (1) that recommender system outputs are based solely on past evaluative standards; (2) that the outputs are based on solely on computable information; (3) that these algorithms are taught using proxies for good; and (4) the algorithms are black boxed – meaning we cannot know the considerations which lead to the output – which as I will argue is what matters when evaluating these evaluative judgments. These distorted outputs then feed into our future evaluative judgements. This creates a vicious cycle whereby humans are taken out of the driver's seat when it comes to the evaluative. We are in danger of losing meaningful human control over what we ought to do and how the world ought to be. We must ensure that we retain control over how the world ought to be and then only use technology to help us realize that world.

8.2.1 Past Evaluative Standards

If we take our personal aesthetic or ethical standards, we know that some of them have changed over time. What we once thought was a good movie is no longer a good movie (can you imagine what it would be like if your standards on films didn't change since you were a child?). Many across the world have come to the decision that eating meat is a moral wrong. The changes brought by the pandemic have introduced many novel moral norms regarding wearing masks and coming to work sick.

What this points to is that our evaluative standards have changed. Not simply subjectively, but the context over time has changed which has necessitated new evaluative standards. Recommender systems are built on the premise that past evaluative standards were good in the first place. But this is simply going to be untrue in many cases.

Amazon used an algorithm to recommend applicants for jobs. The algorithm used past hiring habits – which could be translated as past evaluative standards regarding who would be a good employee. Of course, it was quickly understood that those past evaluative standards included the idea that men were better employees for higher up positions than women (Dastin 2018). The past evaluative standards of Amazon's hiring were not good – and therefore the algorithm's standards were also not good.

It is also easy to see that some evaluative standards that were good may not continue to be good in the future. Continuing to use the example of hiring, it will be the case that the profile and character of what constituted a good employee for a specific role in the past will be different from what constitutes a good employee in the future. People with gaps in their CV, for example, were often seen negatively.

However, it has since been pointed out that this benefits men who historically have taken little time off to take care of a newborn. Coming to this understanding changes how we evaluate potential hires. In the academic world there is a constant debate over what standards academics should be judged by (Hicks et al. 2015). These will necessarily evolve. Simply training an algorithm on data from the past[4] will cement in place an evaluative standard that is most likely incorrect.

8.2.2 Reducing to Computable Information

The data used to train recommender system algorithms must be machine readable data. This means that either information that is not machine readable must be left out or that information that is not machine readable must be converted into machine readable information. Both have problems and will be taken in turn.

Leaving non-machine-readable information out means that a lot of the information that we use to evaluate is deemed unimportant. When evaluating people – whether it is for a job, for prison sentencing, or for their ability to pay back a loan – it is necessary to reduce these people to machine readable format. Their criminal history, financial data, job history, etc. can all be used to evaluate. However, there is a reason that processes of evaluating people often involve open conversations in the form of interviews, depositions, etc. in order to understand the data in context. Someone's CV may look bad because they have a two-year employment gap, but their explanation for that gap could be good (e.g., they had to take care of a dying family member). When information like this doesn't make it to the machine, it is implicitly deemed unimportant – and will therefore affect people with unusual circumstances. This privileges those that have 'normal' lives.

Things get worse when non-machine-readable information is claimed to be made machine readable. This happens with emotion detection, lie detection, pain detection, etc. Often things like this are done by analyzing video or image data and reading people's facial expressions. These expressions are thought to show what people are feeling. However, studies have shown that the scientific basis for this is nonexistent (Barrett et al. 2019). This has not stopped companies and academics from claiming that such methods work. Students are rated for engagement in classes (Goldberg et al. 2021). HireVue claims that their AI-powered video interviewing service can read a candidate's empathy level: "E-Motions measures empathy, defined as an individual's ability to read and recognize emotions in others" (HireVue 2019). Companies and academics also claim to be able to detect 'suspicious' people (Arroyo et al. 2015; Gorilla Technology 2019). There is even software that claims

[4] While data must necessarily come from the past it is an important point to highlight. It should cause us to ask which outputs and contexts where this aspect of algorithms be helpful – and where do the pitfalls lie due to this fact. Societal behavior during the pandemic shifted so drastically, that many algorithms failed (Heavon 2020). Knowing this should give us pause before implementing algorithms.

to provide pain assessment based on facial recognition (PainCheck n.d.). However, these systems are known to be seriously flawed. Studies have shown that negative emotions are assigned more often to people of color (Rhue 2018). The AI Now institute concluded in 2019 that "there remains little to no evidence that these new affect-recognition products have any scientific validity" (Crawford et al. 2019).

The point is that attempting to reduce non-computable information like emotions and character traits to computable information is, to date, not possible. When recommender systems try to make recommendations that require such information, they must either leave it out or fake it. Either way will distort its results in an undesirable way.

8.2.3 Proxies for 'Good'

When ML algorithms are trained, there needs to be some goal built into it for it to know what it is supposed to be getting closer to. For example, a chess playing algorithm has the goal of winning built into it. When it plays a game and loses, it adjusts the statistical weights for the moves made in that game to reflect that loss. With recommender systems powered by ML, a goal is also needed. When a social media feed 'recommends' a post by placing it at the top of your feed, it may have the goal of getting you to click on it, re-tweet it, reply to it, etc. The overall goal of these platforms is to keep their app in the foreground – the focus of your attention (BBC News 2021).

The point is that recommender systems need something to aim for. Whatever that something is deserves scrutiny, as it is – in some sense – a proxy for 'good'. The simplistic logic is that if a recommendation keeps you engaged – then whatever was recommended was good for you. However, this is obviously not necessarily the case. The very opposite may be true. For example, if an algorithm recommends for me to eat French fries, and I indeed order French fries – that does not mean that the French fries are good for me. Steamed broccoli would be much better – even if I do not end up ordering it. When engagement is used as the goal for an ML algorithm, on, for example, a news feed, then it is an implicit assumption that the more engaged the user is with the news the better that news is for them. Any given person might be more engaged with news stories that are written to be misleading – giving people a false impression of what is going on in the world.

Moving away from platforms, we can see something similar with, for example, RSs that recommend job candidates. The algorithm succeeds when the top recommended candidates get hired. But this is simply circular. The entire purpose of the system is to recommend 'good' candidates; however, 'good' is simply a measure of whether or not the candidate gets hired. The company wants to hire good candidates – but whomever they hire will be considered 'good'. The reader here may ask how the algorithm determines the top ranked candidates. This may be based on past hiring decisions by the employer (see problems with this in the preceding section) or even on video analysis of candidates asking questions – which has raised ethical issues surrounding cultural, racial, and gender differences biasing the results. In the next sect. I get more into the opacity of the considerations that lead to an algorithm's output – and why that is a problem. Here it is important to note that because of this opacity we are forced to determine the success of the algorithm based on the subject's engagement with the top recommendations – rather than an evaluation of the recommendation itself.

With recommender systems in evaluative contexts, there is a danger that simplistic goals like 'engagement' or 'clicks' will drive us away from what is good for us in those contexts. Platforms may claim that the goals of their algorithms have nothing to do with the good – they are neutrally trying to get you to engage more with their platform so that they can sell ads. However, determining which job candidates are rated highest, which convicted criminals are rated the riskiest, which news stories are prominently displayed, and social media post is at the top of your feed is a huge responsibility with normative implications.

8.2.4 Black Boxed

If someone were to tell you to that you should quit your job, they are making an evaluative statement. Something to the effect of "it would be good for you to quit your job." I imagine that if you had yet to think about the prospect of quitting your job, then you would not just follow this person's advice without further inquiry. Instead, it would be appropriate to ask why they think that. What are their *reasons* for thinking that you should quit your job. Without those reasons the judgment is worth little. For it is the reasons that need scrutiny. For example, their reasoning might be that academia doesn't pay enough and you could make much more in the private sector. This reason may or may not be a good reason *for you*.

With AI powered recommender systems, we have judgements without the reasons. They are black boxed because, as it stands, it is not possible to understand the reasoning for an output of machine learning systems. While this is often portrayed as ML's biggest problem, it is also the source of its power. ML is not restricted to reasons that are articulable to humans – they have access to patterns and considerations that would be impossible for us to comprehend (Robbins 2019, 2020). If we restricted machines to human articulable reasons, like we did with expert systems or good old-fashioned AI (GOFAI), then we would not have had the breakthroughs in AI we have had today.

However, a decision about whether to quit your job seems to require human articulable reasons. In fact, it is difficult to find ethical or aesthetic decisions that do not require justifying reasons (which are human articulable reasons that justify a particular decision). This presents a problem for ML powered recommender systems used for such outputs. For using these recommender systems implies that reasons are not important for the outputs. To take another example, ML systems used for recommending people to be hired for positions implies that the reasons for hiring a particular person are not important. It does this because whatever internal logic that the recommender system is using is opaque to us. There may be reasons (human inarticulable) that could justify the output. However, without access to these reasons we cannot question their ability to justify the output. When we take such recommendations without the ability to question the reasons which lead to that recommendation, we are implicitly disregarding the importance of those reasons. This is a problem because the reasons justifying why you hire one person over another are of the utmost importance. Amazon's ill-fated hiring tool mentioned earlier in this chapter highlights this point. A reason for hiring one person rather than another was gender. Knowing that gender is a reason precludes someone that wants to make ethical decisions from accepting the output of the system.

While using gender as a reason to hire one person over another is almost always unethical there are other reasons that are seemingly irrelevant that also have ethical import. If the number of letters in someone's name or the number of lines on someone's CV were to be used to hire one person over another it would also be unethical. These are not good reasons to hire or not to hire someone. A hiring committee could decide to use such considerations because they were going with a random approach – which would not be unethical; however, a machine doing this without our knowledge that these were considerations would cause us to place higher evaluative weight on a candidate based on reasons that are irrelevant to their candidacy. If a human being recommended someone for a job based on the number of letters in their name without telling you that was a consideration – then it would be deceptive at best. Knowing that this situation is possible with algorithms should give us pause. When we do not know what the reasons are for a particular ML recommendation, then we are left without a way to evaluate whether the output is acceptable. In cases like hiring, not having the reasons for why a particular person was hired is unacceptable.

This has led some to argue that what is needed is so-called 'explainable AI' (XAI) (Floridi et al. 2018; Wachter et al. 2017). Many researchers are now working on trying to make ML explainable (see e.g. Adadi and Berrada 2018; Linardatos et al. 2021). While there has been some success, we are a long way from knowing the reasons that justify a particular output – which is what is needed in situations described above.

This all goes to show that ML based recommender systems should probably not be used for evaluative outputs like hiring and anywhere reasons are important. What I want to argue for now is that it is worse than it looks. These systems also have the capacity to influence our evaluative reasoning. To highlight this let us look at news recommender systems.

What is going on in the world is important. We form opinions about where resources should be allocated (frequent stories about traffic jams may make one conclude that we need to fund a light rail), who to vote for (a candidate may be involved in a scandal reported by the news causing you to vote for someone else) and gives you an overall picture of what is currently going on in the world. We can easily see the ethical import of this in the U.S. right now (July 2022). The hearings are going on in in Washington D.C. regarding the pro-Trump violence in the capital on January 6, 2020. All but one news network is broadcasting the hearings live. Fox News has decided it is not important enough to broadcast (Peters and Koblin 2022).

8.3 Changing Human Values

The ultimate concern that this chapter is focusing in on is that recommendation algorithms are not simply trying to figure out what is good for you. They are, intentionally or not, influencing what you think is good. They are changing your behavior to realize their goals. When I say that a recommender algorithm has a goal, I do not mean that the algorithm is an agent with its own goals. These goals are given to the algorithm. They are 'surrogate agents.' Despite not being conscious or intelligent, they are able to act as surrogate agents and act on behalf of those human agents that gave them their goals (Johnson and Powers 2008).

Information released by former employees at YouTube has shown that the designers of these recommender systems understand that their algorithms change user behavior – and that this is the point. Platforms like YouTube make money by selling advertisements. They have a financial incentive to keep you engaged on their platform for as long as possible so that you see a maximum number of ads. This incentive has driven changes in the recommender algorithm. For example, in response to users getting bored of watching recommended videos that were simply very similar to things they had already watched, Google built an RS called ReinForce. This algorithm was designed to "maximize users' engagement over time by predicting which recommendations would expand their tastes" (Roose 2019).

The idea that algorithms could have the power to 'expand our tastes' should be of the utmost concern. Remember – the algorithm is not driving you towards some agreed upon 'good'. The algorithm is simply maximizing user engagement. So, in sum, the algorithm is designed to change what you value so that you spend more time on YouTube. The entire project is premised upon the idea that an algorithm can take some control over what a person values. In other words, recommendations can impact how an individual values. This is an extremely important point. While ReinForce was designed to change values, other recommendation systems may change values without such malicious intent.

The concern is that algorithms could influence how we come to view what a good X is – in light of its recommendations. For example, if you were on a hiring committee and were given 20 CVs to review and you picked your top 5 which had zero overlap with the 5 that were selected by a hiring algorithm, you could either believe that the algorithm was wrong, or you might adjust the standards you use considering the recommendations (though you may not consciously do this). Though many of us are sure to claim that we would never simply take the algorithm's recommendations as truth that would override our own intuitions and standards, the situation is far from clear.

Humans suffer from various biases which may cast doubt as to whether we, as humans, will be able to prevent recommender algorithms from influencing our evaluative standards. Automation bias, for example, "occurs when a human decision maker disregards or does not search for contradictory information in light of a computer-generated solution which is accepted as correct" (Cummings 2012). This has resulted in disasters like the 1983 Korean Air flight which was shot down in Soviet airspace after the automated system was incorrectly setup and not monitored by the flight crew (Skitka et al. 1999). We humans, tend to be biased towards machine solutions and do not tend to verify every solution offered by these machines. Piling onto that bias is confirmation bias, which biases us towards looking for confirming data rather than disconfirming data (Cummings 2012).

This gets much more complicated with ML powered recommender systems – as we have very little information to go on to confirm, disconfirm, or in any way check the output of such a system. In evaluative cases, there may be simply no way to check. When YouTube recommends you a video to watch next – there is little by which we can check to see if that is indeed the best video to follow the video that we previously watched. Having an algorithm recommend to us the five best candidates in a pool of 1000 is near impossible to verify. This is in part because the real evaluation – as discussed earlier – should be made regarding the evaluative standards. Are these the right standards? With ML we will not know what those standards are. I am not trying to say that there is some objective set of standards that we have that are perfect. What I mean to say is that what we need to do is evaluate the standards themselves. Whatever process we use to, for example, hire someone – we must do it in a way that we can critique and question the evaluative standards used.

This does not stop us from being influenced by those standards. It will be difficult to understand the effect that these multitude of recommendations will have on us – especially children. The news, music, videos, people, products, etc. that are recommended to us will no doubt inform our understanding of what 'good' is in whatever context. The hypothesis is that this will affect our behavior – which will feedback into the algorithms that make recommendations. This vicious cycle would wrest control over the evaluative from human beings and give it to machines. It is necessary to study how people are influenced by recommender systems, and how we can mitigate those influences before we have lost control over the evaluative.

8.4 Same Problem with Humans?

Here, there may rise an objection due to the idea that our access to humans' reasons for recommendations are also not accessible to us. This is what Jocelyn Maclure calls the argument from "limitations of the human mind":

> Decision-making, either by human beings or machines, lacks transparency. As was abundantly shown by researchers in fields such cognitive science, social psychology, and behavioural economics, real world human agents are much less rational than imagined by either some rationalist philosophers or by rational choice theorists in the social sciences. (Maclure 2021)

Top AI researchers, including Geoffery Hinton have made a point like this:

> When you hire somebody, the decision is based on all sorts of things you can quantify, and then all sorts of gut feelings. People have no idea how they do that. If you ask them to explain their decision, you are forcing them to make up a story. Neural nets have a similar problem (Simonite 2018)

It is an intuitive point – especially in light of some of the research that has been done by Daniel Kahneman (2013; Kahneman et al. 2021) and Jonathan Haidt (2001). They have made the empirical case that humans do not simply act in light of the reasons that they claim to have used. Rather, it is the other way around. They make up the reasons for their actions after the fact. Their actions were influenced by several situational factors that were outside of their control and unknown to them. This is supported by a number of studies that, for example, show that judges give harsher sentences before lunch than afterwards (Danziger et al. 2011). The implication is that how hungry a judge is affects sentencing decisions – though no judge has used that as a reason to hand down harsh sentences. So, though it feels like we have good reasons for human decisions, it is simply a myth. Therefore, there is no reason to not use machines simply because the process for reaching outputs is opaque.

However, this misses some crucial points. First, if expensive machines are simply re-creating the problems that we have with humans – then there seems to be no reason to use the machines. The burden is on those pushing for the adoption of these systems to show a good reason to use them. This is especially true considering the wealth of ethical issues associated with these systems. Most importantly are the environmental costs of these systems (Crawford 2021; van Wynsberghe 2021; Robbins and van Wynsberghe 2022).

More to the heart of the matter, Maclure makes the point that institutions and processes are designed to account for human biases and deficiencies: "in non-ideal normative theory, none of these institutions are seen as perfectly capable of neutralizing human foibles, but they can be criticized and continuously improved" (Maclure 2021). The fact that, for example, judges hand down harsher sentences when they are hungry can be mitigated with better scheduling. Institutions and people can be criticized for their failures and can change due to public scrutiny. Also, individuals – especially professionals like judges – must take moral and legal responsibility for their decisions. This is something that RSs – and machines in general – cannot accept. So, an argument would have to be made that decisions like these can be taken by agents that cannot accept moral or legal responsibility before we delegate such decisions to them.

8.5 Conclusion

Recommender systems are increasingly being used for many purposes. With this chapter I have shown that this is creating a deeply problematic situation. First, recommender systems are likely to be wrong when used for these purposes because there are distorting forces working against them. RS's are based on past evaluative standards which will often not align with current evaluative standards. RS's algorithms must reduce everything to computable information – which will often, in these cases, be incorrect and will leave out information that we normally consider to be important for such evaluations. The algorithms powering these RSs also must use proxies for the evaluative 'good'. These proxies are not equal to the 'good' and therefore will often go off track. Finally, these algorithms are opaque. We do not have access to the considerations that lead to a particular recommendation. I have argued that it is precisely these considerations that are used to evaluate whether or not a particular recommendation is good. Without these considerations we are taking the machine's output on faith.

Second, I have shown that these algorithms can modify how we evaluate. YouTube has modified its algorithm explicitly to 'expand our tastes'. This is an extraordinary amount of power – and one that if my first argument goes through, shows that these algorithms will be expanding our tastes in a manner that is likely to take us away from the good. This influences our behavior which feeds back into the algorithms that make recommendations. It is important that we establish some meaningful human control over this process before we lose control over the evaluative.

Finally, I have anticipated that readers may say that the way we receive recommendations without RSs is also problematic. There is no way to verify that someone's recommendation of a job candidate, movie, prison sentence, etc. is 'good'. To this I have replied in two ways. First, that if all things were equal, and we can't verify either, why would we use a machine which has environmental and economic costs over a human being? Second, that things are not equal – machines cannot accept the moral responsibility required to make evaluative choices that cannot be verified by human beings. Giving up this control is giving up control over the evaluative – something that requires good reasons which have yet to be offered.

References

Adadi, Amina, and Mohammed Berrada. 2018. Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI). *IEEE Access* 6: 52138–52160. https://doi.org/10.1109/ACCESS.2018.2870052.

Ainger, John, and Alberto Nardelli. 2022, July 26. EU Nations Reach Agreement to Reduce Gas Use for Next Winter. *Bloomberg.Com*. https://www.bloomberg.com/news/articles/2022-07-26/eu-nations-reach-agreement-to-reduce-gas-use-for-next-winter.

Arroyo, Roberto, J. Javier Yebes, Luis M. Bergasa, Iván G. Daza, and Javier Almazán. 2015. Expert Video-Surveillance System for Real-Time Detection of Suspicious Behaviors in Shopping Malls. *Expert Systems with Applications* 42 (21): 7991–8005. https://doi.org/10.1016/j.eswa.2015.06.016.

Barrett, Lisa Feldman, Ralph Adolphs, Stacy Marsella, Aleix M. Martinez, and Seth D. Pollak. 2019. Emotional Expressions Reconsidered: Challenges to Inferring Emotion From Human Facial Movements. *Psychological Science in the Public Interest: A Journal of the American Psychological Society* 20 (1): 1–68. https://doi.org/10.1177/1529100619832930.

BBC News. 2020, September 26. Molly Russell Social Media Material 'Too Difficult to Look At.' London. https://www.bbc.com/news/uk-england-london-54307976.

———. 2021, September 6. TikTok Overtakes YouTube for Average Watch Time in US and UK. *BBC News*. Technology. https://www.bbc.com/news/technology-58464745.

Braganza, Oliver. 2022. Proxyeconomics, a Theory and Model of Proxy-Based Competition and Cultural Evolution. *Royal Society Open Science* 9 (2): 211030. https://doi.org/10.1098/rsos.211030.

Cohen, Rachel, Toby Newton-John, and Amy Slater. 2017. The Relationship Between Facebook and Instagram Appearance-Focused Activities and Body Image Concerns in Young Women. *Body Image* 23 (December): 183–187. https://doi.org/10.1016/j.bodyim.2017.10.002.

Crawford, Kate. 2021. *The atlas of AI: Power, politics, and the planetary costs of artificial intelligence*. Yale University Press.

Crawford, Kate, Roel Dobbe, Theodora Dryer, Genevieve Fried, Ben Green, Elizabeth Kaziunas, Amba Kak, et al. 2019. *AI Now 2019 Report*. New York: AI Now. https://ainowinstitute.org/AI_Now_2019_Report.pdfw2hwdlsKFcce1B1wW0ucWRL.

Cummings, Mary. 2012. Automation Bias in Intelligent Time Critical Decision Support Systems. In *AIAA 1st intelligent systems technical conference*. American Institute of Aeronautics and Astronautics. https://doi.org/10.2514/6.2004-6313.

Danziger, Shai, Jonathan Levav, and Liora Avnaim-Pesso. 2011. Extraneous Factors in Judicial Decisions. *Proceedings of the National Academy of Sciences* 108 (17): 6889–6892. https://doi.org/10.1073/pnas.1018033108.

Dastin, Jeffery. 2018, October 10. Amazon Scraps Secret AI Recruiting Tool That Showed Bias against Women. *Reuters*. https://www.reuters.com/article/us-amazon-com-jobs-automation-insight-idUSKCN1MK08G.

Floridi, Luciano, Josh Cowls, Monica Beltrametti, Raja Chatila, Patrice Chazerand, Virginia Dignum, Christoph Luetge, et al. 2018. AI4People – An Ethical Framework for a Good AI Society: Opportunities, Risks, Principles, and Recommendations. *Minds and Machines* 28 (4): 689–707. https://doi.org/10.1007/s11023-018-9482-5.

Goldberg, Patricia, Ömer Sümer, Kathleen Stürmer, Wolfgang Wagner, Richard Göllner, Peter Gerjets, Enkelejda Kasneci, and Ulrich Trautwein. 2021. Attentive or Not? Toward a Machine Learning Approach to Assessing Students' Visible Engagement in Classroom Instruction. *Educational Psychology Review* 33 (1): 27–49. https://doi.org/10.1007/s10648-019-09514-z.

Gorilla Technology. 2019. IVAR Edge AI from Gorilla. Gorilla Technology – Products – IVAR. https://www.gorilla-technology.com/IVAR.

Haidt, Jonathan. 2001. The Emotional Dog and Its Rational Tail: A Social Intuitionist Approach to Moral Judgment. *Psychological Review* 108 (4): 814–834. https://doi.org/10.1037/0033-295X.108.4.814.

Heavon, Will. 2020, May 11. Our Weird Behavior During the Pandemic Is Messing with AI Models. *MIT Technology Review*. https://www.technologyreview.com/2020/05/11/1001563/covid-pandemic-broken-ai-machine-learning-amazon-retail-fraud-humans-in-the-loop/.

Herman, Edward S., and Noam Chomsky. 2011. *Manufacturing consent: The political economy of the mass media*. Knopf Doubleday Publishing Group.

Hicks, Diana, Paul Wouters, Ludo Waltman, Sarah de Rijcke, and Ismael Rafols. 2015. Bibliometrics: The Leiden Manifesto for Research Metrics. *Nature* 520 (7548): 429–431. https://doi.org/10.1038/520429a.

HireVue. 2019. HireVue Delivers Game-Based Assessments for Measuring Job-Related Emotional Intelligence. *Hirevue.Com*. https://www.hirevue.com/press-release/hirevue-delivers-game-based-assessments-for-measuring-job-related-emotional-intelligence.

Johnson, Deborah, and Thomas Powers. 2008. Computers as Surrogate Agents. In *Information technology and moral philosophy*, Cambridge studies in philosophy and public policy, ed. Jeroen van den Hoven and John Weckert, 251–269. Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9780511498725.

Kahneman, Daniel. 2013. *Thinking, fast and slow*. Reprint edition. New York: Farrar, Straus and Giroux.

Kahneman, Daniel, Olivier Sibony, and Cass R. Sunstein. 2021. *Noise: A flaw in human judgment*. William Collins.

Linardatos, Pantelis, Vasilis Papastefanopoulos, and Sotiris Kotsiantis. 2021. Explainable AI: A Review of Machine Learning Interpretability Methods. *Entropy* 23 (1): 18. https://doi.org/10.3390/e23010018.

Maclure, Jocelyn. 2021. AI, Explainability and Public Reason: The Argument from the Limitations of the Human Mind. *Minds and Machines* 31 (3): 421–438. https://doi.org/10.1007/s11023-021-09570-x.

Milano, Silvia, Mariarosaria Taddeo, and Luciano Floridi. 2020. Recommender Systems and Their Ethical Challenges. *AI & SOCIETY* 35 (4): 957–967. https://doi.org/10.1007/s00146-020-00950-y.

Milmo, Dan. 2021, December 6. Rohingya Sue Facebook for £150bn Over Myanmar Genocide. *The Guardian*, Technology. https://www.theguardian.com/technology/2021/dec/06/rohingya-sue-facebook-myanmar-genocide-us-uk-legal-action-social-media-violence.

Mozur, Paul. 2018, October 15. A Genocide Incited on Facebook, With Posts From Myanmar's Military. *The New York Times*, Technology. https://www.nytimes.com/2018/10/15/technology/myanmar-facebook-genocide.html.

PainCheck. n.d. PainChek Pain Assessment Tool – Pain Management Software | PainChek. https://www.painchek.com/. Accessed 13 Sep 2022.

Peters, Jeremy W., and John Koblin. 2022, June 7. Fox News Doesn't Plan to Carry Jan. 6 Hearings Live. *The New York Times*, Business. https://www.nytimes.com/2022/06/07/business/media/fox-jan-6-hearings.html.

Rhue, Lauren. 2018. Racial Influence on Automated Perceptions of Emotions. SSRN Scholarly Paper ID 3281765. Rochester: Social Science Research Network. https://papers.ssrn.com/abstract=3281765.

Robbins, Scott. 2019. A Misdirected Principle with a Catch: Explicability for AI. *Minds and Machines* 29 (4): 495–514. https://doi.org/10.1007/s11023-019-09509-3.

———. 2020. AI and the Path to Envelopment: Knowledge as a First Step Towards the Responsible Regulation and Use of AI-Powered Machines. *AI & SOCIETY* 35 (2): 391–400. https://doi.org/10.1007/s00146-019-00891-1.

———. 2021. Machine Learning & Counter-Terrorism: Ethics, Efficacy, and Meaningful Human Control (Doctoral thesis). Delft, The Netherlands: Technical University of Delft. https://repository.tudelft.nl/islandora/object/uuid:ad561ffb-3b28-47b3-b645-448771eddaff.

———. 2022. Machine Learning, Mass Surveillance, and National Security: Data, Efficacy, and Meaningful Human Control. In *The Palgrave handbook of National Security*, ed. Michael Clarke, Adam Henschke, Matthew Sussex, and Tim Legrand, 371–388. Cham: Springer International Publishing. https://doi.org/10.1007/978-3-030-53494-3_16.

Robbins, Scott, and Aimee van Wynsberghe. 2022. Our New Artificial Intelligence Infrastructure: Becoming Locked into an Unsustainable Future. *Sustainability* 14 (8): 4829. https://doi.org/10.3390/su14084829.

Roose, Kevin. 2019, June 8. The Making of a YouTube Radical. *The New York Times*, Technology. https://www.nytimes.com/interactive/2019/06/08/technology/youtube-radical.html, https://www.nytimes.com/interactive/2019/06/08/technology/youtube-radical.html.

Simonite, Tom. 2018. Google's AI Guru Wants Computers to Think More Like Brains. *Wired*. https://www.wired.com/story/googles-ai-guru-computers-think-more-like-brains/.

Skitka, Linda J., Kathleen L. Mosier, and Mark Burdick. 1999. Does Automation Bias Decision-Making? *International Journal of Human-Computer Studies* 51 (5): 991–1006. https://doi.org/10.1006/ijhc.1999.0252.

The New York Times. 2004, May 26. From the Editors; The Times and Iraq. *World*. https://www.nytimes.com/2004/05/26/world/from-the-editors-the-times-and-iraq.html.

van Wynsberghe, Aimee. 2021. Sustainable AI: AI for Sustainability and the Sustainability of AI. *AI and Ethics* 1 (3): 213–218. https://doi.org/10.1007/s43681-021-00043-6.

Wachter, Sandra, Brent Mittelstadt, and Luciano Floridi. 2017. Transparent, Explainable, and Accountable AI for Robotics. *Science Robotics* 2 (6): eaan6080. https://doi.org/10.1126/scirobotics.aan6080.

Part III
Designing and Evaluating Recommender Systems

Chapter 9
Ethical and Legal Analysis of Machine Learning Based Systems: A Scenario Analysis of a Food Recommender System

Olga Levina and Saskia Mattern

Abstract Laws are the reflection of the ethical and moral principles of the society. While the use of technology influences users' behavior in a pace that is affected by the technology introduction to the market, legal activities can be driven by the society as the results of such interactions. This scenario analysis- based research focuses on a classic but fictional food recommender system and the ethical issues that might occur from its usage. The recommender system is taken here as an example of machine learning-based systems (MLS) that can often be found in the individual, business and administrative applications. The research compares the existing legal solutions, with the focus on the GDPR legislation, and the discovered ethical issues. The ethical analysis is led along the ALTAI principles suggested by the European Commission, the common good approach as well as the general principles constituted in human rights. While the GDPR-based analysis showed that this data- and privacy-based legislation addressed most of the identified ethical issues, questions related to the common good approach in the context of environment and mobility that arise due to the wide spectrum of the MLS usage require further legal discussion. The application of the two approaches shows that conducting the ethical and legal analysis is beneficial for both the designers of such MLS as well as the legal actors. The findings can enhance the design and functions of a user-facing MLS as well as influence or validate legal activities.

Keywords Ethical analysis · Legal analysis · GDPR · IT impact · Affected user

O. Levina
Brandenburg University of Applied Sciences, Brandenburg, USA
e-mail: olga.levina@th-brandenburg.de

S. Mattern (✉)
FZI Forschungszentrum Informatik, Berlin, Germany
e-mail: mattern@fzi.de

S. Genovesi et al. (eds.), *Recommender Systems: Legal and Ethical Issues*, The International Library of Ethics, Law and Technology 40,
https://doi.org/10.1007/978-3-031-34804-4_9

9.1 Introduction

The presence and application of digital services has become an integral part of the personal daily routine as well as within business processes. In the last two decades an increasing number of companies have adopted a business model based on digital services attained from user data. Accompanying these developments are not only the changes in the business processes but also societal attention to the effects and the constitution of these services.

Moral and ethical demands for the design and application of digital technologies, especially in the machine learning context, are steadily rising and are resulting in regulatory activities in some countries. In this paper, we use a scenario case study of a recommender system to identify potential ethical issues within the composition and usage of this digital artifact, i.e., software or an information system, as well as the correspondent legislation in force. We explain what ethical aspects are addressed by legal requirements. We show that without social awareness and responsibility, legal regulations alone cannot guarantee socially compliant information technology (IT). Here, we use the scenario of a food recommender system, the FoodApp, that demonstrates the multi-dimensional effects that such a data-based information system can have. The widespread use of smartphones for purchasing goods and services has led to a debate in research and public about the effects of the recommender systems on users' behavior and thus the potential for ethical considerations. The food recommender systems encapsulated in a digital application delude the relation between the consumer and the market by providing a physical good (food) in a short time after a virtual interaction (app-based selection and purchase). The effects of this interaction on the environment, the business and on the stakeholders involved is largely invisible for the consumer making this scenario helpful for the analysis of ethical issues. Since legal measures incorporate ethical values and norms of a society, this paper uses ethical analysis to identify issues that might occur when using a food recommendation system. The results of the legal analysis of this scenario are then used to identify the ethical issues that lack legal equivalents. These results can be used for public discussion about the scope and necessity of future legal regulations.

Hence, the contribution of the research is twofold. First, the analysis method using a combination of ethical and legal assessment is described as a part of the development process of an IT system. Second, the application of the combined ethical and legal analysis of a Machine Learning-based System (MLS) is presented with the focus on recommender systems and its implications are outlined. For the legal analysis, legislation of the European Union (EU), especially the General Data Protection Regulation (GDPR), is considered. The ethical aspects are analyzed using the data-oriented ethical analysis (Levina 2020) and relying on the ALTAI requirements as suggested by the European Commission (European Commission 2020) and subsequently restated as the suggestion for legislation in April 2021 (European Commission 2021). The ethical analysis looks at the data process within the system's design and identifies some of the relevant points in the development process, where ethical questions arise and influence the design of the system. The legal analysis reveals where the emerging ethical questions are already legally regulated. While the presented results cannot be considered exhaustive as to the incorporation of ethical and legal issues during MLS design and use, it provides an exploration framework for designers, developers and users of such systems as well as for Legal and Information Systems Researchers (ISR) by offering an addition to the ISR research methods.

The paper is structured as follows: A brief overview of the state-of- the-art research on ethical issues related to recommender systems follows the scenario introduction of the FoodApp in Sect. 9.3. The ethical analysis and the legal analysis are conducted in Sects. 9.4 and 9.5 respectively, before the combination of their results is presented and discussed in Sect. 9.6. The paper concludes with an outlook on research perspectives.

9.2 An Example Application: FoodApp- the Application for Meal Delivery

For demonstration purposes of the application of the ethical and legal analysis, a scenario of a fictional food recommendation application, the FoodApp, is used here. The fictional app description is adopted from Levina (2022).

The FoodApp is a fictional application based on a three-sided digital platform that is implemented as a mobile app. It is a branch of the fictional large company Acima that offers on demand individual transportation provided by freelancing drivers. To further explore the transportation market, Acima started the FoodApp, a fast-growing food delivery platform connecting the customer, restaurant owner and the delivery partner. It allows the customer to choose from a large database of participating restaurants and order a menu to be delivered to the customer's address via delivery partners. The eater can choose a specific delivery partner based on the ratings of the currently available partners. The payment process is integrated into the platform as is the real-time tracing of the order delivery.

The platform business goal is "fast and easy food delivery whenever, wherever". To achieve this goal, a MLS, a recommender system, is used to provide the best food suggestions for the user in accordance to the indicated preferences and the order history. The business performance indicators for the FoodApp include the return and re-order customer rates, as well as customer number growth rates. The implemented ML-model is thus optimized to drive user's re-ordering on the platform.

To use the FoodApp, the customer downloads it on the mobile device granting permissions for it to access the location of the device. Further, a profile including information on delivery address, name, e-mail and phone number is required. Payment methods and logging in to the payment provider is further required. No manual modifications concerning data collection by the app are possible. Then, the meal preferences such as preferred cuisine or menu item need to be indicated or a meal can be chosen from the provided suggestions. The first suggestions are based on the historical frequency of the orders made within the community in the area of eater's location. A rating system for restaurant and delivery partner performance is implemented.

The platform gains revenue from the customer via convenience charge, fixed commissions and marketing feeds from the restaurants, while providing the assignments and the payment to the delivery partners, as well as the technical infrastructure for the platform participants. The application is a key driver of Acima's revenue and is a fast- growing meal delivery service with over 15 million users worldwide. Additionally, the platform includes an app for delivery partners that provides the possibility to accept or decline a specific delivery job, monitor the revenues, rate the restaurant's delivery process, as well as provide directions to the restaurant and to the eater.

9.3 Current Approaches to Ethical Analysis of Recommender Systems

Ethical issues in the context of IT-artifacts have gained increasing attention in research over the last decade. Paraschakis explores e-commerce recommender applications and identifies five ethically problematic areas: user profiling, data publishing, algorithms design, user interface design and online experimentations, i.e. exposing selected groups of users to specific features before making them available for everybody (Paraschakis 2016, 2017).

Milano, Taddeo, and Floridi conduct an exhaustive literature review of the research on recommender systems and their ethical aspects and identify six areas of ethical concern: ethical content, i.e. content that is or can be filtered according to societal norms; privacy as one of the primary challenges of a recommender system; autonomy and personal identity, opacity, i.e. lack of explaining how the recommendations are generated; fairness, i.e. the ability to not reflect social biases; polarization and social manipulability by insulating users from different viewpoints or specifically promoting one-sided content (Milano et al. 2019).

Milano, Taddeo and Floridi also show that the recommender systems are designed with the user in mind, neglecting the interests of the variety of other stakeholders, i.e., interest groups that are being directly or indirectly affected by the recommendation (Milano et al. 2019). Polonioli presents an analysis of the most pressing ethical challenges posed by recommender systems in the context of scientific research (Polonioli 2020). He identifies the potential of these systems to isolate and insulate scholars in information bubbles. Also, popularity biases are identified as an ethical challenge potentially leading to a winner-takes-all scenario and reinforcing discrepancies in recognition.

Karpati, Najjar and Ambrosio analyze food recommendation systems and identify several ethically questionable practices (Karpati et al. 2020). They name the commitment to already given preferences and thus to the values of the designers as a contradiction to the potential for ethical content. Privacy, autonomy and personal identity that the authors identify as potentially vulnerable and hence suggest need to be realized via an informed concern and a disclosure about the business model used. Opacity about the origin of the recommendations as well as of the criteria and algorithms used to generate the recommendations. Fairness, polarization and social manipulability as well as robustness of the system complete the list of identified ethical issues for a food recommender.

These approaches discuss ethical impacts of recommender systems from the perspective of the receivers of the recommendations. Milano, Taddeo and Floridi argue that the social effects such as manipulability and personal autonomy of the user are hard to address, as their definitions are qualitative and require the implementation of the recommender system in the context they operate (Milano et al. 2019), while Karpati, Najjar and Ambrosio offer a multi-stakeholder approach to address these issues (Karpati et al. 2020).

The data processing-centered approach to analyzing ethical issues suggested by Levina (2020) identifies the decision points during the MLS development, while advocating the inclusion of a laboratory phase into the system design to assess the potential consequences (see also Coravos et al. 2019). This research applies a data process-centered combination of ethical and legal analysis in the attempt to identify how or whether the identified threads to ethical values that can be realized via an MLS are being already covered by the legislation of the EU and where in the artifact design these can be addressed.

9.4 Ethical Analysis

Here the ethical analysis following the steps of data processing in MLS is applied. To understand ethical implication of a process or technology, specific values that are being affected by the technology deployed need to be taken into account. There are a number of values that can be considered, but here, we look at the values represented as requirements in the "Assessment List for Trustworthy AI (ALTAI)" (European Commission 2020) by the High-Level Expert Group on Artificial Intelligence of the European Commission and that also are the fundament for the regulation proposal for governing artificial intelligence technologies (European Commission 2021) being: Human agency and oversight; technical robustness and safety; privacy and data governance; transparency; diversity, non-discrimination and fairness; societal well-being; accountability. Those have been chosen specifically to assess ethics aspects of AI technology, and are therefore appropriate to our goal.

The data processing-centered ethical analysis puts privacy and data governance in its focus and thus allows a stringent analysis of the legal aspects along the General

Data Protection Regulation (GDPR), EU legislation. The specific stages of the data processing considered here are: *sense*, including: collect data, detect data, and select (problem-related) data; *transform*, including: clean data, chose machine learning model, train model, integrate model into the software system; *act*, including generate recommendations or more general, a result based on the chosen ML-model; and the *apply* stage that considers effects on stakeholders by the results and the effects introduced by the ML-model definition and application. These data processing stages are based on the data management process by Shutt and O'Neill (2013). While the apply stage is not an integral part of the data process, the effects of the application of the MLS on user behavior are important for its design. Furthermore, it is differentiated here between the MLS-user, i.e., a person using the MLS directly, and an MLS-*affected* user, i.e., a person or a stakeholder that is affected by the effects of the MLS.

In the *sense* phase of data processing, it needs to be assured that the data have been collected with the *informed consent and voluntariness* of the data subject. The legal analysis presented in Sect. 9.5 elaborates further on these considerations. Informed consent also includes the statement of the *purpose* of the data collection implying an opt-in function for data collection. What data are being collected is normally described in the *terms and conditions document of the* recommender system. Their legal sufficiency is further discussed in the legal analysis. Often, under the claim of providing more personalized suggestions, profile data provided by the user as well as e.g., the location data automatically provided by the mobile device are also collected. Nevertheless, it is not clear to the user, *when* the location *data are collected* challenging the value of *privacy*, nor is it clear what role these data paly for recommendation creation, addressing the value of transparency. While *opt-out functions* are often cited by the enterprises producing data-based services as user empowerment (Abdelaziz et al. 2019), they are legally insufficient (see legal analysis) and shift too much responsibility towards the user (Milano et al. 2019).

FoodApp's business goal is to engage the user in the re-ordering of food via its digital platform. The user interacts with the app aiming for a comfortable provision of the favorite food in an efficient way. Therefore, the user is inclined to give up some autonomy within this process. Nevertheless, in the digital realm the user is often not aware of what elements of his/her *autonomy* are jeopardized when the digital service, here food selection and ordering via a digital platform, is performed. While the user can still change the app's settings, s/he is often unaware of the default access requirements. Data are the fundament for the further model building for the recommendation algorithm and their amount and sources are strongly defined by the business model that this MLS is supposed to support.

As FoodApp would like their users to return to the app, it will need among other factors, very good recommendation results as well as a frictionless ordering process together with a reliable problem handling mechanisms to fulfill basic customer expectations (Karpati et al. 2020). The first requirement, i.e., very good recommendation results in terms of user's preferences, can be realized using a recommendation algorithm based on the collected data from the user as well as from the users with similar preferences or history on the platform (Karpati et al. 2020). Since the user activity data might provide additional patterns for the recommendation, it also provides a potential reason to keep the user engaged on the app for the longest possible time, which might involve the use of dark patterns in the app design (Gray et al. 2018). FoodApp's user profile provides the information that is, among others, needed for the algorithms in the MLS to derive food recommendations. The user does not have any information about the exact *purpose* of the provided datasets, the *data lifecycle*, nor about who has *access* to the (possibly) un-anonymized profile or historical data and about the *data state timeline*, i.e., when the data are transferred or deleted. These aspects can be categorized as "*transparency* issues", since the user does not have the information about FoodApp's processes s/he might need or would like to have.

Aside from the user data, FoodApp's database would include data on the restaurants available for ordering and delivery through the platform. Addressing the restaurants is part of Acima's business model and might also be part of the business focus of the restaurants, as they can be included on the platform according to specific *criteria*, e.g., reviews on other platforms, personal preferences, number of years in business, etc. leading to a potential pre-selection of available food choice on the platform. Additionally, the delivery network of partners that will pick up food at the restaurants and deliver it to the customer's door need to be established and equipped with the means to be contacted, paid and managed by the platform. Hence, FoodApp needs to establish an *ecosystem*, similar to a classic supply chain, to be able to fulfill its business goal or even to be able to operate according to its business model. Building up such an ecosystem as well as the potential to manage the orders for delivery, provides Acima as a digital platform with a specific power over the delivery partners as well as the restaurants that can have extensional effects on the partners involved in the ecosystem as well as the bigger area of stakeholders (Levina 2019).

The user can filter the suggestions within the FoodApp using the provided *filter categories*. These categories, defined by the MLS-engineers and designers, include cuisine and menu item names, as well as the ratings of the accordant restaurants. In future interactions with the FoodApp its home screen offers the meals and food items that are most frequently ordered by the eater or users that were identified to have a similar ordering behavior, thus nudging the eater to order the same or similar kind of food (Zhou et al. 2010).

Also, how the data are analyzed, i.e., anonymously or connected to the profile might raise further *privacy* and *transparency* questions. In cases when the recommender system dictates what data are collected, the mode of data collection and their importance within the algorithm are also decided without user participation. Hence, the question about the *diversity* and *quality* of data and consequently the ones of the recommendations arise. The lack of diversity can be manifested in other aspects that are indirectly related to the business goal, such as feedback from the stakeholders, the effects on the ecosystem the MLS is acting in as well as the potential impact of its recommendations on the environment. In a case of a food recommender as in (Karpati et al. 2020), these can be a considerable environmental impact in terms of air pollution (Saner 2020) besides the generated plastic waste (Zhong and Zhang 2019).

The *transformation* phase is the last phase within the data process where the fundamental decision to apply machine learning instead of a less resource-intensive solution to realize a business goal can be revised. Training and using an MLS requires increased power usage. Parameters within the MLS that lead to recommendations are designed and configured by data analysts and software engineers, thus reflecting their values and reducing the *diversity* of the recommendations as well as promoting decisional de-skilling (Floridi 2016) and increasing *trust* into the algorithmic suggestions (Gille et al. 2020; Krügel et al. 2021; Fritz et al. 2020) by the user.

To reduce energy and time *resources* to train the particular model in use, an existing pre-trained model form the application domain can be used. Although, the choice of the model often does not consider its *explainability* (Kamiran and Calders 2012) or its requirements for the resources. Its accuracy is often the main criterion, which might neglect user's preferences. Also, observations on the effects of MLS application on user's *behavior* rarely precede the going live of the MLS.

The FoodApp has based its business model on the data-based provision of food recommendations and the forwarding of the recommendations to the restaurants and delivery partners. Thus, being data-based, these business questions would require the use of data analysis tools, although the added value of the neuronal networks for the recommendations depends on the quality of data and the accuracy thresholds defined by the product designers.

The *model quality* is in the center of the ethical inquiry in the transformation phase. The set thresholds define mathematical methods, e.g., neural networks vs., e.g., support vector machines, and thus the resources needed to train the model as well as to generate the recommendation. The transform phase does not only include the training and optimization of the models used for the recommendation, but it also considers the inclusion of the ML-models into the information systems context.

While definition of food categories as well as the selection of the included cuisines and restaurants is part of the *sense* phase and especially the *select* sub-phase, questions in the transform phase focus on the mathematical transformation of these selected details. Inclusion of, e.g., nudging techniques is also part of the sense phase and the collect sub-phase, but it is strongly defined by the *business model*.

The *act* phase comprises the definite initiating of an action based on the provided MLS recommendation. The interaction with the recommender is often implemented as a "human-in-the-loop" interaction pattern that puts the human in *control* only over the recommendation-based action itself. It often leaves the user oblivious about the following process that was initiated by this action, leaving the user oblivious about the consequences that were induced by the action, as no feedback from outside the user-focused process is taken into account.

For Acima, the value is created when the food delivery order is completed in the FoodApp. Hence, the ordering process is organized in a way that no extended explications or additional information are given so that the user does not have to choose, decide or react during the interaction process. This design allows a *fast phase-out* between opening the FoodApp and ordering the food. This effect can be expected to contribute to user satisfaction and thus re-visiting the platform for the next order.

The process efficiency offered by the FoodApp is also built on the lack of decision possibilities and a limited items selection that is based on the historic and profile preferences for the user. Additionally, the gained comfort for the user in terms of food selection and delivery has implications on the *ecosystem* of the FoodApp. The restaurant partners will be faced with an increased amount of reviews from delivery customers, potentially forcing them to concentrate on robust packaging to ensure the sound condition of the meal for delivery. More robust packaging means more damage to the environment but potentially better ratings from the FoodApp users (Zhong and Zhang 2019).

Furthermore, the food recommendations based on historic and similar orders might lead to homogenization of the food offered in the participating restaurants, as menu items that are ordered less often might not be prepared by the restaurants anymore, potentially leading to the decrease or shift of skills of the cooking staff. The individual delivery of the food orders requires reliable and efficient delivery partners. Acima relies here on its network of drivers for personal transportation that are also incentivized to transport food orders via reward programs. This efficient and effortless process of ordering food for individual consumption can and does cause significant environmental damage in terms of air pollution through traffic and waste (Zhong and Zhang 2019).

Further effects on the *social* environment can also occur. The eater rates the restaurant on the food quality and the delivery partner on the quality of the delivery. The rating is based on eaters' satisfaction with the end result, whereas the traffic situation and other external effects of the recommendation process are not considered. This relationship pattern causes societal effects that are visible in the traffic situation, environmental damages as well as reduction of labor costs and conditions and also affect user's behavior (De-Arteaga et al. 2020).

While the analyzed aspects in previous phases were based on the data process as an enabler of the business value creation, the *apply* phase leaves the realm of the software system as an artifact and enters into the realm of its usage that can directly or indirectly influence behavior of individuals. One potential effect might be the reduced diversity in the choices that survive in such ecosystems due to the focus of the recommender system to provide suggestions based on e.g. item's popularity, leading to the extinction of less favored choices as well as impacts on the collateral effects, e.g. effects on air quality, quantity of waste and quality of life via increased delivery traffic are commonly observed in the cities where food recommender are active (Karpati et al. 2020; Saner 2020; Zhong and Zhang 2019). Thus, even if the MLS is not directly critical to the human life, rights or *well-being*, there is a necessity to consider the societal effects of its implementation as outlined in the criticality assessment of algorithmic systems by the German Ethics Commission (Germany. Datenethikkommission 2018) that takes the effects of the application of an MLS on critical goods such as human life and wellbeing into consideration.

In the FoodApp scenario, rating of the delivery partners results in an increasing number of orders for high ranked drivers and in a reduction of delivery orders for the worse ranked drivers. Hence promoting the reviews into the main factor for job acquisition, and thus income, for the drivers. This type of job market is known as the *gig economy* (Friedman 2014). It provides income potential for the workers while creating an interdependency between the platform customer and the gig worker. This relation seems to remain unclear for the platform customer and is often debated by the platform owner (Susser and Grimaldi 2021). Consequently, the OECD stated in 2016 that digital platforms need social values to be reflected in the platform governance (OECD 2016).

9.5 Legal Considerations

Using MLS can cause novel challenges for the legal compliance. Ideally, there is a common goal between legal experts, ethicists and technicians: an optimized and legally compliant design and implementation of machine learning-based artifacts. To further this, some main aspects are presented using the food recommender system as an exemplary scenario.

9.5.1 Data Protection Law

The availability of large amounts of data are essential for machine learning. But when *personal data are* involved, the *European General Data Protection Regulation (GDPR)* comes into play. There are actually very few data, such as pure machine data, which no longer have any personal reference. Personal data refers to all information relating to an identified or identifiable natural person.[1] This broad definition covers all information that can somehow be attributed to a specific person. Even *pseudonymous data,* like an IP address, is classified as personal data, because, although not directly, it can be attributed to a natural person by means of a transfer.[2]

In the FoodApp, user, restaurant and delivery data are collected, selected, processed and stored. All this constitutes *data processing* in the sense of data protection law. It is clear – not only because of the tremendous possible *penalties and fines*[3] imposed by the GDPR that it is absolutely necessary to plan MLS such as FoodApp in compliance with data protection regulations. Albeit, the issues of big data, Machine Learning (ML) and algorithms were not explicitly addressed by the GDPR. Here very few references can be found, e.g., Art. 22 GDPR, which regulates automated individual decision-making, including profiling. Even though Art. 22 GDPR appears to be extremely relevant to ML- processes at first sight, its legal

[1] Definition in Art. 4 No. 1 GDPR.

[2] Recital 50, Kühling/Buchner/Klar, GDPR Comment, Art. 4 No. 1 marginal no. 28.

[3] See Art. 83, 84 GDPR.

scope is actually rather limited. According to the prevailing and convincing view,[4] Art. 22 GDPR applies only to those automated processes, which are intended to evaluate certain characteristics and features of a person, despite its rather open wording.[5] Hence, the data process itself is subject to the general provisions of the GDPR, which were not specifically designed for ML-processes. Subsequently, some points of relevance are presented.

9.5.2 General Principles and Lawfulness of Processing Personal Data

First, Art. 5 GDPR is to be mentioned, which lays down the general principles of the data processing such as lawfulness, fairness and transparency. Art. 5 No. 1 a GDPR reads: "Personal data shall be processed lawfully, fairly and in a transparent manner in relation to the data subject". Equally relevant is Art. 6 GDPR, which regulates the lawfulness of processing. Art. 5 No. 1 a GDPR reads: "Personal data shall be processed lawfully, fairly and in a transparent manner in relation to the data subject". The main scope of the transparency requirements is defined by the information obligations in Art. 12 and Art. 13 GDPR, for the exercise of the rights[6] of the affected *data subject*. Art. 12 GDPR initially states that all information and communications must be provided in a precise, transparent, comprehensible and easily accessible form in clear and simple language. When data for the usage in ML are collected, affected persons need to be explicitly informed about this fact. Thus, the FoodApp should provide explicit mentions of the data being collected and the purpose the data is being used for.

However, *transparency* seems to be relatively difficult to fully accomplish in this context, because this also requires explainability of the decision process in algorithmic decision procedures. It is essential that users of algorithmic systems are at least able to understand, explain and control their operation, and that those affected receive sufficient information to enable them to exercise their rights in Art 12–23 GDPR properly. It appears to be excluded, that the process can be reproduced in all details afterwards.[7] So it is to be assumed that only the principle underlying the algorithm, i.e. the basic assumptions of the logic underlying the algorithm, but not the concrete computation formula or the algorithm itself must actually be presented

[4] If you consider Art. 4 No. 4 GDPR and recital 71 of GDPR.

[5] Recital 71 of GDPR; Art. 4 No. 4 GDPR; Abel: Automatisierte Entscheidungen im Einzelfall gem. Art. 22 DS-GVO (Automated decisions in individual cases according to Art. 22 GDP9", ZD 2018, p. 304.

[6] See Art. 12–23 GDPR.

[7] Rosenthal: GDPR vs. AI, LR 2018, 173; Bitkom, 2018: Machine Learning and the Transparency requirements of the GDPR.

to fulfil the information obligations in Art. 12 and 13 GDPR.[8] The biggest difficulty of transparent presentation here probably lies in describing the complexity of the analytical procedures used in such a clear and easy way that their scope is somehow tangible for the person affected.[9] On the other hand, companies also may claim that the disclosure of the algorithm and data categories would affect their business model by hinting to their business secrets.[10] In order to balance the interests of the MLS-user and the MLS- affected user, the provisions of Art. 42 Paragraph 1 GDPR, as well as data protection certification procedures and data protection seals and test marks, might represent a suitable instrument for proving compliance with the GDPR by the person responsible or their contract processors.[11]

9.5.3 Lawfulness

In addition to the transparency and fairness of the data processing, it must also be lawful under Art. 5 and Art. 6 GDPR. Processing personal data are generally prohibited, unless the data subject has consented to the processing or one of the other legal bases stated in Art. 6 GDPR apply. It seems to be tempting to regulate data processing for ML-Systems in general terms of use or general terms and conditions. If customers want to use a certain service or app, which uses ML, he or she simply has to agree to the terms and thus also to the processing of data for ML-purposes.

Nevertheless, an opt-in or even opt-out in terms of use is not equivalent to *informed consent*. The basic requirements for the effectiveness of valid legal informed consent are defined in Art. 4 and Art. 7 GDPR and specified further in recital 32 of the GDPR. In order to obtain freely given consent, it must be given on a voluntary (free) basis, which implies a real choice by the data subject. Therefore, the tying prohibition[12] applies: the fulfilment of a contract may not be made dependent on the consent to the processing of further personal data which are not necessary for the fulfilment of this contract. Though one can already doubt the voluntary aspect here, there is again a major problem in particular with regard to the transparency of information. For consent to be informed and specific, the data subject must

[8] Art. 29 Data Protection Working Party (WP 251 rev.01) – Guidelines on the Automated individual decision-making and Profiling for the purposes of Regulation 2016/679; Rosenthal: GDPR vs. AI, LR 2018, 173, Ehmann/Selmayr: GDPR Comment, Art. 15 Rn. 12; Paal/Hennemann, GDPR Comment, Art. 13 Rn. 31; Kamlah, in: Plath, BDSG GDPR Comment, Art. 13 No. 28.

[9] Bitkom/DFKI, 2017: KI – Wirtschaftliche Bedeutung, gesellschaftliche Herausforderungen, menschliche Verantwortung (Artificial intelligence: economic significance, social challenges, human responsibility) p. 134; Gausling: Artificial Intelligence and GDPR, DSRITB 2018, p. 519.

[10] In German legislation it ist he „Gesetz zum Schutz von Geschäftsgeheimnissen (GeschGehG)“that is meant to protect the business secrets in legal sense.

[11] Bitkom/DFKI, 2017: “Artificial intelligence: economic significance, social challenges, human responsibility”, p. 134; Gausling: Artificial Intelligence and GDPR, DSRITB 2018, p. 519.

[12] Recital 43 of GDPR, Gola GDPR Comment/Schulz, Art. 7, No. 21–33.

at least be notified what kind of data will be processed, how it will be used and the purpose of the processing operations.[13] If all this is applicable, an opt-in to general terms of use/terms and conditions could be sufficient. An opt-out, on the other hand, is never sufficient, as it must be an unambiguous action.[14]

Secondly, it is questionable whether another legal basis would apply unless consent can be assumed. It would be conceivable in particular that the use of ML is necessary within the meaning of Art. 6 No. 1 b GDPR for the performance of a contract to which the data subject is party. That means, the ML algorithm must be used and also be necessary to enable the service provision to a customer, e.g., food recommender. However, ML will probably be used more often to optimize processes, but not to make them possible in the first place. Another commonly used reason for lawful data processing can be find in Article 6 paragraph 1 lit. f: "processing is necessary for the purposes of the legitimate interests pursued by the controller […]." If a company argues that it has a legitimate interest in using ML, because its services would otherwise not be competitive, for example, this might actually be sufficient in an individual case. As the legal assessment depends on the individual case and interpretation, many uncertainties remain.

9.5.4 Purpose Limitation and Access to Data

Art. 5 No. 1 b GDPR requires a *purpose* limitation, which means that data may only be processed for specified, explicit and legitimate purposes and must not be processed beyond those purposes. Thus, the purpose legitimates the processing of the data in terms of necessity, adequacy, completeness and duration of the processing.[15] Once the purpose is fulfilled, the collected data must be deleted, since the purpose cannot be modified. This is only exceptionally possible if the present and initial purposes are compatible (Art. 6 paragraph 4 GDPR). In order to ascertain whether a purpose of further processing is compatible with the initial purpose, the controller should take several factors into account, e.g., the context in which the personal data have been collected, the nature of the personal data and the consequences of the intended further processing for data subjects.[16] This is interpreted rather restrictively. In this context, it is important to also point out, that the GDPR does not take a stance on the issue of "data ownership". Therefore, the question arises, who actually "owns" the original and generated data, has access to them and can use or even sell them. According to the prevailing opinion "data ownership" does not exist.[17]

[13] Recital 32 of GDPR.

[14] Recital 32 of GDPR.

[15] Paal/Pauly/Frenzel DS-GVO Art. 5 Rn. 23–25.

[16] Recital 50 of GDPR.

[17] Among others: Czychowski/Siesmayer, Computer Law Manual, para. 20.5, margin no. 42; Zech, GRUR 2015, 1151, margin no. 1157; Hoeren, MMR 2019, p. 5; Determann, ZD 2018, p. 503; Working Group "Digital Reboot" of the Conference of the Ministers of Justice of Germany, report

Rather, it seems to be the current consensus to realize data sovereignty in each individual case through contractual agreements between the parties involved in a data exchange. This further underlines and demonstrates the absolute importance of purpose limitations as it restricts the power of disposal and the data subject retains his or her sovereignty.

9.5.5 Data Minimization and Storage Limitation

In order to derive accurate results using MLS, considerable amounts of data are required. These data need to be collected, selected and also stored to be usable. Therefore, the data collection in the FoodApp is already *conceptually opposed* to the principles of data minimization and *storage limitation* in Art. 5 No. 1 c, e GDPR as the data is being collected without the naming of the purpose and without its clear attachment to the business model. According to the principles mentioned above, only as much data as absolutely necessary may be collected and stored for as long as absolutely necessary. So, how can a balance be struck between the conflicting goals of large data collection as exercised by some data-based information systems and data protection? The principle of *data minimization* obliges data controllers in Art. 25 GDPR to design their systems technically in such a way that the risks for data subjects are minimized (privacy by design) and that default settings ensure that only personal data that are necessary for the purpose are processed (privacy by default). As stated in Art. 25 (2) GDPR privacy by default should ensure that personal data is processed with the highest privacy protection. Hence, personal data is made accessible to a definite number of persons and only personal data that is necessary for a specific reason shall be obtained. The principles of data minimization and purpose limitation relate to the concept (Ježová 2020).

As a consequence, the FoodApp would have to provide a data-protection-friendly default setting and offer the user a detailed selection option. In addition, an individual setting by the user, which can be made at any time, must be possible.

Yet here another divergence is evident – manufacturers or software developers de facto have influence on the data protection conformity of data processing processes and are able to implement "Privacy by design". However, manufacturers or software developers, are actually not responsible in the sense of the GDPR, but only possibly commissioned data processor (Art. 24 GDPR). The Controller, the one that uses MLS, stays responsible. In this context Art. 28 GDPR sates that the controller shall use only processors providing sufficient guarantees to implement appropriate technical and organizational measures in such a manner that processing will meet the

of 15.5.2017 https:// jm.rlp.de/fileadmin/mjv/Jumiko/Fruehjahrskonferenz_neu/Bericht_der_AG_Digitaler_Neustart_vom_15._Mai_2017.pdf; German Federal Ministry of Transport and Digital Infrastructure, "Ownership Regulations" for Mobility Data? – A study from a technical, economic and legal perspective", August 2017, https://www.bmvi.de/SharedDocs/DE/Publikationen/DG/eigentumsordnung-mobilitaetsdaten.html requirements of the GDPR and ensure the protection of the rights of the data subject. But this actually does not seem to be sufficient as no real obligation is imposed. The manufacturer or software developers should thus be somehow included in the scope of this provision.[18]

Art. 25 paragraph 1 GDPR describes that *pseudonymization* might help to effectively implement these data protection principles and thus to protect the rights of the data subjects. The exact way in which the legislator envisages this is not explained in any detail, not even in the recitals. Art. 25 GDPR does not provide a solution for the principle of storage limitation, but the legislator has written specific requirements in the recitals of the GDPR: The personal data should be limited to what is necessary for the purposes for which they are processed. In order to ensure that the personal data are not kept longer than necessary, time limits should be established by the controller for erasure or for a periodic review.[19] So, the principle requires erasure routines that take effect at regular intervals and prevent endless storage.

9.5.6 Accuracy, Security and Impact Assessment

Data allow a reconstruction of an individual's characteristics or information and must therefore be accurate in order to permit such a reconstruction. The principle of accuracy is concretized in Art. 5 No.1 d. Accordingly, every reasonable step must be taken to ensure that personal data that are inaccurate, with regard to the purposes for which they are processed, are erased or rectified without delay. Therefore, there should be an obligation to evaluate the systems on a regular basis, as is the case for the security measures under Article 32 GDPR.[20] In its recitals, the legislation makes clear at various points that strict security measures are important, because personal data processing can lead to severe physical, material or non-material damage.[21] In order to maintain security and to prevent processing in infringement of this regulation, the controller or processor should evaluate the risks inherent in the processing and implement measures to mitigate those risks, such as encryption.[22] In Art. 35 GDPR, the legislation provides an instrument for the evaluation of risks, the so called "*data protection impact assessment*". A data protection impact assessment must be established for sensitive processes before they are introduced. The legislator has hereby provided for a risk-based "technology impact assessment", which serves to take a closer look at the effects of the use of new technology on society and

[18] See also Conrad: Künstliche Intelligenz – Die Risiken für den Datenschutz (Artificial intelligence – The risks for data protection), DuD, 12/2017, p.743.

[19] Recital 39 of GDPR.

[20] Conrad: Künstliche Intelligenz – Die Risiken für den Datenschutz (Artificial intelligence – The risks for data protection), DuD, 12/2017, p.743.

[21] Recitals 75, 76, 77, 78, 79, 83 of GDPR.

[22] Recital 83 of GDPR.

the environment.[23] The outcome of the assessment should then be taken into account when determining the appropriate measures to be taken in order to demonstrate that the processing of personal data complies with the GDPR.[24] However, MLS that are trained using external data, might be very easily manipulated. Since the MLS decision criteria are often unknown, it is difficult to foresee and block all conceivable manipulation attempts. Accordingly, it is to be expected that MLS will be subject to special security requirements. Here, special certifications and audits could be helpful in the future.

9.6 Results of the Combined Ethical and Legal Analysis Approach

The ethical analysis was conducted along the data processing steps of an MLS and under the consideration of the ethical values suggested in ALTAI. Ethical aspects discussed here accorded to the ones previously identified by Paraschakis (2016, 2017) for recommender systems in e-commerce, such as user profiling, privacy and online experimentation and Karpati, Najjar and Ambrosio for a food recommender system (Karpati et al. 2020).

Also, some of the identified ethical issues can be associated with the ethical areas of concern found by Milano et al. (2019) such as social manipulability. Additionally, using the FoodApp scenario some aspects were found that cannot be easily added to one of the categories in the previous work, such as environmental concerns induced by the increasing traffic due to the individual food delivery and packaging waste. Questions related to the use of digital applications for individual services resulting in the appearance and stimulation of the gig economy were also evident in the FoodApp scenario. Here, the delivery workers were a major part of the business model, while also facing uncertain labor conditions and minimal autonomy within their employment (Rosenblat and Stark 2016; Doteveryone 2019; Aguiléra et al. 2018). The last issue is being addressed by the European Commission and resulted in a directive that defines platform work as well as the workers' rights, especially in the context of algorithmic decisions (European Commission 2021).[25]

Using the FoodApp scenario several ethical issues were identified. The legal analysis showed that most of these issues have corresponding legal requirements that need to be implemented into the digital product for compliance. On the other hand, the combined ethical and legal analysis identified some ethical issues that are not yet included in the considered legislation.

[23] Conrad: Artificial intelligence – The risks for data protection, DuD, 12/2017, p.743.

[24] Recital 84 GDPR.

[25] Directive Of The European Parliament And Of The Council, 9.12.2021, https://ec.europa.eu/commission/presscorner/detail/en/ip_21_6605

The FoodApp scenario demonstrated the lack of transparency in the collection and use of user data. The purpose and the lifecycle are not visible for the user, depriving him or her from autonomous decision making about the data collection. An informed choice about the impact on user's privacy is thus not possible. These ethical issues are addressed by the GDPR in the Art. 5 that requires specified, explicit and legitimate purpose for data collection. Art. 12 GDPR requires transparent information, communication and modalities for the exercise of the rights of the data subject. The lack of possibility for opting out of specific data collection and therefore the ethical issue of autonomous control over the algorithm parameters of a recommender system. Articles 25 and 32 GDPR require privacy by default and opt-in for the specific data aspects to be collected, making the opt-out option not sufficient for the compliance. Art. 17 of GDPR states the right to erasure by the data subject, thus allowing the user to end the use of her or his data by the recommender system algorithm. Another transparency and privacy aspect identified in the FoodApp scenario is that the FoodApp is part of the Acima enterprise. Since the user does not have any information about the data lifecycle, it is fair to assume here that Acima stakeholders or products can have access to FoodApp user data. Legal analysis showed that this ethical issue of accessibility is addressed by the GDPR Art. 29 and 5 that require contractual agreement between data subject and controller or controller and processor.

The FoodApp scenario demonstrated the lack of a feedback loop from the app's stakeholders. The ordering and delivery processes are not reflected upon with the restaurants involved. The waste and traffic issues are not included into the recommender algorithms as well as are not monitored together with the local authorities. On the other hand, FoodApp's algorithm is optimized for re-ordering, hence for the increase of individual deliveries and waste associated with these deliveries. Hence, the values of social and environmental well-being and the associated ethical issues such as waste reduction, reduction of traffic and fuel consumption are not considered. Also, no legal consideration was identified that considers these issues in the digital consumer context. While FoodApp's business model heavily relies on the network of the delivery partners, the job assignment algorithm is not known for the delivery persons and the effect of the user reviews on the job assignments is not transparent neither for the users nor the delivery partners. These ethical issues of fairness and job accessibility are also not addressed by a legal requirement.

Autonomous decision making as well as the tendency for ethical consumption is endangered in the context of the FoodApp by the optimization goal of re-ordering and the given choice of restaurants and cuisines. The user is not enabled to make suggestions about restaurants or cuisines or forage for new options, engraving the skills and tastes within the app. The working conditions of the delivery persons are not visible or known to the user as well as the rewarding mechanism, e.g., tips, are not made available. In this process, the user is detached from the physical part of the service.

Combining legal and ethical analysis shows that some of the identified ethical issues are already covered by existing legislation. Nevertheless, bigger negative effects such as the effects on the environment or the society are part of the social awareness and responsibility that are not (and maybe should not be) regulated, but can be supported by socially acceptable IT artefacts.

As the legal norms can be implemented differently by those responsible for business and design decisions, as they do not provide specific processes for the implementation into the digital technologies. Ethical issues and views are more complex and disperse than legal norms and as such their implementation into the digital products can be less explicit. Therefore, we introduce and use the terms of socially aware IT, as such a system would consider and integrate the legal and ethical requirements into the design of the information system. The added effort could lead to a socially acceptable IT product. To ensure the remaining and homogeneous quality adherence, inter-company assessment mechanisms could be put in place.

Ethical issues that occur due to the use and implementation of digital products have been identified in research over the years. The FoodApp scenario analysis has demonstrated some of the complex effects that MLS can introduce as well as the resulting ethical issues. While some of these ethical issues were converted into legislation in some countries, e.g., within the European Union, others are actively debated in terms of the regulatory needs. ALTAI provides the set of ethical values that need to be considered when an MLS is being designed or used. While ethical values are not as binding as legal requirements, some of them are already incorporated into the legislation, i.e., the ones that are provided by the European Commission. The ethical analysis would uncover the ethical issues that might not yet have a legal equivalent. It is a matter of democratic discussion to decide whether and which ethical issues will need legal regulation and which ones can rely on the social contract.

9.7 Conclusion and Outlook

In this paper a combination of legal and ethical analyses was presented and applied to use case of a food recommender system, the FoodApp. This analysis approach showed aspects for ethical concern, such as decisional deskilling, emergence of structures of economic dependencies as well as additional effects on the environment induced by the individualization of the recommended service. The data-based approach chosen here presents an actionable radius for the system engineer and data analyst to include ethical and legal compliance during the design process of the ML-component by identifying operationalized ethical issues within software development.

Legal analysis of the FoodApp scenario showed that many of the ethical concerns are already addressed by the European GDPR legislation. However, targeting a broad area of data processing applications, the GDPR depends on interpretation, future jurisprudence or even new, more detailed legislation. Especially, the values of social and environmental well-being and even some of the aspects addressing human autonomy and oversight might need a legal fundament.

Although, the study showed that users of digital services may have expectations that are in part already covered by legal regulations but some of the identified ethical issues also rely on the ethical awareness of the company and thus go beyond legal compliance. While the ethical analysis revealed issues that go beyond the existing regulations on data protection, legal analysis showed the range for the interpretation of the legal regulation. Also, issues that are a matter of business ethics rather than legal regulation such as the awareness of the environmental impact and on the labor market originating from the broad usage of the digital platform were identified.

The combined analysis showed that MLS, specifically the examined food recommendation systems, have effects that do not only concern data processing, but that are rather beyond the direct interaction between the user and the system. While the GDPR addresses the data processing aspects such as user privacy and transparency, the effects of the usage of a food recommender system require an interdisciplinary discussion about the need of further regulation.

References

Abdelaziz, Y., D. Napoli, and S. Chiasson. 2019. End-Users and Service Providers: Trust and Distributed Responsibility for Account Security. In *Proceedings of the 2019 17th international conference on privacy, security and trust, PST 2019*, 1–6. Fredericton: IEEE eXpress Conference Publishing. https://doi.org/10.1109/PST47121.2019.8949041.

Aguiléra, A., L. Dablanc, and A. Rallet. 2018. L'envers et l'endroit Des Plateformes de Livraison Instantanée: Enquěte Sur Les Livreurs Micro-Entrepreneurs à Paris. *Réseaux* 6: 23–49. https://doi.org/10.3917/res.212.0023.

Coravos, A., I. Chen, A. Gordhandas, and A. D. Stern. 2019, February 14. We Should Treat Algorithms Like Prescription Drugs. *Quartz*. https://qz.com/1540594/treating-algorithms-like-prescription-drugs-could-reduce-ai-bias/.

De-Arteaga, M., R. Fogliato, and A. Chouldechova. 2020. A Case for Humans-in-the-Loop: Decisions in the Presence of Erroneous Algorithmic Scores. In *Proceedings of the 2020 CHI conference on human factors in computing systems*, 1–12. Honolulu: Association for Computing Machinery. https://doi.org/10.1145/3313831.3376638.

Doteveryone. 2019. *Survival of the fittest, piecemeal work and management by algorithm*. https://doteveryone.org.uk/2019/10/insights-gig-economy-research/.

European Commission. 2020. Assessment List for Trustworthy Artificial Intelligence (ALTAI) for Self-Assessment | Shaping Europe's Digital Future. https://ec.europa.eu/digital-single-market/en/news/assessment-list-trustworthy-artificial-intelligence-altai-self-assessment.

———. 2021. Proposal for a Directive of The European Parliament and of the Council on Improving Working Conditions in Platform Work. https://ec.europa.eu/eures/public/eu-proposes-directive-protect-rights-platform-workers-2022-03-17_en.

Floridi, L. 2016. Tolerant Paternalism: Pro-Ethical Design as a Resolution of the Dilemma of Toleration. *Science and Engineering Ethics* 22 (6): 1669–1688. https://doi.org/10.1007/s11948-015-9733-2.

Friedman, G. 2014. Workers without Employers: Shadow Corporations and the Rise of the Gig Economy. *Review of Keynesian Economics* 2 (2): 171–188. https://doi.org/10.4337/roke.2014.02.03.

Fritz, A., B. Wiebke, H. Gimpel, and S. Bayer. 2020. Moral Agency without Responsibility? Analysis of Three Ethical Models of Human-Computer Interaction in Times of Artificial Intelligence (AI). *De Ethica* 6 (1): 3–22. https://doi.org/10.3384/de-ethica.2001-8819.20613.

Germany. Datenethikkommission der Bundesregierung. 2018. Gutachten der Datenethikkommission. https://www.bundesregierung.de/breg-de/service/publikationen/gutachten-der-datenethikkommission-langfassung-1685238

Gille, F., A. Jobin, and M. Ienca. 2020. What We Talk About When We Talk About Trust: Theory of Trust for AI in Healthcare. *Intelligence-Based Medicine* 1 (November): 100001. https://doi.org/10.1016/j.ibmed.2020.100001.

Gray, C.M., Y. Kou, B. Battles, J. Hoggatt, and A.L. Toombs. 2018. The Dark (Patterns) Side of UX Design. In *Proceedings of the 2018 CHI conference on human factors in computing systems*, 1–14. Montreal: Association for Computing Machinery. https://doi.org/10.1145/3173574.

Ježová, D. 2020. Principle of Privacy by Design and Privacy by Default. *Regional Law Review* 127: 127–139. https://doi.org/10.18485/iup_rlr.2020.ch10.

Kamiran, F., and T. Calders. 2012. Data Preprocessing Techniques for Classification without Discrimination. *Knowledge and Information Systems* 33 (1): 1–33. https://doi.org/10.1007/s10115-011-0463-8.

Karpati, D., A. Najjar, and D. A. Ambrossio. 2020. Ethics of Food Recommender Applications. https://doi.org/10.1145/3375627.3375874.

Krügel, S., A. Ostermaier, and M. Uhl. 2021. Zombies in the Loop? People Are Insensitive to the Transparency of AI-Powered Moral Advisors. *Philosophy & Technology* 35 (1): 1–37.

Levina, O. 2019. Digital Platforms and Digital Inequality – An Analysis from Information Ethics Perspective. In *Proceedings of the Weizenbaum conference 2019 "Challenges of digital inequality – digital education, digital work, digital life*, 4. Berlin: Weizenbaum Institute for the Networked Society – The German Internet Institute. https://doi.org/10.34669/wi.cp/2.4.

———. 2020, March 08–11. A Research Commentary- Integrating Ethical Issues into the Data Process. In Paper presented at *15th international conference on Wirtschaftsinformatik.* https://www.researchgate.net/publication/339677955_A_Research_Commentary-Integrating_Ethical_Issues_into_the_Data_Process.

———. 2022. Implementing Ethical Issues into the Recommender Systems Design Using the Data Processing Pipeline. *Advances in Intelligent Systems and Computing* 14 (1): 153–163. https://www.researchgate.net/publication/358280889_Implementing_Ethical_Issues_into_the_Recommender_Systems_Design_Using_the_Data_Processing_Pipeline.

Milano, S., M. Taddeo, and L. Floridi. 2019. Recommender Systems and Their Ethical Challenges. *Minds and Machines* 2: 187–191. https://philpapers.org/archive/MILRSA-3.pdf.

OECD. 2016. Protecting Consumers in Peer Platform Markets: Exploring the Issues Background Report for Ministerial Panel 3.1. https://unctad.org/system/files/non-official-document/dtl-eWeek2017c05-oecd_en.pdf.

Paraschakis, D. 2016. Recommender Systems from an Industrial and Ethical Perspective. In *Proceedings of the 10th ACM conference on recommender systems – RecSys '16*, 463–466. Boston: Association for Computing Machinery. https://doi.org/10.1145/2959100.2959101.

———. 2017. Towards an Ethical Recommendation Framework. In *Proceedings of the international conference on research challenges in information science*, 211–220. Brighton: IEEE. https://doi.org/10.1109/RCIS.2017.7956539.

Polonioli, A. 2020. The Ethics of Scientific Recommender Systems. *Scientometrics* 126 (2): 1841–1848. https://doi.org/10.1007/s11192-020-03766-1.

Rosenblat, A., and L. Stark. 2016. Algorithmic Labor and Information Asymmetries: A Case Study of Uber's Drivers. *International Journal of Communication* 10: 3758–3784. https://doi.org/10.2139/ssrn.2686227.

Saner, E. 2020. Delivery Disaster: The Hidden Environmental Cost of Your Online Shopping. *The Guardian*. https://www.theguardian.com/news/shortcuts/2020/feb/17/hidden-costs-of-online-delivery-environment.

Schutt, R., and C. O'Neil. 2013. *Doing data science- straight talk from the frontline*. Sebastopol: O'Reilly Media.

Susser, D., and V. Grimaldi. 2021. Measuring Automated Influence: Between Empirical Evidence and Ethical Values. In *Proceedings of the AAAI/ACM conference on AI, ethics, and society*, 242–253. New York: Association for Computing Machinery.

Zhong, R., and C. Zhang. 2019. Food Delivery Apps Are Drowning China in Plastic. *The New York Times*. https://www.nytimes.com/2019/05/28/technology/china-food-delivery-trash.html.

Zhou, R., S. Khemmarat, and L. Gao. 2010. The Impact of YouTube Recommendation System on Video Views. In *Proceedings of the 10th ACM SIGCOMM conference on internet measurement*, 404–410. New Delhi: Association for Computing Machinery. https://doi.org/10.1145/1879141.1879193.

Chapter 10
Factors Influencing Trust and Use of Recommendation AI: A Case Study of Diet Improvement AI in Japan

Arisa Ema and Takashi Suyama

Abstract To use AI systems that are trustworthy, it is necessary to consider not only AI technologies, but also a model that takes into account factors such as guidelines, assurance through audits and standards and user interface design. In this paper, we conducted a questionnaire survey focusing on (1) AI intervention, (2) data management, and (3) purpose of use. The survey was conducted on a case study of an AI service for dietary habit improvement recommendations among Japanese people. The results suggest that how the form of communication between humans and AI is designed may affect whether users trust and use AI.

Keywords Trustworthy AI model · Data management · Health recommendation · HCI · AI ethics

10.1 Society 5.0 and Recommendation AI in Japan

In the wake of social issues where knowledge and information are not sufficiently shared in a society, Japan's Cabinet Office has proposed a vision of a future society called Society 5.0. As a follow-up to the hunting and gathering (Society 1.0), agricultural (Society 2.0), industrial (Society 3.0) and information societies (Society 4.0), Society 5.0 aims to develop the economy and solve social issues by integrating cyberspace and physical space (Cabinet Office 2016).

Society 5.0 was announced as part of Japan's Fifth Science and Technology Basic Plan in 2016 and continues to be referred to as the vision of the Japanese government and business entities. For example, in 2018, 2020, and 2022, Keidanren (Japan Business Federation) issued its 'Healthcare in the Age of Society 5.0' proposal, presenting a vision of a society that offers new options for diverse needs in health management and medical treatment (Keidanren 2022). It proposes that

A. Ema (✉) · T. Suyama
The University of Tokyo, Tokyo, Japan
e-mail: ema@ifi.u-tokyo.ac.jp

S. Genovesi et al. (eds.), *Recommender Systems: Legal and Ethical Issues*, The International Library of Ethics, Law and Technology 40,
https://doi.org/10.1007/978-3-031-34804-4_10

‘health management can be achieved by making appropriate recommendations suited to the situation at the time, without having to make a difficult decision’, and it is expected that artificial intelligence (AI) will be customized to recommend public actions.

While recommendation AI is expected to be used in many fields to realize Society 5.0, there are also many challenges. Some people may psychologically resist recommendations made by AI. It has been pointed out that that especially in Japan, elderly respondents showed negative attitudes towards recommendation AI services (Ikkatai et al. 2022). Data management is also a challenge for AI to provide information customized to individuals. For example, there was a controversy in Japan surrounding the use of data of job-seeking students by companies without the students’ consent in 2019 (Kudo et al. 2020). Whether data is appropriately managed or used for purposes for which it was not intended can influence people’s decisions to use AI services. In this paper, we present the results of a survey that examined the perspectives of users in deciding whether they would like to use the recommendation AI services.

10.2 Model for Ensuring Trustworthiness of AI Services

Ensuring trustworthiness of AI services has been an important topic in recent years. In Europe, a report on Trustworthy AI was released by the European Commission in March 2019. Simultaneously, Japan’s Cabinet Office released ‘Social Principles of Human-Centric AI’, which include building a mechanism to ensure the trustworthiness of AI and the data or algorithms that support it (Cabinet Secretariat 2019).

Incorporating various AI principles into practical processes has been gaining increasing attention in recent years (Jobin et al. 2019). Trusted AI is achieved not only through technology but also through an ‘AI governance ecosystem’ that includes various elements such as guidelines, auditing, standardization, user interface design, literacy and education (Japan Deep Learning Association 2021), and various models for ensuring trustworthiness have been proposed.

The DaRe4TAI framework by Thiebes et al. examines whether privacy protection and discriminatory judgements occur at all stages of data input and output and AI model building (Thiebes et al. 2020). In Japan, several frameworks have been published to promote trusted AI in society. For example, the Ministry of Economy, Trade and Industry (METI) released the ‘Governance Guidelines for Implementation of AI Principles, ver.1.1’ in 2022, which presents action goals for AI providers to implement (METI 2022). An increasing number of companies are establishing their own AI principles and guidelines; Fujitsu Limited, for example, has developed a method to disseminate trustworthy AI to society and published its procedures and application examples (Fujitsu Limited 2022). Research is also being conducted at universities, and the Risk Chain Model (RCModel) is structured into three layers: AI model (e.g., accuracy and robustness), system (e.g., data and system infrastructure), service provider (e.g., behavioral norms and communication) and user (e.g., understanding, utilization and usage environment) (Matsumoto and Ema 2020). The framework guide and case studies[1] are available for considering ways to ensure the reliability and transparency of AI services by structuring them into these three layers.

By establishing such ethics policies and guidelines, AI service providers are trying to ensure social trust, and publicize that they are developing and using technology in an appropriate manner. However, having policies and guidelines is not the same as being a truly trustworthy organization. The issue of 'ethics washing' has been pointed out: an organization may claim to behave ethically, but it may be just a façade or a 'smokescreen to hide their transgressions' (Dand 2021). Therefore, while it is important to have these principles in place, it is also important to communicate appropriately with users.

10.3 Components of a Trustworthy AI Model

While there are several models and frameworks for ensuring trustworthiness of AI services, this paper proposes a model that places particular emphasis on the interface between the service provider and the user. In particular, we investigate not from the perspective of what technical requirements are necessary for AI technology to be trustworthy, but rather from the hypothesis that how the form of communication between humans and AI is designed may affect whether users trust and use it. For example, even if the AI services offered by companies are the same, user preferences will vary depending on how purpose of use is explained and how much freedom of choice users are given.

There are many studies on AI-human interface design, especially on the general bases for human trust in automated machines such as AI (Madhavan and Wiegmann 2007). Research points to the performance of the machine (e.g., abilities), the process of how the machine works, and whether its design achieves the designer's purposes as the bases for trust (Lee and See 2004). To gain the trust of users as well as designers, it is important that AI is in accordance with their preferences. Trust in machines is not only related to the machine, but also to the governance of the company, such as whether data is appropriately managed. Therefore, we conducted a survey based on the hypothesis that three aspects of interaction are important: (1) AI intervention, (2) data management and (3) purpose of use. Although a variety of AI services are currently available, this paper specifically investigates recommendation AI as a case study.

[1] Case studies are published at The University of Tokyo website: https://ifi.u-tokyo.ac.jp/en/projects/ai-service-and-risk-coordination/

10.3.1 AI Intervention

Service providers need to understand users' preferences for AI services. For example, some users may not want AI to make decisions and prefer human judgement, while others may prefer to use AI services only as a reference but only if a human makes the final decision. Conversely, others may prefer to have AI alone make decisions without any human intervention.

Thus, the degree of AI intervention varies widely, and it may be difficult to tell the difference between a pattern in which decisions are made by humans alone and a pattern in which AI is used but final decisions are made by humans unless the process is clearly explained. Therefore, from the viewpoint of transparency, service providers are expected to log the process by which decisions are made and to establish organizational governance to provide AI services. Users are also expected to use AI based on information and explanations provided by developers and service providers in some cases, as considered in the principles of proper utilization in the 'AI Utilization Guidelines' proposed by the Japanese government (Ministry of Internal Affairs and Communications 2019).

10.3.2 Data Management

To ensure the reliability of AI services, it is imperative for users and service providers to discuss whether data management is appropriate. Some users may ask service providers to explain the security measures they have employed while learning how they manage data, while others may trust the company and refrain from checking the security measures.

Therefore, service providers are expected to indicate what data management they adopt. While there are various certifications and standards for data management, some companies have adopted their own standards, and management methods are diversifying.

10.3.3 Purpose of Use

When users are deciding whether or not to use an AI service, it is important for them to know what information will be taken from the service and for what purpose it will be used. The general means of communicating the purpose of use is the terms of service agreements.

To avoid too-long and too-complex terms of service agreements that go unread (Maronick 2014), service providers should provide agreements that are easy to understand and of appropriate length. Currently, there are formats allowing users to choose what they agree or disagree with, rather than just agreeing or disagreeing with all of them at once. In addition, some terms are designed so that there are no negative consequences, such as services being provided even if users do not agree with all of them. The way the terms of use are written and presented is also becoming more diverse.

10.4 Verification of Trustworthy AI Model: A Case Study of AI for Dietary Habit Improvement Recommendations

There are various types of recommendation AI services, but this paper takes recommendation AI for dietary habit improvement utilizing users' dietary data as the case study. As Japan's Keidanren (Japan Business Federation) has proposed a vision of what health management should be in Society 5.0, there is a high affinity and need for AI and healthcare services. Applications in which AI provides menu and nutrition management support and interactive AI that analyses meal menus in real time and guides people towards appropriate eating habits are available in Japan.

A survey was conducted from January to March 2021 to examine what kind of AI services, data management methods and purpose of use would make users willing to use AI for dietary habit improvement recommendations.

10.4.1 Subjects

A research firm was commissioned to select respondents to ensure that Japanese men and women and their ages (20–60s) were equally represented, and in-depth interviews were first conducted with nine of them. For the subsequent survey, the same group of respondents was commissioned to a research firm, and 500 respondents were included. Since the purpose of the survey was to target general users, those who responded in the pre-survey that they were 'familiar with AI' and 'not at all reluctant to use AI services' were excluded from the survey.

10.4.2 Verification 1: AI Intervention

In-depth interviews were conducted to determine the degree of AI intervention in recommending services. It was found that respondents tend to prefer services in which 'AI can be used, but the final recommendation is made by a human' rather than 'AI making the recommendation' for dietary improvement advice. Two trends were obtained as reasons for this: AI performance is considered to be not as good as that of humans, and the ability to ask questions and have conversations is considered important in healthcare.

Therefore, four patterns were displayed and tested in the quantitative survey as demonstrated in Table 10.1: a dietitian with AI (No. 1) and AI (No. 2); a dietitian with AI (No. 1) and a high-performance AI that makes suggestions equivalent to those of a dietitian (No. 3); a dietitian with no questions or conversations (No. 4) and AI (No. 2); and a dietitian with no questions or conversations (No. 4) and a high-performance AI (No. 3). Respondents were asked to indicate which of these four patterns of food improvement advice they would prefer to use.

Table 10.1 AI intervention descriptions

No.	Types of AI services	Descriptions
1	A dietitian (human) with AI	**Our professional dietitian will suggest the best diet improvement methods for you.** Eating habits are analyzed by AI, which learns the relationship between eating habits and health based on a vast amount of data. Based on data calculated from nutrients, calorie intake, age, height and weight, a dietitian will guide you to a suitable diet, and you can ask the dietitian questions.
2	AI	**AI will suggest the best diet improvement methods for you.** The AI learns the relationship between diet and health based on a vast amount of data and will guide you to a diet that suits your age, height and weight based on nutrients and client calories calculated from your diet (we cannot accept questions about how the AI analysis works).
3	A high-performance AI that makes suggestions equivalent to those of a dietitian	**Cutting-edge AI suggests the best diet improvement methods for you.** (*Assume that all information on your diet, pre-existing conditions will be converted into data if necessary, and that you will receive the same suggestions as a dietitian) The AI learns the relationship between diet and health based on a vast amount of data and will guide you to a diet suitable for your age, height and weight based on nutrients and client calories calculated from your diet (we cannot accept questions about how the AI analysis works).
4	A dietitian (human) with AI with no questions or conversations	**Our professional dietitian will suggest the best diet improvement methods for you.** (*Assume that you cannot ask the dietitian any questions) Eating habits are analyzed by AI, which learns the relationship between eating habits and health based on a vast amount of data. Based on data calculated from nutrients, calorie intake, age, height and weight, a dietitian will guide you to a suitable diet.

10.4.3 Verification 2: Data Management

When we conducted an in-depth interview to determine how users evaluate whether data management is properly implemented, we received several comments indicating that there is an emphasis on whether some kind of certification is obtained rather than a detailed explanation of what kind of technology is used to ensure security measures.

Therefore, we displayed three patterns of questions in the survey, as Table 10.2 shows: a comparison between ISO[2] and a company's own standards as a representative example of certification, and an explanation of those certifications and specific security technologies to verify which service users would prefer.

10.4.4 Verification 3: Purpose of Use

In-depth interviews were conducted to investigate what type of explanation of the purpose of use is preferred in the terms of service agreements for AI services and how best to obtain consent. From the interviews, 'the items for obtaining consent are explained in detail' and 'users can customize what they agree or do not agree to' tended to be preferred.

Therefore, in the survey, we created terms of service agreements (Table 10.3) and three patterns of questions were displayed: users agree collectively by checking one

Table 10.2 Data management option descriptions

Types of data management	Descriptions
ISO	**Security that has passed strict audits** (*ISO is an international standard) ISO-certified cloud security Anonymization using ISO-certified encryption technology
In-house standards	**Security that has passed strict audits** (*Assume a well-known major Japanese company) Highest rated cloud security (in-house standards) Anonymization using in-house proprietary encryption technology
Specific security technologies	**Cutting-edge security technology** Cloud security using XX method Anonymization using YY encryption technology

[2] ISO is a private organization headquartered in Switzerland and stands for International Organization for Standardization. Since differences in product size, quality, safety, and functionality from country to country can hinder international trade, the ISO is intended to create standards. Currently, ISO/IEC JTC1 is the forum for discussing international de jure standards for AI. There are several working groups and trustworthiness of AI is also discussed.

Table 10.3 Description of purpose of use in terms of service agreements

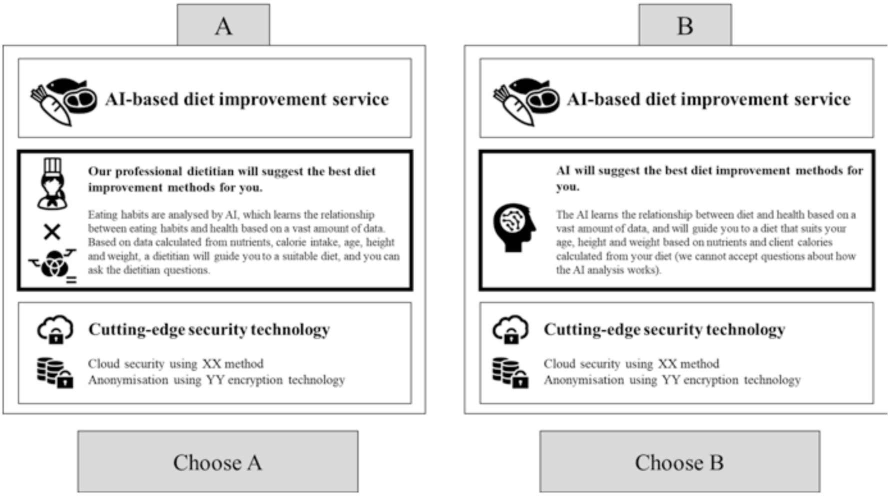
Agreement regarding personal data to be collected
1. Data to be collected
Name, address, telephone number, e-mail address, nickname, gender, date of birth
Dietary information such as the content, portion size and time of meals eaten for breakfast, lunch, dinner and snacks
Physical information such as height, weight, body fat percentage
Exercise information such as exercise time, exercise content and exercise habits
Information on the terminal used and usage logs
2. Purpose of use
(1) Operation of services
To accept registrations related to the service, to verify the user's identity and to provide and maintain the service
To respond to inquiries regarding the service
(2) To provide the service to users
To provide information on advice automatically generated by the program based on physical and dietary exercise information
To provide counselling services such as dietary guidance based on physical and dietary exercise information
(3) To provide information related to the service
To send information regarding the service or related events
To request surveys regarding the service or related events
(4) Marketing
To use the information for marketing activities by the company in a manner that does not allow identification of specific individuals
To use the information to improve the delivery of advertisements, after processing in such a way that specific individuals cannot be identified
(5) For research use
To be used by our subcontractors for research purposes after processing in such a way that specific individuals cannot be identified
To be used by our business partners for research purposes after processing in which specific individuals cannot be identified

item; users check five items per category, and users check all (15 items), to verify whether users' preferences change with the number of checked items. In the cases of checking five and 15 items, we also created a pattern with optional check items and compared it with a pattern in which check items are required.

10.4.5 Method

The quantitative survey format was created using the maze web tool,[3] and the two patterns were displayed simultaneously on the left and right sides of the screen, with the different patterns highlighted. The respondents were asked to choose which service they would prefer to use (Figs. 10.1 and 10.2).

[3] maze, https://maze.co/



A

AI-based diet improvement service

Our professional dietitian will suggest the best diet improvement methods for you.

Eating habits are analysed by AI, which learns the relationship between eating habits and health based on a vast amount of data. Based on data calculated from nutrients, calorie intake, age, height and weight, a dietitian will guide you to a suitable diet, and you can ask the dietitian questions.

Cutting-edge security technology

Cloud security using XX method
Anonymisation using YY encryption technology

B

AI-based diet improvement service

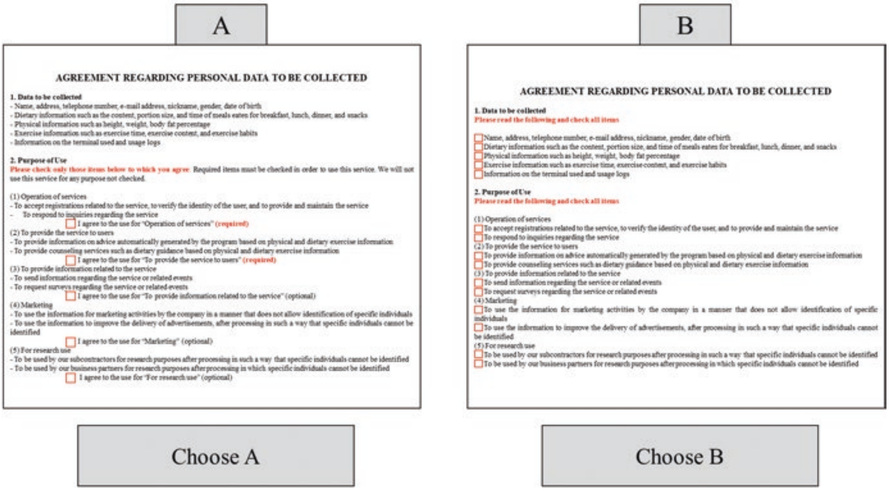
AI will suggest the best diet improvement methods for you.

The AI learns the relationship between diet and health based on a vast amount of data, and will guide you to a diet that suits your age, height and weight based on nutrients and client calories calculated from your diet (we cannot accept questions about how the AI analysis works).

Cutting-edge security technology

Cloud security using XX method
Anonymisation using YY encryption technology

Fig. 10.1 Example of screens with different 'AI intervention'



A

AGREEMENT REGARDING PERSONAL DATA TO BE COLLECTED

1. Data to be collected
- Name, address, telephone number, e-mail address, nickname, gender, date of birth
- Dietary information such as the content, portion size, and time of meals eaten for breakfast, lunch, dinner, and snacks
- Physical information such as height, weight, body fat percentage
- Exercise information such as exercise time, exercise content, and exercise habits
- Information on the terminal used and usage logs

2. Purpose of Use
Please check only those items below to which you agree. Required items must be checked in order to use this service. We will not use this service for any purpose not checked.

(1) Operation of services
- To accept registrations related to the service, to verify the identity of the user, and to provide and maintain the service
- To respond to inquiries regarding the service
□ I agree to the use for "Operation of services" **(required)**
(2) To provide the service to users
- To provide information on advice automatically generated by the program based on physical and dietary exercise information
- To provide counseling services such as dietary guidance based on physical and dietary exercise information
□ I agree to the use for "To provide the service to users" **(required)**
(3) To provide information related to the service
- To send information regarding the service or related events
- To request surveys regarding the service or related events
□ I agree to the use for "To provide information related to the service" (optional)
(4) Marketing
- To use the information for marketing activities by the company in a manner that does not allow identification of specific individuals
- To use the information to improve the delivery of advertisements, after processing in such a way that specific individuals cannot be identified
□ I agree to the use for "Marketing" (optional)
(5) For research use
- To be used by our subcontractors for research purposes after processing in such a way that specific individuals cannot be identified
- To be used by our business partners for research purposes after processing in which specific individuals cannot be identified
□ I agree to the use for "For research use" (optional)

B

AGREEMENT REGARDING PERSONAL DATA TO BE COLLECTED

1. Data to be collected
Please read the following and check all items

□ Name, address, telephone number, e-mail address, nickname, gender, date of birth
□ Dietary information such as the content, portion size, and time of meals eaten for breakfast, lunch, dinner, and snacks
□ Physical information such as height, weight, body fat percentage
□ Exercise information such as exercise time, exercise content, and exercise habits
□ Information on the terminal used and usage logs

2. Purpose of Use
Please read the following and check all items

(1) Operation of services
□ To accept registrations related to the service, to verify the identity of the user, and to provide and maintain the service
□ To respond to inquiries regarding the service
(2) To provide the service to users
□ To provide information on advice automatically generated by the program based on physical and dietary exercise information
□ To provide counseling services such as dietary guidance based on physical and dietary exercise information
(3) To provide information related to the service
□ To send information regarding the service or related events
□ To request surveys regarding the service or related events
(4) Marketing
□ To use the information for marketing activities by the company in a manner that does not allow identification of specific individuals
□ To use the information to improve the delivery of advertisements, after processing in such a way that specific individuals cannot be identified
(5) For research use
□ To be used by our subcontractors for research purposes after processing in such a way that specific individuals cannot be identified
□ To be used by our business partners for research purposes after processing in which specific individuals cannot be identified

Fig. 10.2 Example of screens with different 'terms of service agreements'. (A has five check items with customization, B has 15 check items with no customization)

10.4.6 Results

10.4.6.1 AI Intervention

Among the recommended services, Fig. 10.3 shows the results of user preferences for the pattern in which a person or AI ultimately recommends the service, the pattern in which AI is as high performing as a person, and conversely, the pattern in

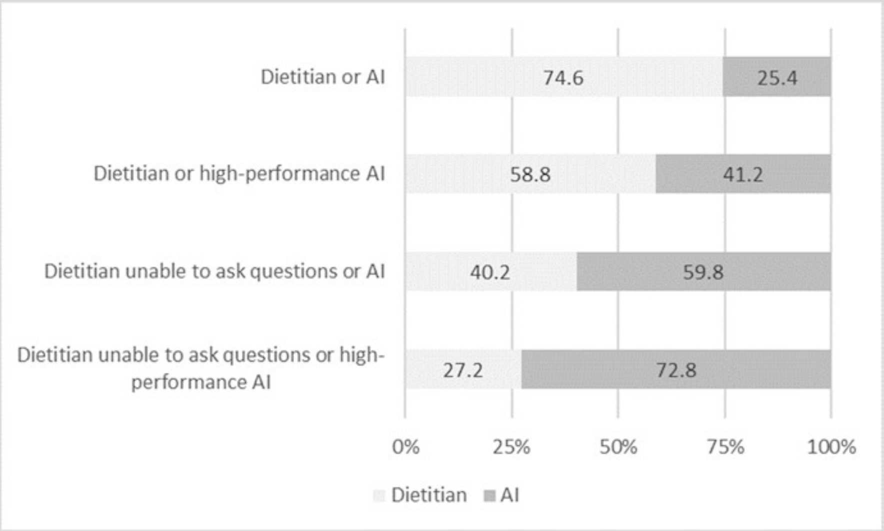


Fig. 10.3 Preference for AI intervention (n = 500)

which a user cannot talk to a dietitian (human). The figure shows that the strongest preference is for services that are ultimately recommended by a dietitian (human) using AI. In addition, when a dietitian (human) is compared with a high-performance AI that can make judgements similar to a human, the human recommendation nevertheless prevails. Conversely, the ratio of those who prefer to use the AI recommendation increases, reversing the result for the recommendation by a dietitian (human) who cannot ask questions. These results suggest that users emphasize the interactivity of the recommendation service, such as the ability to have a conversation, rather than the high-performance AI. This tendency was more pronounced among the group that responded that they are 'resistant to AI services' (Fig. 10.4).

10.4.6.2 Data Management

What do users tend to look for in a recommendation AI to determine if the data is being appropriately managed? Figure 10.5 shows the results of user preference for patterns in which ISO or a company's in-house standards are clearly stated and in which there is an explanation of specific security technology. The figure suggests that some kind of certification, such as ISO or acquisition of a company's own standards, is more important for gaining user trust than an explanation of specific security technologies.

Compared to in-house standards, users tend to prefer ISO. This suggests that socially recognized certification is more reliable. In this survey, it was set as 'major well-known Japanese companies' that established their own in-house standards. Contrary to the assumption that the standard set by a major well-known company

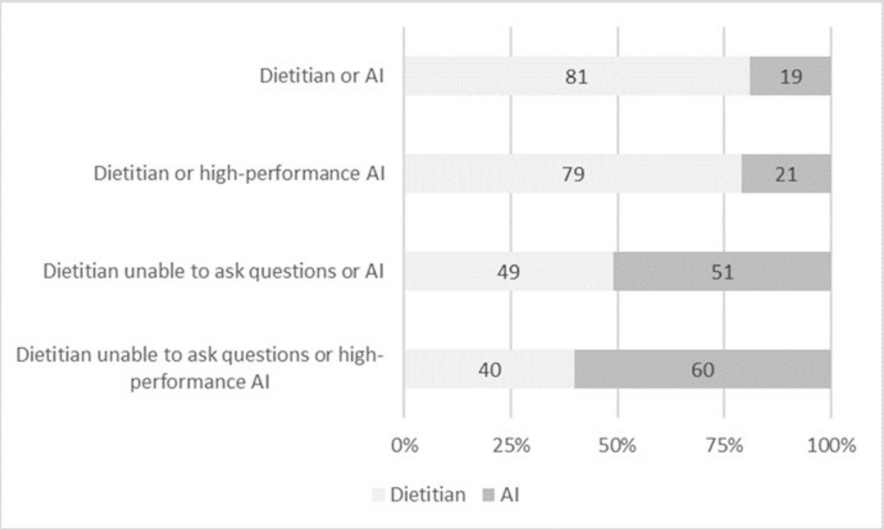


Fig. 10.4 AI service preference of those with AI service resistance (n = 500)

would be considered more reliable, users preferred ISO, suggesting that it is less significant to display one's own standards, regardless of company size or name recognition. Contrarily, the results of the study suggest that companies would be more likely to earn socially recognized certification such as ISO, which would contribute to user preference.

10.4.6.3 Purpose of Use in Terms of Service Agreements

Which consent formats are preferred by users? Figure 10.6 shows the results of user preferences, which are classified into patterns based on the number of consent check items and customizability. The figure shows that users prefer the one with a higher number of consent check items. This suggests that a large number of items is preferred, even though it is troublesome to read through them.

Meanwhile, users preferred the customizable pattern with 'optional' check items in addition to the 'required' ones, regardless of the consent items. This suggests that users prefer patterns that allow them to choose the contents of consent by themselves. However, in this survey, users do not actually click the mouse to check the item. The results of this survey should be verified further, as the cumbersomeness of the operation may prevail if users are actually required to click to check the boxes.

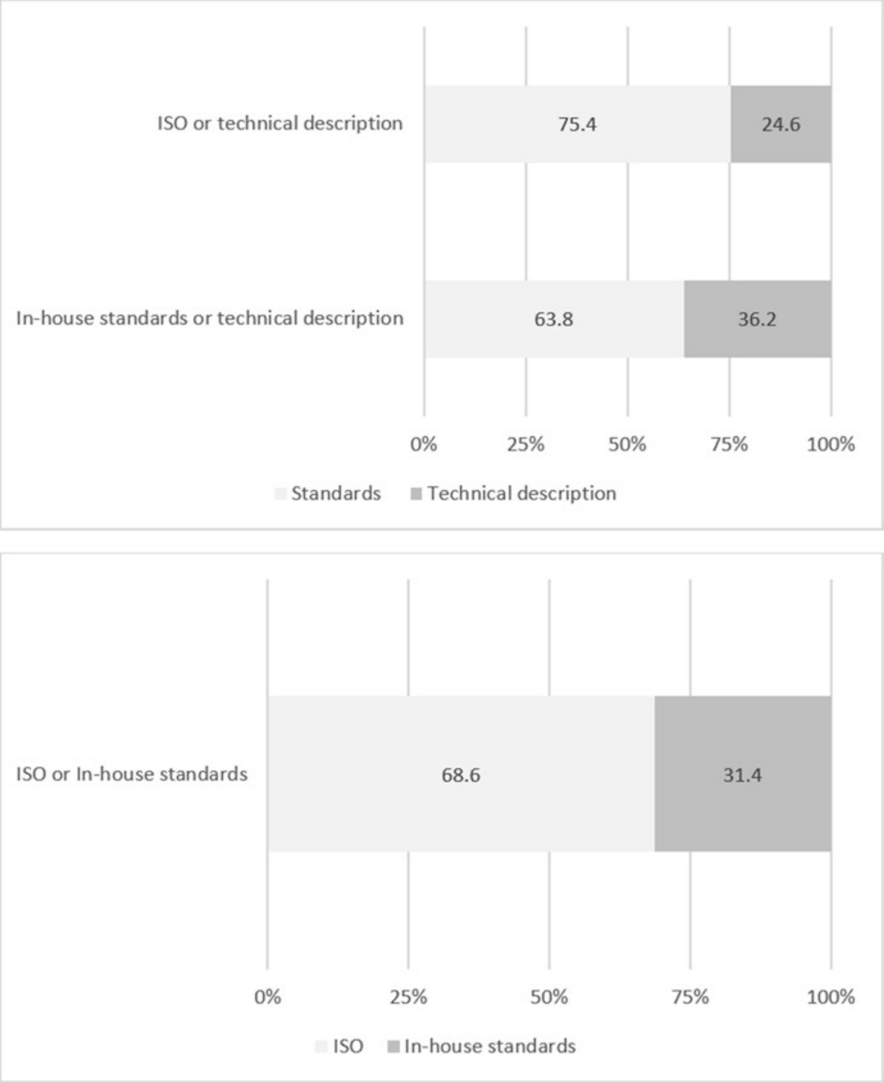


Fig. 10.5 Preferences of data management (n = 500)

10.5 Necessary Elements for Trusted AI

Technical developments such as explainable and fair AI are considered important for trusted AI services. This paper focuses on communication between users and service providers among non-technical aspects. To facilitate communication, we investigated what kind of AI intervention, data management and purpose of use would be considered acceptable by users for dietary habit improvement recommendation AI.

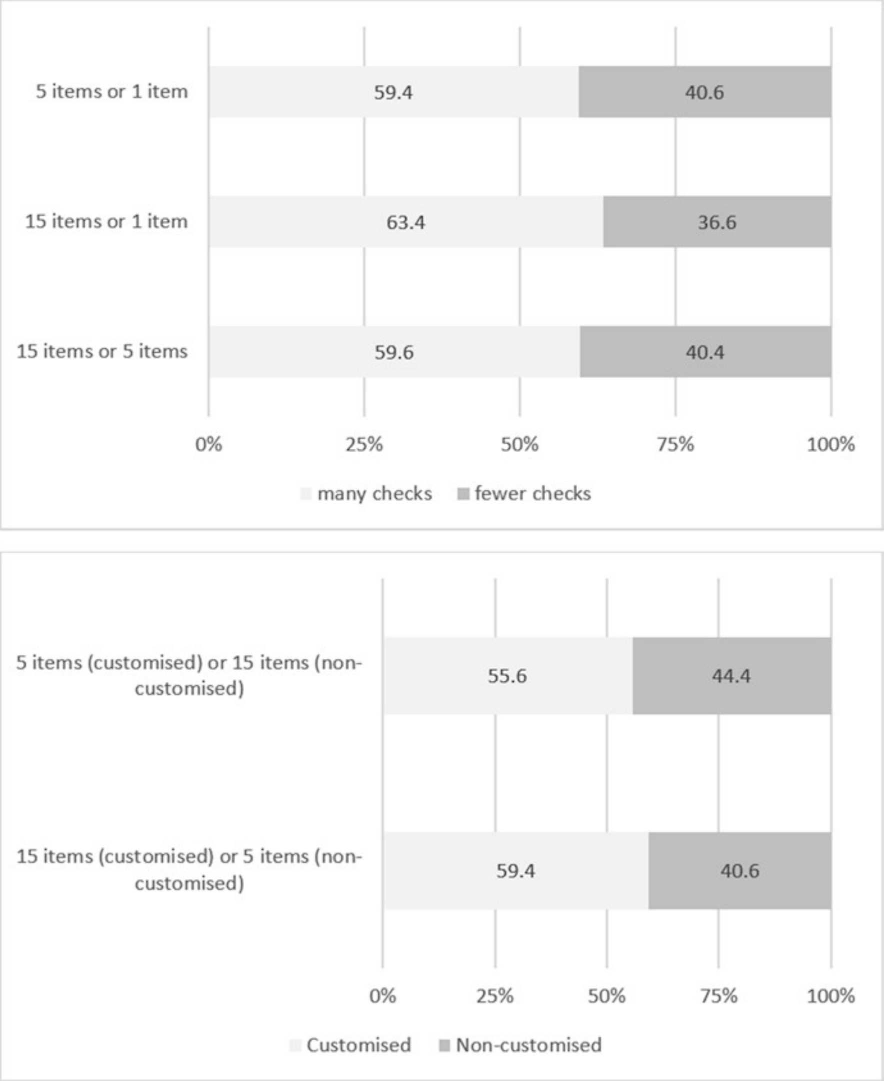


Fig. 10.6 Preferences of terms of use (n = 500)

Reliability of services in general depends on users understanding an accurate description of how they work. However, AI is subject to changes in the environment in which it is used and in the algorithms that result from the learning process. Considering that predictability of behavior is difficult not only for users but also for service developers, transparency and accountability are important for users to trust AI services. Specifically, communication is required so that there is no discrepancy in understanding between users and service providers regarding AI intervention and data management.

The results of this survey suggest that people tend to trust recommendation AI that is communicative, allowing users to ask questions and engage in dialogue when it is recommended. For those who are resistant to AI, it is also expected that rather than promoting the technical perspective of 'being a high-performance AI', it would be more acceptable to develop a service that allows users to talk and ask questions to people while using AI in the background for analysis. Advertising high-performance AI that can provide advice equivalent to that of a human being or providing detailed information about the technology used for data management is not likely to have a significant effect on user preference.

In addition, the pattern of ISO for data management was highly rated. However, when the respondents who preferred the pattern with ISO in the in-depth interview were asked if they could explain what ISO is, few had extensive knowledge except that it is a standard of some type. Furthermore, the results showed that the more detailed the items for obtaining agreement to the terms of service agreements, the better, but other studies show that people don't read terms of agreements (F-secure 2014; Japan Fair Trade Commission 2021). Therefore, this suggests that the detailed content itself is a sign of trust. These results also indicate that non-technical aspects, such as socially formed understandings, have a greater influence on user preferences than the actual existence of technical guarantees and the fact that they are explained. This study suggests that service providers should not ignore these non-technical perspectives when it comes to trusted AI.

This is a case study of AI recommendation for dietary improvement services. However, AI recommendation services are employed in various fields, and not all recommendation AIs are similar to those in this survey. It can be assumed that health management and medical-related fields are areas where interaction with people is particularly important. For example, users may prefer recommendation AI rather than people for complex route guidance because AI can better perform complex processing. Therefore, an international trustworthiness survey focusing on AI intervention, data management, and purpose of use in various service cases is needed to consider the requirements for trustworthy AI.

Acknowledgements This paper was supported by the Toyota Foundation D18-ST-0008 'Building platform on ethics and governance of artificial intelligence'.

References

Cabinet Office. 2016. Society 5.0. https://www8.cao.go.jp/cstp/english/society5_0/index.html. Accessed 2 May 2022.

Cabinet Secretariat. 2019. Social principles of human-centric AI. https://www.cas.go.jp/jp/seisaku/jinkouchinou/pdf/humancentricai.pdf. Accessed 2 May 2022.

Dand, Mia. 2021. The diversity and ethics crisis in AI. AI principles to practice: An introduction to international activities. https://ifi.u-tokyo.ac.jp/wp/wp-content/uploads/2021/04/JSAI_report_210427.pdf. Accessed 24 Aug 2022.

F-secure. 2014. Tainted love: How Wi-Fi betrays us. https://www.studocu.com/en-gb/document/manchester-metropolitan-university/professional-development/wi-fi-experiment-uk-2014/1115305. Accessed 2 May 2022.

Fujitsu Limited. 2022. Fujitsu delivers a new resource toolkit to offer guidance on ethical impact of AI systems. https://www.fujitsu.com/global/about/resources/news/press-releases/2022/0221-01.html. Accessed 2 May 2022.

Ikkatai, Yuko, Tilman Hartwig, Naohiko Takanashi, and Hiromi M. Yokoyama. 2022. Octagon Measurement: Public Attitudes toward AI Ethics. *International Journal of Human-Computer Interaction*. https://doi.org/10.1080/10447318.2021.2009669.

Japan Deep Learning Association. 2021. Report "AI governance ecosystem – Trusted AI with industrial structure." https://www.jdla.org/en/en-about/en-studygroup/en-sg01/. Accessed 21 Aug 2022.

Japan Fair Trade Commission. 2021. Final report regarding digital advertising. https://www.jftc.go.jp/en/pressreleases/yearly-2021/February/210217.html. Accessed 2 May 2022.

Jobin, Anna, Marcello Ienca, and Effy Vayena. 2019. The Global Landscape of AI Ethics Guidelines. *Nature Machine Intelligence* 1: 389–399.

Keidanren. 2022. Healthcare in society 5.0 age. https://www.keidanren.or.jp/policy/2022/005.html (in Japanese). Accessed 2 May 2022.

Kudo, Fumiko, Hiromi Arai, and Arisa Ema. 2020. Ethical Issues Regarding the Use of AI profiling Services for Recruiting: The Japanese Rikunabi Data Scandal. Arxiv. https://arxiv.org/abs/2005.08663

Lee, John D., and Katrina A. See. 2004. Trust in Automation: Designing for Appropriate Reliance. *Human Factors* 46 (1): 50–80.

Madhavan, Poornima, and Doublas A. Wiegmann. 2007. Similarities and Differences Between Human–Human and Human–Automation Trust: An Integrative Review. Theoretical Issues in Ergonomics Science 8(4), 277–301

Maronick, Thomas J. 2014. Do consumers read terms of service agreements when installing software? – A two-study empirical analysis. *International Journal of Business and Social Research, MIR Center for Socio-Economic Research* 4 (6): 137–145.

Matsumoto, Takashi, and Arisa Ema, 2020. RCModel, a Risk Chain Model for Risk Reduction in AI Services. Arxiv. https://doi.org/10.48550/arXiv.2007.03215

Ministry of Economy, Trade and Industry. 2022. Governance guidelines for implementation of AI principles. https://www.meti.go.jp/shingikai/mono_info_service/ai_shakai_jisso/20220128_report.html. Accessed 2 May 2022.

Ministry of Internal Affairs and Communications. 2019. AI utilization guidelines. https://www.soumu.go.jp/main_content/000658284.pdf. Accessed 24 Aug 2022.

Thiebes, Scott, Sebastien Lins, and Ali Sunyaev. 2020. Trustworthy artificial intelligence. *Electronic Markets* 31: 447–464.

Chapter 11
Ethics of E-Learning Recommender Systems: Epistemic Positioning and Ideological Orientation

Lisa Roux and Thierry Nodenot

Abstract Recommender systems are increasingly used in e-learning to provide users with personalized services and advice. Depending on the specific context for which the system is implemented (e.g., homework on a specific subject for university students, new training courses for life-long learners), the objectives and proposed items, the chosen recommendation techniques, the features that are considered, the way the recommendations are presented to the users are closely related to the designers' perception of learners and knowledge. The various approaches reflect different epistemic and ethical viewpoints; for example, representing people using fixed models is easier to process, diagnose, predict and explain, but presents a partial view of reality and obscures the fact that they are complex and evolving individuals. Similarly, some filtering methods can restrict the view of available courses to items considered similar to those that the learner has already followed, thus promoting specialization rather than diversification and openness. This aspect is closely related to fundamental issues involved in the theory of knowledge, questioning the notions of utility and purposes of science, as well as a key issue for academic change and, more fundamentally, that of modern societies. Indeed, these issues should be seen in a broader context of reflection about the economic changes and ideological transformations of a society grounded on neoliberal capitalism. The main goal of this study is to explain how the design of recommender systems in e-learning has both ethical and practical implications since it reflects an ideological conception of science and techniques, thus requiring a previous examination of these issues in order to define the theoretical model of knowledge in which it takes place. For that purpose, we study the certain visions of teaching and learning that can be brought about by algorithms and models used by existing recommender systems in e-learning.

Keywords Recommender system · e-learning · Ethics · Implications of design choices

L. Roux (✉) · T. Nodenot
Université de Pau et des Pays de l'Adour, E2S UPPA, LIUPPA, Anglet, France
e-mail: thierry.nodenot@iutbayonne.univ-pau.fr

S. Genovesi et al. (eds.), *Recommender Systems: Legal and Ethical Issues*, The International Library of Ethics, Law and Technology 40,
https://doi.org/10.1007/978-3-031-34804-4_11

11.1 Introduction

Recommender systems are widely used in e-commerce, social networks, Video On Demand, etc. They are based on collected information about user preferences, which can be acquired implicitly (e.g., collecting the data describing the users' behavior) or explicitly (e.g., collecting the users' ratings, using social and demographic information). They aim at guiding users within the wide range of products offered by e-platforms so that they can find the items that they are most likely to engage with. Similarly, recommender systems for e-learning are used to help learners to deal with the abundant learning resources and activities available on e-learning platforms. They are used to support individual learning and provide the learner with the appearance of personalized tutoring in environments of large classes but limited human contact (i.e., reduced teacher to student ratio). In the context of the ongoing COVID-19 pandemic, the use of such recommender systems appeared crucial for the student's involvement and engagement, and to help teachers optimize the assistance they provide. They are also expected to efficiently help students to choose their learning paths, efficient pedagogical activities, and suitable course material. Indeed, many papers present good results concerning the effectiveness of their recommendation approach, whether it be based on task difficulty ranking (Segal et al. 2014), learning material or path recommendation (Mbipom et al. 2018; Ye et al. 2015; Shi et al. 2020), friend recommendation (Rafaeli et al. 2005), learning object recommendation (Gallego et al. 2012; Fraihat and Shambour 2015), or performance evaluation (Amasha et al. 2020; Moore et al. 2019). Furthermore, from the institution's perspective, they can be used to increase their economic benefits, since they enable larger volumes of students with fewer teachers.

However, if lucrative economic interests come first, this may raise a number of ethical concerns, both theoretically and practically (Beach and Dovemark 2009). Recommender systems are widely used for e-commerce and entertainment (i.e. music, video on demand (VOD), video game platforms, etc.). In early 2000, the first notable applications appeared in the domain of education (Manouselis et al. 2012). The development of the Internet has stimulated the research in recommender systems in order to improve the filtering and assessment methods, in particular through the involvement of economic actors who encourage them to increase their benefits (e.g. the Netflix prize that challenged machine learning and data mining). Such practices not only promote algorithmic development but also affect our relationship to education, culture, etc. (Hallinan and Striphas 2016).

First of all, as Catherine Hayles (1999) shows, new technological contexts (thus by inclusion new algorithmic contexts) change our relationship to our environment and, more fundamentally, alter our meaning of humanity, culture, sociality, etc. Moreover, the methods developed for VOD, e-commerce, e-learning etc. are the same; thus, the methods used to sell products and increase the benefits of an enterprise are also used to recommend learning items to students, which may cause problems. Indeed, recommendations in e-commerce can be accused of being more meant to help sellers than customers, just as advertising is. In e-commerce, the sellers use recommender systems to increase the number of sales and their income, sometimes with very little concern about what their customers really need. From this standpoint, they can be considered as the real customers of recommender systems, instead of the sellers' customers. Some studies have shown that recommender systems in e-commerce not only help the sellers' customers to get what they prefer: they also contribute to shaping their preferences (Adomavicius et al. 2019). Yet, even though this very practice is already highly questionable in e-commerce, it is even more critical in the case of e-learning. In particular, recommender systems in e-learning cannot pretend to stand apart from the academic debates about university, notably concerning its aims and values. For example, should university train future workers with competencies consistent to the job market, or whole citizens who will be able to take an active part in the social and political life of the society? (Florian 2018; Brighouse and McPherson 2015) Is university threatened by a commodification of knowledge process? (Jacob 2003).

As part of the reflection about education trends and policies, and besides all its potential benefits, the development of recommender systems in e-learning gives rise to ethical and social issues, including privacy problems, lack of control, etc. Just as the other personalized e-learning functions, the recommender systems use personal information, generally using algorithms poorly understood by the users, in order to offer them services. Consequently, the opacity of such systems raises a main issue because the users are recommended items without being provided with the knowledge to understand these recommendations, thus knowing when to choose to follow them or not. In order to understand why it can be problematic, we will explain how recommender systems can promote particular visions about education, pointing out that these assumptions need to be disclosed.

The paper is organized as follows: Section 11.2 presents the main methods used in recommender systems; Section 11.3 exposes the potential problems posed by applying them in the field of pedagogical recommender systems; the problem statement is exposed in Sect. 11.4, which is followed by an outline of proposals to overcome some harmful consequences of the use of recommender systems for e-learning (Sect. 11.5).

11.2 Methods of Recommender Systems

In the literature, recommender systems are generally classified into two types, depending on how the recommendations are made: content-based and collaborative filtering. Also other methods, such as knowledge-based filtering, more complex and difficult to implement, receive attention in specific fields, such as e-learning.

Content-based filtering (Van Meteren and Van Someren 2000; Pazzani and Billsus 2007) is a very common method for recommendations in e-commerce, VOD, etc. because it is both efficient and easy to implement. The principle is very simple, and entirely centered on the comparison between the user's interests and the item's features. The user is recommended items that most resemble the items that they have already liked, or just consulted. For that purpose, the system analyzes the user's favorite items and the features of these items to identify the user's preferences. It only requires building a model of the user preferences, using the item features and a history of the user's consultations, likes and dislikes, etc. But other information can be taken into account to build the user model, such as demographic characteristics (e.g. age, nationality, gender).

Collaborative filtering (Schafer et al. 2007; Afoudi et al. 2018) is another common method, in which the user is recommended the favorite items of the users with similar tastes. For that purpose, the system identifies the active user's closest users and the items that they have liked but that the active user has never consulted yet, in order to recommend them. The recommendation can be based on the values of recorded interactions only (memory-based approaches) or on a generative model that explains the user-item interactions in order to make predictions (model-based approaches). This method has been popularized by Amazon, considering that whether a user has consulted a product x, they should be also interested in a product z, but it is mostly used by social networks to recommend new social connections (e.g. friends, groups).

Knowledge-based filtering (Aggarwal 2016; Bouraga et al. 2014) is far more difficult to implement. It consists in recommending items based on explicit recommendation criteria, information about the user preferences, and all the characteristics of every available item. The system resorts to knowledge representation, for example in the form of rules about how an item meets a particular user. In e-commerce, they are generally used for items that are not purchased very often, such as a car, tourist destination, etc.

Since every method has drawbacks and limits (e.g. cold-start problem of collaborative filtering: in case of new users, the system does not have information about their preferences in order to make recommendations (Lika et al. 2014)), hybridization (Burke 2007; Isinkaye et al. 2015) is increasingly used to overcome them. It consists in coupling different filtering methods to make better recommendations. Many hybrid methods can be used, such as cascade (i.e. one recommender produces recommendations that will be refined by the other) or switching (i.e. the system uses a criterion dependent on the situation to switch between recommendation techniques).

Different evaluation metrics can be used to assess how a recommender system performs (Isinkaye et al. 2015). Among the most commonly used are the methods based on precision (i.e. number of selected items that are relevant) and recall (i.e. number of relevant items that are selected). In such methods, the assessment uses the user's ratings to know which items are relevant. For example, the system tries to predict the score that a user would grant to an item, then calculate how much the predicted score is from the actual score.

11.3 Recommender Systems in e-Learning

E-learning recommender systems are based on the same methods as recommender systems for e-commerce, VOD, etc. (i.e. content-based, collaborative, and knowledge-based filtering, or hybrid methods). Due to the numerous and various courses, programs, pedagogical resources, and activities available online, more and more research is conducted to propose recommender systems able to support learners, either students in school or long-life learners, in deciding which courses to follow, what resources to consult, depending on their preferences, their needs, their expectations, their skills, at a given time.

We consider that a recommender system in e-learning can be described by the following dimensions:

1. The features used to describe the learner (e.g., learning styles, explicit preferences, implicit preferences, demographic information, current skills).
2. The nature of the items to recommend (e.g., courses, friends, learning materials, graduate programs, keywords) and the features used to describe them (e.g., general topic, main concepts, format, authors, popularity).
3. The filtering techniques.

Some recommender systems are described by an additional dimension:

4. The teaching model that describes how the recommendations are provided to the student (e.g. the rules of item selection, the granularity of recommendations, the communication acts).

To a certain extent, recommender systems in e-learning have similarities with recommender systems in e-commerce. For example, they use the same filtering techniques (i.e. content-based filtering, collaborative filtering, knowledge-based filtering, or hybridization). But we notice that they also have some specificities:

1. The information can be imprecise and ambiguous. For example, learners may know what job they want to do but may not know which skills and knowledge are required.
2. The course sequencing may be crucial because some pedagogical activities need prerequisites.
3. Acquiring some knowledge is not like acquiring a concrete object: it is not enough to purchase it to get it. The context is particularly important in learning and it is crucial to determine which item is best adapted to a given situation. For example, long-life learners may be unable to devote more than fragmented periods to learning, and thus prefer short activities, while full-time students may prefer intensive and long courses.

There is not one good way to model a recommender system in e-learning. The efficiency and the relevance of the methods and features used may vary depending on the problems (i.e. the learning objectives, the intended users, the variety and the degree of specialization of course material, etc.). But, beyond their purely practical aspects (i.e. efficiency and accuracy of the recommendations, system speed and responsiveness, adaptability, etc.), some ethical issues intrinsic to the field of learning are specific to recommender systems in e-learning and have some direct practical implications. In particular, the choices of filtering techniques, learner model, and assessment methods have direct consequences on social and epistemic open-mindedness, the diversity of thinking, and the conception of knowledge and learning. As we will see, these issues are closely related to the aforementioned debates about the function and the utility of knowledge.

11.3.1 Filtering Techniques: What Implications on Social and Epistemic Open-Mindedness?

Recommender systems in e-learning use the same methods and adapt them to the specific educational field; but some of them are more likely to offer the user a choice of tools directly relevant to the current job market, thus promoting a utilitarian vision of science, while others tend to encourage and support the idealist perception of holistic knowledge by proposing more diversified courses, maybe dealing with rarely studied subjects. For example, one shortcoming of content-based methods is that they induce a lack of serendipity, that is, very few encounters with the unexpected when seeking something else. The lack of serendipitous exposure has been denounced as harmful in social media and online newspapers since it causes the apparition of “filter bubbles”, seen as huge threats to democracy and social links, polarizing and fragmenting the social space (Pariser 2011). This method is sometimes used in e-learning recommender systems, such as labelled items (Li et al. 2008) and distance between concept vectors (i.e. measurement of the difference between two concepts that are described through mathematical variables) (Ye et al. 2015),

In e-learning, these methods tend to propose courses directly related to the one that the learner has just finished, encouraging specialization. For example, in the case of content-similarity based on the distance between concept vectors (Ye et al. 2015), only the semantic relatedness is taken into account, using an automatically calculated relatedness matrix. Therefore, in this case, whether a student has just finished a course about Python programming essentials for data analysis, which deals with Python programming language, data science, and object-oriented programming, they could be proposed courses about the basics of Python programming language, dealing with Python programming language and object-oriented programming, and courses about graphics for data visualization, dealing with Python programming language, data science, and graphic tools, rather than a course about the ethical issues of AI. Still, these ethical issues could be very relevant for their whole training. Thus, some learning materials will never be proposed to the students and will even be overshadowed by the propositions of the algorithm. On the one hand, it allows scientific expertise and a good and deep comprehension of the non-trivial scientific objects that are specific to one discipline, thus increasing the learner's understanding of structures and processes. On the other hand, a certain amount of openness may be essential to remind learners of the multiplicity of the reality levels and the necessity to look beyond the disciplines, because knowledge is in essence an opening to the external world: learning is always a coming out from oneself and an act of inhabiting the world, through insights, perception, and intentionality.

That is why it must be an informed designer choice, for example when an online platform is explicitly designed to improve the learners' expertise in one targeted field. Indeed, such platforms can be very useful for learners (e.g. when they need to master a specific tool) and teachers (e.g. in the frame of one specific subject taught by a teacher, with defined resources and activities, and for which the possibility of distance-learning is offered), but they are not appropriate for general-interest e-learning platforms, where learners want to enrich their global knowledge. Since they facilitate in-depth study of one field or subject, they can hinder openness. Not only does serendipity allow one to compare and further up one's knowledge with other forms of knowledge, models, and paradigms, but it also brings interest to the discovery. More generally, the models on which the recommendations are based are important too.

11.3.2 Model Selection: A Risk of Thinking Homogenization?

As mentioned, recommender systems generally use two different information sources: the features of users and the features of items. A user model (i.e. the system's internal representation of the user's preferences, needs, expectations, etc.) can be built mainly based on users' ratings on items, users' previous navigation patterns, or the content features of purchased items. Similarly, in e-learning, a learner model, which corresponds to the user model, is mostly used to make personalized recommendations, also based on the features of the learning material, activities, paths, etc. Although some e-learning recommender systems do not use any (Ye et al. 2015), even most content-based recommender systems are based on a learner model, which expresses the learner's interests (Shu et al. 2018), internet history (Khribi et al. 2008), learning styles (Dwivedi and Bharadwaj 2013; Severac et al. 2012), etc. The choice of what is represented through this model is crucial since it refers to a specific conception of learners and strongly influences the recommendation. For example, in the Fuzzy Tree Matching-Based Personalized E-Learning Recommender System developed by Wu et al. (2015), the learner profile contains the learner's background, learning goals, prior knowledge, and learner characteristics, specified by the learner themselves when they registered; while in the Hybrid Attribute-based Recommender System for E-learning Material Recommendation proposed by Salehi and Kmalabadi (2012), the learner profile refers to their preferences obtained from their ratings. Sometimes, but rarely, educational recommender systems use a teaching model. For example, the adaptive neuro-fuzzy pedagogical recommender designed by Sevarac et al. (2012) integrate a set of modal rules based on the student's knowledge, the course sequencing, etc. in order to determine what the system should recommend to the student in a specific situation. They show that this model is very beneficial for the quality of the recommendations because, although more complex, the learner model is mostly different from the models used, either consciously or unconsciously, by the teachers to make recommendations to their students. Some learner models can encourage overspecialization and pattern reproduction, they can cause a split between different ways of thinking and ignorance of other beliefs, approaches, etc. Since they help learners to filter out information, they prevent them from being exposed to new learning perspectives.

For example, the recommender agent for e-learning systems developed by Zaiane (2002) uses data mining techniques such as association rules mining in order to build a model that represents on-line user behaviors, to suggest activities. In this case, the prediction is based on the current sequence of activities or pages visited by the learner and the other users' frequent sequences of visited pages: basically, the system "learns" from the past activities of one user or a group of users and predicts activities or pages that a given user might be interested in before suggesting them to the user. Thus, learners do not have personalized activity recommendations since these lead to the same learning patterns reproduction. This recommender agent is also based on the course sequencing: if a learner has studied the A course, then they should study the B course. Here again, the recommendations are called "personalized" while it is in fact only an appearance of tailored assistance: the learner model is just based on the history of the followed courses in order to recommend the next one. Other works use a student modelling based on the learning style (Dwivedi and Bharadwaj 2013; Severac et al. 2012). The learning style approach has been widely called into question by recent researches in education (Rohrer and Pashler 2012; Kirschner, 2017), in which it was pointed out it that this approach can have prejudicial effects, such as freezing the students in one single way of learning thus being counterproductive to the development of varied skills. However, since these models remain easy to implement, they are still often used by recommender system's designers. Such systems can reduce the learner's exposition to new ways of learning, new kinds of informational sources. Here again, this can be a pedagogical choice. Indeed, even though differentiated learning is now widely identified as a crucial approach to improve student academic engagement and success and has become a standard requirement in educative policy led in various countries (e.g. England (Mills et al. 2014), France (Kahn 2017)) notably with the aim to promote social justice, there are also strong critics, notably conceptual (Needham 2011), practical (Mahony and Hextall 2009), and ideological (Pykett 2010; Beach and Dovemark (2009)). The problem is that recommendations provided by recommender systems in e-learning are generally misleadingly presented as personalized. This lack of variety can prevent students from getting used to seeking and exploiting new information vehicles, which can be problematic if teachers expect the recommender system to take over students' exposition to multiple kinds of information. Due to their opacity, the models and data on which recommendations are based remain unknown by teachers and students, who do not have information necessary to understand what "personalized" really means in the case of the systems they use. There is a lack of information and transparency that can prevent teachers from choosing a system that really suits their pedagogical strategy and appropriately use it.

11.3.3 Assessment Methods: What Do They Value?

In some papers (Khribi et al. 2008; Wu et al. 2015; Salehi 2013), the limitations of content-based and collaborative filtering are reported in order to propose better recommendation methods, such as hybridization, coupling content-based and collaborative filtering, but they do not all result in more serendipity. For example, when designers choose the cascade method to couple the content-based and collaborative filtering, the cold-start problem is overcome, thus the recommendation results can be better, but the system still encourages overspecialization. Indeed, the collaborative filtering step only refines the results obtained by the content-based filtering step. The main problem comes from the metrics used to evaluate the system and what is actually valued.

Several methods are used to evaluate recommender systems in e-learning, such as surveys (Sevarac et al. 2012), the evolution of the number of requests after cleansing the data (Khribi et al. 2008), MAE (Bobadilla et al. 2009). Depending on the method used, various characteristics can be as assessed: surveys assess the user's satisfaction, the evolution of the number of requests assesses the utility, MAE assesses the accuracy, etc. The choice of the evaluation method and what needs to be evaluated is crucial because it both reflects the purchased main goals of the system and contributes to defining the final system design. Indeed, for example, the assessment methods can be applied to different versions of the system to decide which one to choose. It can also be applied to a single version of the system in order to know whether it can be deployed or whether the design has to be modified.

As previously mentioned, even though some researchers use more qualitative methods to assess their recommender systems in e-learning, it remains that, in many works, the metrics and assessment tools used are those developed for e-commerce recommender systems, which especially value the adequacy between the student tastes and the system suggestions. Thus, most recommender systems base their recommendations on the rating estimations (Wu et al. 2015; Salehi and Kmalabadi 2012; Bobadilla et al. 2009): given the active user's actual ratings, the system predicts the score they would give to the other available items. For instance, according to the standard assessment method, the precision can be calculated using the 80/20 method. 80% of the already rated items are used to train the system and calculate predicted ratings for the remaining 20%. Then, these predictions are compared to the actual ratings of these items. In e-commerce, precision is crucial in order to propose to the consumer items that they are likely to desire to buy. But in e-learning, this assessment method raises various issues, either pedagogical or ideological, which could seem crucial for teachers.

Concerning the pedagogical issues, it could be pointed out notably the fact that, when using this method, it is assumed that the score prediction precision is the most important metric to assess the recommendation quality, that is, the most important consideration is the student's likes or dislikes. However, this assumption could be questioned. There are various reasons why other considerations should be considered. For example, the learning courses, activities, and materials need to be arranged in order to ensure that the prerequisites of some courses are acquired. Moreover, how can we know what the scores provided by the learners are based upon? Did they assess the form or the content? If they have scored the content, did they find it interesting, easily understood, useful for their whole training, or relevant to their current concerns? Were they happy to acquire new knowledge, face new and challenging issues, or have a quick and easy task? Yet, whether learners have to like their course material and learning activities is a philosophical, educational, and sociological issue, which could be seriously questioned. For example, even though it is proven that enjoyment positively influences the didactic process and the memorization of information (Hernik and Jaworska 2018; Pekrun et al. 2009) in particular by increasing motivation, positive emotions do not always result in efficient learning: for example, relaxation can affect learning by causing over-confidence (Pekrun et al. 2011). In particular, Henritius et al. (2019) argue that students' satisfaction can be a misleading indicator of learning, which often requires stepping outside one's comfort zone, thereby creating emotions (such as discomfort) rarely associated with satisfaction. Furthermore, efficiency of the encoding and memorization of information may not be the more essential criterion. Indeed, more fundamentally, the very aim (or aims) of following courses has to be examined: it can be defended that the pursued goal is acquiring knowledge, or mastering mental tools (e.g. scientific methods) to understand and analyze the world, developing a critical mind, etc. Yet, in this latter perspective, the confrontation of students to questions, information, and ideas that they do not like could be also useful, and even necessary. Moreover, the context of learning is important too, and the learner's needs can change over time, depending on their projects, the knowledge they acquire, the new challenges they meet, the techniques they master, etc.

Secondly, from an ideological point of view, teachers could fear that such a method can result in the commodification of knowledge, which is a process accused of valuing knowledge in relation to its economic productivity, and transforming students into consumers who can expect that a commodity offers the service for which it has been produced. This issue is closely related to the concept of "Cognitive capitalism" (Blondeau and Latrive 2020; Fumagalli and Lucarelli 2010) which involves the transformation of an intellectual good into a commodity and/or a resource. It could seem problematic for several reasons. First of all, it may be argued that it is hardly compatible with the idea of a free and disinterested activity, an activity with no defined goal, a *praxis* that is summoned whatever the purpose is. The economic rationale is built on an instrumental rationale, on predictable behaviors and results. A second fear could be that the creation of a knowledge market requires regulating and restricting access to knowledge, leading to inequality issues since access to knowledge depends, among other things, on the economic conditions of an individual. Finally, and above all, according to Blondeau and Latrive, cognitive capitalism encourages eliminating the knowledge regarded as useless or even counterproductive because of time-wasting (Blondeau and Latrive 2020): if only knowledge that supports growth is useful and deserves to be transmitted, the rest leads to nothing but sterile debates and questionings. From this standpoint, the knowledge about emerging technologies, for example, is valued, since it can lead to a continuous producing and selling of endlessly improved items. On the contrary, philosophical concepts such as ethics are often and prejudicially neglected, as shown in some papers (Lauer 2021), and social issues are considered as secondary. Some studies have shown that, although social sciences are essential to address climate change and energy transition, these fields receive very little funding for climate-related research (Overland and Sovacool 2020). This is a crucial and structural problem; yet, it is well-known now that artificial intelligence tends to reproduce biases inherent to data used for machine learning.

In addition to the ideology and the perception of education conveyed by such practices, they have effects on student learning, which should be known before choosing to integrate them in a pedagogical platform or strategy. For example, some students take online courses with the only purpose of being awarded a certification. Similarly, in order to increase their revenue, some institutions offer online courses so that they can reduce the number of actual human teachers while proposing courses to a higher number of students. From this standpoint, providing the students with quick access to resources and activities can be regarded as beneficial. But the other side of the coin is that learners develop a habit to trust the recommendations of the algorithms, instead of seeking and selecting the appropriate information on their own, yet this cognitive routine should be encouraged and trained. This is particularly true with recommender systems that propose keywords for the student's requests (Li et al. 2008) or ordered items in terms of supposed relevancy. Indeed, education is not only transmission of learning content and vocational knowledge and know-how, but also about teaching critical thinking and reasoning, seeking information and comparing sources, debating and discussing. Both are not mutually exclusive and reuniting them is maybe one of the main challenges that universities must face in these times of quest for profits and efficiency.

11.4 Problem Statement

According to the previous analyses, we have identified three main issues that could seem problematic for teachers and institutions when they choose to integrate a recommender system in their learning strategy or platform:

- The commodification of knowledge, which is pointed out for valuing it in relation to its economic productivity, subjecting knowledge to the market law and transforming students into consumers.

- The specialization effect, that is the fostering of students' exposure to close learning contents or forms.
- The rationalization process, which is conceptualized by those who oppose it a trend to establish as indisputable truth the results coming from dominating models, promote the idea of a scientific consensus on many subjects and break with the prolific scientific method and philosophical methods, which prompt caution, comparison, verification, and dialectics.

The preoccupations raised by these issues are mainly ideological and pedagogical. We think that there are three sides of the problem:

1. The rigidity of recommender systems, which are generally unable to adapt their recommendations to the teacher's pedagogical approach and needs
2. The lack of specialists in the science of education in recommender system designer teams
3. The lack of transparency about how recommender systems deployed for students' learning work

Since these issues seem very controversial in scientific literature about education, we argue that addressing these both problems is a crucial ethical matter.

For this purpose, we propose these main lines of improvement: fostering systems able to adapt to the teacher's specific pedagogical approach requirements, ensuring an epistemic liability of the models used to design recommender systems, allowing users to understand the underpinning reasons of recommendations, their potentialities and their limits.

11.5 Some Proposals

11.5.1 Knowledge-Based Recommendations

In the case of e-learning platforms proposing a wide range of courses, knowledge-based methods appear to be a promising way to connect concepts and give them meaning, thus drawing a network of possible paths. Ontologies can be a good solution to encode the semantic and modal relationships between the concepts. In this way, the system could be able to recommend a variety of different courses and contents, in a relevant proportion, given the degree of correspondence between the course that the student is currently following and the available items.

It can also be interesting within the scope of a specific course, when the recommender system is used to help the student find appropriate resources or activities when they feel blocked, want to go further, etc. Besides, in order to implement a teaching model (i.e. implementation of the teachers' pedagogical strategies), some works already propose knowledge-based recommender systems. It is a very interesting approach since it allows us to integrate and use more complex information and bond them in order to propose a more tailored assistance. For example, Sevarac (2012) proposes to define high-level rules, easily understood and used by teachers, so that they can decide what activities will be recommended for every set of learners. The fuzzy sets describe the student's knowledge of some topics and the preferred learning style. In this proposed solution, teachers do not have much room for maneuver yet and the student model (i.e., current knowledge and learning styles) still appears limited, all the more as the use of learning style is questionable. But it gives a good idea of what can be done, and what should be improved to provide the students with assistance tailored to suit the teacher's approach, thus ensuring pedagogical continuity, while doing their homework for example. This requires close interactions between recommender systems and teachers, by means of meaningful feedback, and easy to learn and use setting-up tools.

11.5.2 A Learner Model Coming from Cognitive and Educational Sciences

Intelligent tutoring systems are often based on learning and teaching models, chosen according to several characteristics such as the pedagogical goals and available resources. Proposing a suitable learner model, not judgmental and able to meet the real and precise needs of every student, is a major challenge for e-learning recommender systems. Indeed, conceiving an AI-system, which necessarily works using categories and labelling, whose recommendations could suit the individual characteristics of human students, seems very difficult and requires careful design. For that purpose, a solution can be to use learner models that come from the cognitive and educational sciences, since a deep reflection has already been led to conceive models that at best allow to express and represent the learners' specificities. Of course, there is not any ideal model that enables a representation of all the most relevant characteristics of learners, and a meticulous analysis should be systematically conducted to find the most accurate and appropriate one, depending on the specific aims of every recommender system and the kind of assistance that is expected to be provided. For example, in intelligent tutoring systems dedicated to assisting learning, the cognitivist approach (Anderson and Gluck 2001), as opposed to behaviorism, aims to explain the learner's behavior changes through mental operations. Indeed, in the behaviorist approach, learning is viewed through the prism of the stimulus-response relationship. The complexity of the learner's cognition is not denied but it is considered as a black box that is not intended to be opened (Skinner 1974). On the contrary, in the cognitivist approach (Anderson 1996), the cognitive box is opened: the learner's cognitive processes are broken down into interconnected subprocesses and stages.

11.5.3 A Teaching Model Based on Empiric Analyses

Teachers have various strategies to take decisions about the right didactic action to take in response to the student's observed situation, in particular basing their choice on some epistemic factors such as the knowledge at stake and the learner's knowledge state. The teacher's diagnosis of this situation is crucial for their recommendations. In recommender systems, the learning models allow for the representation of these learning situations, but they have thereafter to be adequately interpreted to produce recommendations with real pedagogical value. In such a perspective, building pedagogical scenarios is not enough, since they do not ensure that the epistemic dimensions (e.g., the knowledge organization and acquisition) will be taken into account. A teaching model has therefore to be used in order to organize knowledge acquisition in a given learning situation. Teaching models are generally given by empiric analyses conducted by teachers. For example, they can be based on generic models such as the Socratic method (i.e., a dialogue about the studied issue is carried on by the system with the student, for example by presenting them with different cases, probing for relevant factors, asking for predictions, entrapping the student when they have not identified all necessary factors, presenting counterexamples, etc.), implemented in several versions (Collins and Stevens 1991; Lepper et al. 1993) or the analysis of the teacher's expertise (Lajoie et al. 2001; Heffernan and Koedinger 2002). Schoenfeld (1998) studies the teacher's behavior without proposing an automatic model but by investigating, for example, the role played by beliefs, knowledge, goals, etc. in school management, teaching practice, adaptation, etc.

11.5.4 Explainable Recommendations

In e-commerce, recommender systems are mostly designed to act alone. Yet, in e-learning, it would seem very relevant to use a recommender system to team up / collaborate with humans, either teacher or learner, or both. It could serve many objectives. First, regarding learners, a recommender system could be used to both assist them and guide them through the wide range of proposed resources, just as they already do, and improve learner agency. For that purpose, the system should be explainable. In e-commerce, some systems are described as explainable because the reason why the items are suggested is made explicit. For example, on marketplaces, customers can read explanations such as "people who liked this item also liked this one", "you liked this item, you may be interested in this one", etc. The main goal is to increase the user's attention and interest. But, in e-learning, on the student side, the main goals would be to actively engage the students in their learning, provide them with the tools to understand their learning behavior, thus think over it and adapt it, and help them to decide whether they want to follow the recommendations or not. It is about making the learners able to understand and deal with their metacognitive strategies. Indeed, this approach, coupled with the use of learner models coming from educational sciences, could help students to understand their mental mechanisms and act on their motivation, learning strategies, etc.

On the teacher side, the goal would be to increase their understanding of their students' behavior and improve following-up and monitoring through scoreboards and alerts for example, as well to have control over the models used and the recommendations made, so that they can implement their teaching model. Indeed, there are reasons to believe that teachers should be assisted instead of excluded from the recommendation process (i.e. AI-systems should be used to team up with humans instead of replacing them). First, recent research has demonstrated that enhanced AI-systems accuracy does not always lead to better system performance (Yin et al. 2019; Lai and Tan 2019) because performance is rather closely linked to the quality of the relationship between the human and the AI as a team (i.e. trust, knowledge of the limits and the potentials of the AI, understanding of system operation, etc.). Moreover, as we have mentioned, even though there are techniques to assess system accuracy, the teacher remains the most qualified to evaluate the relevance of a didactic recommendation and its ability to match to their expectations. For example, they can examine precisely whether the system has taken into account the suitable epistemic factors to make the right decision, and whether the recommendation fits their own pedagogical model. Finally, despite teacher shortage (Flynt and Morton 2009; Hutchison 2012; Ingersoll and May 2011; Martino Rezai-Rashti 2010), various arguments emerge for slowing down the excessive automation of education, including the issues about data privacy, lack of control, responsibility, etc. For example, in studying the impact of artificial intelligence on learning, teaching, and education, Tuomi (2018) explains that AI can limit the domain where humans express their agency. They also remind us that there may be fundamental theoretical and practical limits in designing AI systems that can explain their behavior and decisions so that it is important to keep humans in the decision-making loop. Similarly, Selwyn (2019) expresses his mistrusts (e.g. concerns about inaccuracies, misrecognition, and faulty decision-making) with regard to the very fast spread of AI-systems in every sphere of our lives and the strong enthusiasm, insufficiently supported by philosophical questioning, though, for its potentialities.

Finally, on the system side, the aim is to be able to learn, not only automatically but also with the human feedback and settings up, as well as the integration of teaching rules. For example, in the system that we are currently developing with my team, teachers receive information about a given student (i.e., their general profile, their current on-task behavior, the improvements that should be done by the student on the current task, the recommendations provided by the system to the student, and the student's feedback about the recommendations and their explanation). In this way, the teacher has full information to decide whether the recommendations were appropriate or not and readjust the settings of the recommender system if necessary (Roux et al. 2021).

11.6 Discussion and Conclusion

In the complex environment in which we live, it appears important to be able to think and question the different objects we have to deal with. Many matters about AI for example (but it is true for other fields as well) cannot be reduced to the technological perspective: they also require philosophical, sociological, economic, anthropological, etc. views since experts may have to put different concepts and reality levels into dialogue. Making recommendations of learning objects for students goes beyond technical issues, thus it is not enough to rely on the methods that work for sales and VOD: such project require knowledge and understanding of pedagogical issues in order to choose the algorithms and models that appear the more relevant for a given pedagogical context and purpose. Moreover, this context and purpose should be well-defined and made clear to users, so that teachers can make informed choice when selecting a recommender system and integrating it in their pedagogical strategy.

Thus, the current questionings about the lack of transparency and fairness of recommender systems dedicated to e-commerce are all the more crucial in education because it deeply affects the individuals' relationship to the world. Our study show that any existing educational recommender systems encourage overspecialization and the reproduction of the same behavioral patterns, at the expense of openness and diversity. Some of them also result in reducing the individual abilities to seek and compare information, verify the sources, and make their own informed choices. The use of recommender systems can be beneficial for reducing the inequalities in learning, for example, to enable working students to access online courses and help teachers in monitoring; but studies have shown that it also can have harmful effects if their design is only driven by economic imperatives, or the ethical and social consequences are not carefully examined. One of the main problems is that e-learning recommender systems use the same methods (e.g. filtering and assessment methods) as the recommender systems designed for e-commerce, whose main purposes are profits and speed, even if this means deteriorating the forum of public discourse and amplifying patterns of discrimination and disadvantage (Milano et al. 2021).

While using the usual techniques of recommender systems without questioning their implications is a root source of these problems, some solutions, both technical and educational, can be proposed to address it. Designing knowledge-based methods, using learner models based on careful educational and cognitive studies, and providing explained recommendations that enable learners to be actively involved in their training can be part of a solution. Other avenues could be explored, for example investigating the most appropriate items to recommend (e.g., keywords for queries, friends, pedagogical activities) and the way to present them (e.g., top-ranked lists, ordered lists, a spontaneous recommendation for one single item). All of these suggestions raise technical, educational, and social questions that should be the objects of debate and careful examination when designing a recommender system. Finally, and above all, systematic ethical and epistemic questioning should become the guiding principles of any technical research.

References

Adomavicius, G., J. Bockstedt, S.P. Curley, J. Zhang, and S. Ransbotham. 2019. The Hidden Side Effects of Recommendation Systems. *MIT Sloan Management Review* 60 (2): 1.

Afoudi, Y., M. Lazaar, and M. Al Achhab. 2018. Collaborative Filtering Recommender System. In *Proceedings of the International Conference on Advanced Intelligent Systems for Sustainable Development*, 332–345. Cairo/Cham: Springer. https://doi.org/10.1007/978-3-030-11928-7_30.

Aggarwal, C.C. 2016. Knowledge-Based Recommender Systems. In *Recommender Systems*, 167–197. Cham: Springer. https://doi.org/10.1007/978-3-319-29659-3_5.

Amasha, M.A., M.F. Areed, S. Alkhalaf, R.A. Abougalala, S.M. Elatawy, and D. Khairy. 2020. The Future of Using Internet of Things (loTs) and Context-Aware Technology in E-learning. In *Proceedings of the 2020 9th International Conference on Educational and Information Technology*, 114–123. Oxford: Association for Computing Machinery. https://doi.org/10.1145/3383923.3383970.

Anderson, J.R. 1996. ACT: A Simple Theory of Complex Cognition. *American Psychologist* 51 (4): 355. https://doi.org/10.1037/0003-066X.51.4.355.

Anderson, J. R., and K. Gluck. 2001. What Role Do Cognitive Architectures Play in Intelligent Tutoring Systems. In *Cognition & Instruction: Twenty-Five Years of Progress*, ed. Carver, vol. 35, issue 6, 689–704.

Beach, D., and M. Dovemark. 2009. Making 'right'choices? An ethnographic account of creativity, performativity and personalised learning policy, concepts and practices. *Oxford Review of Education* 35 (6): 689–704.

Blondeau, O., and F. Latrive. 2020. *Libres enfants du savoir numérique. Une anthologie du "Libre"*. Paris: l'Éclat.

Bobadilla, J., F.J. Serradilla, and A. Hernando. 2009. Collaborative Filtering Adapted to Recommender Systems of E-Learning. *Knowledge-Based Systems* 22 (4): 261–265. https://doi.org/10.1016/j.knosys.2009.01.008.

Bouraga, S., I. Jureta, S. Faulkner, and C. Herssens. 2014. Knowledge-Based Recommendation Systems: A Survey. *International Journal of Intelligent Information Technologies (IJIIT)* 10 (2): 1–19. https://doi.org/10.4018/ijiit.2014040101.

Brighouse, H., and M. McPherson. 2015. *The Aims of Higher Education*. Chicago: University of Chicago Press.

Burke, R. 2007. Hybrid Web Recommender Systems. In *The Adaptive Web*, ed. Peter Brusilovsky, Alfred Kobsa, and Wolfgang Nejdl, 377–408. Berlin/Heidelberg: Springer. https://doi.org/10.1007/978-3-540-72079-9.

Collins, A., and A.L. Stevens. 1991. A Cognitive Theory of Inquiry Teaching. In *Teaching Knowledge and Intelligent Tutoring*, ed. Peter Goodyear, 203–230. Norwood: Ablex Publishing.

Dwivedi, P., and K.K. Bharadwaj. 2013. Effective Trust-Aware E-Learning Recommender System Based on Learning Styles and Knowledge Levels. *Journal of Educational Technology & Society* 16 (4): 201–216. https://www.jstor.org/stable/jeductechsoci.16.4.201.

Florian, O. 2018. Le rôle attribué par les étudiants aux études: un utilitarisme dominant? *Cahiers de la recherche sur l'éducation et les savoirs* 17: 239–258. http://journals.openedition.org/cres/3807.

Flynt, S.W., and R.C. Morton. 2009. The Teacher Shortage in America: Pressing concerns. *National Forum of Teacher Education Journal* 19 (3): 1–5. http://www.nationalforum.com/Electronic%20Journal%20Volumes/Flynt,%20Samuel%20Teacher%20Shortage%20in%20America.pdf.

Fraihat, Salam, and Qusai Shambour. 2015. A Framework of Semantic Recommender System for e-Learning. *Journal of Software* 10 (3): 317–330. https://doi.org/10.17706/jsw.10.3.317-330.

Fumagalli, A., and S. Lucarelli. 2010. Cognitive Capitalism as a Financial Economy of Production. In *Cognitive Capitalism and its Reflections in South-Eastern Europe*, ed. Vladimir Cvijanovic, Andrea Fumagalli, and Carlo Vercellone, 9–40. Frankfurt: Peter Lang.

Gallego, D., E. Barra, S. Aguirre, and G. Huecas. 2012. A Model for Generating Proactive Context-Aware Recommendations in e-Learning Systems. In *Proceedings of the 2012 Frontiers in Education Conference*, 1–6. Seattle: IEEE. https://doi.org/10.1109/FIE18277.2012.

Hallinan, B., and T. Striphas. 2016. Recommended for You: The Netflix Prize and the Production of Algorithmic Culture. *New Media & Society* 18 (1): 117–137. https://doi.org/10.1177/1461444814538646.

Hayles, N.K. 1999. *How We Became Posthuman: Virtual Bodies in Cybernetics, Literature, and Informatics*. Chicago: University of Chicago Press.

Heffernan, N.T., and K.R. Koedinger. 2002. An Intelligent Tutoring System Incorporating a Model of an Experienced Human Tutor. In *Proceedings of the 6th International Conference "Intelligent Tutoring System"*, 596–608. Biarritz: Springer.

Henritius, E., E. Löfström, and M.S. Hannula. 2019. University Students' Emotions in Virtual Learning: A Review of Empirical Research in the 21st Century. *British Journal of Educational Technology* 50 (1): 80–100. https://doi.org/10.1111/bjet.12699.

Hernik, J., and E. Jaworska. 2018. The Effect of Enjoyment on Learning. In *Proceedings of the INTED2018 Conference*, 508–514. Valencia: IATED. https://doi.org/10.21125/inted.2018.

Hutchison, L.F. 2012. Addressing the STEM Teacher Shortage in American Schools: Ways to Recruit and Retain Effective STEM Teachers. *Action in Teacher Education* 34 (5–6): 541–550. https://doi.org/10.1080/01626620.2012.729483.

Ingersoll, R.M., and H. May. 2011. The Minority Teacher Shortage: Fact or Fable? *Phi Delta Kappan* 93 (1): 62–65. https://doi.org/10.1177/003172171109300111.

Isinkaye, F.O., Y.O. Folajimi, and B.A. Ojokoh. 2015. Recommendation Systems: Principles, Methods and Evaluation. *Egyptian Informatics Journal* 16 (3): 261–273. https://doi.org/10.1016/j.eij.2015.06.005.

Jacob, M. 2003. Rethinking Science and Commodifying Knowledge. *Policy Futures in Education* 1 (1): 125–142. https://doi.org/10.2304/pfie.2003.1.1.3.

Kahn, S. (2017). *Pédagogie différenciée: Guide pédagogique*. De Boeck (Pédagogie et Formation).

Khribi, M.K., M. Jemni, and O. Nasraoui. 2008. Automatic Recommendations for E-Learning Personalization Based on Web Usage Mining Techniques and Information Retrieval. In *Proceedings of the 2008 Eighth IEEE International Conference on Advanced Learning Technologies*, 241–245. Santander: IEEE Computer Society. https://doi.org/10.1109/ICALT.2008.198.

Kirschner, P.A. 2017. Stop Propagating the Learning Styles Myth. *Computers & Education 106*: 166–171.

Lai, V., and C. Tan. 2019. On Human Predictions with Explanations and Predictions of Machine Learning Models: A Case Study on Deception Detection. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, 29–38. New York: Association for Computing Machinery. https://doi.org/10.1145/3287560.3287590.

Lajoie, S.P., J. Wiseman, and S. Faremo. 2001. Identifying Human Tutoring Strategies for Effective Instruction in Internal Medicine. *International Journal of Artificial Intelligence in Education*: 293–309. http://users.sussex.ac.uk/~bend/its2000/lajoie.pdf.

Lauer, D. 2021. You Cannot Have AI Ethics Without Ethics. *AI and Ethics* 1 (1): 21–25. https://doi.org/10.1007/s43681-020-00013-4.

Lepper, M.R., M. Woolverton, D.L. Mumme, and J. Gurtner. 1993. Motivational Techniques of Expert Human Tutors: Lessons for the Design of Computer-based Tutors. In *Computers as Cognitive Tools*, ed. Susanne P. Lajoie and Sharon J. Derry, 75–105. Mahwah: Lawrence Erlbaum Associates.

Li, L., Z. Yang, L. Liu, and M. Kitsuregawa. 2008. Query-URL Bipartite Based Approach to Personalized Query Recommendation. In *Proceedings of the Twenty-Third AAAI Conference on Artificial Intelligence*, 1189–1194. Chicago: AAAI Press. https://www.aaai.org/Papers/AAAI/2008/AAAI08-188.pdf.

Lika, B., K. Kolomvatsos, and S. Hadjiefthymiades. 2014. Facing the Cold Start Problem in Recommender Systems. *Expert Systems with Applications* 41 (4): 2065–2073. https://doi.org/10.1016/j.eswa.2013.09.005.

Mahony, P., & I. Hextall. 2009. Building schools for the future and the implications for becoming a teacher. Paper presented at the European Conference on Educational Research, Vienna, 28–30 September.

Manouselis, N., H. Drachsler, K. Verbert, and E. Duval. 2012. *Recommender Systems for Learning*. New York: Springer.

Martino, W., and G.M. Rezai-Rashti. 2010. Male Teacher Shortage: Black Teachers' Perspectives. *Gender & Education* 22 (3): 247–262. https://doi.org/10.1080/09540250903474582.

Mbipom, B., S. Massie, and S. Craw. 2018. An E-learning Recommender that Helps Learners Find the Right Materials. In Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, vol. 32, issue 1. Palo Alto: AAAI Press. https://doi.org/10.1609/aaai.v32i1.11389.

Milano, S., B. Mittelstadt, S. Wachter, and C. Russell. 2021. "Epistemic Fragmentation Poses a Threat to the Governance of Online Targeting." *Nature Machine Intelligence* 3 (6): 466–472. https://doi.org/https://doi.org/10.1038/s42256-021-00358-3.

Mills, M., S. Monk, A. Keddie, P. Renshaw, P. Christie, D. Geelan, and C. Gowlett. 2014. Differentiated Learning: From Policy to Classroom. *Oxford Review of Education* 40 (3): 331–348.

Moore, P., Z. Zhao, and H. van Pham. 2019. Towards Cloud-Based Personalised Student-Centric Context-Aware e-Learning Pedagogic Systems. In *Proceedings of the 13th Conference on Complex, Intelligent, and Software Intensive Systems*, 331–342. Cham: Springer. https://doi.org/10.1007/978-3-030-22354-0_30.

Needham, C. 2011. Personalization: From Story-Line to Practice. *Social Policy and Administration* 45 (1): 54–68.

Overland, I., and B.K. Sovacool. 2020. The Misallocation of Climate Research Funding. *Energy Research & Social Science* 62: 101349.

Pariser, E. 2011. *The Filter Bubble: What the Internet is Hiding from you*. London: Penguin UK.

Pazzani, M.J., and D. Billsus. 2007. Content-Based Recommendation Systems. In *The Adaptive Web*, ed. Peter Brusilovsky, Alfred Kobsa, and Wolfgang Nejdl, 325–341. Berlin: Springer. https://doi.org/10.1007/978-3-540-72079-9_10.

Pekrun, R., A.J. Elliot, and M.A. Maier. 2009. Achievement Goals and Achievement Emotions: Testing a Model of their Joint Relations with Academic Performance. *Journal of Educational Psychology* 101 (1): 115–135. https://doi.org/10.1037/a0013383.

Pekrun, R., T. Goetz, A.C. Frenzeld, P. Barchfeld, and R.P. Perrye. 2011. Measuring Emotions in Students' Learning and Performance: The Achievement Emotions Questionnaire (AEQ). *Contemporary Educational Psychology* 36 (1): 36–48. https://doi.org/10.1016/j.cedpsych.2010.10.002.

Pykett, J. 2010. Personalised governing through behaviour change and re-education. PSA Conference Paper, Edinburgh.

Rafaeli, S., Y. Dan-Gur, and M. Barak. 2005. Social Recommender Systems: Recommendations in Support of e-learning. *International Journal of Distance Education Technologies (IJDET)* 3 (2): 30–47. https://doi.org/10.4018/jdet.2005040103.

Rohrer, D., and H. Pashler. 2012. Learning Styles: Where's the Evidence? *Online Submission 46* (7): 634–635.

Roux, L, P. Dagorret, T. Etcheverry, T. Nodenot, C. Marquesuzaa, and P. Lopisteguy. 2021. A Multi-Layer Architecture for an E-Learning Hybrid Recommender System. Paper presented at the 18th International Conference on Cognition and Exploratory Learning in Digital Age. IADIS Press.

Salehi, M. 2013. Application of Implicit and Explicit Attribute Based Collaborative Filtering and BIDE for Learning Resource Recommendation. *Data & Knowledge Engineering* 87: 130–145. https://doi.org/10.1016/j.datak.2013.07.001.

Salehi, M., and I.N. Kmalabadi. 2012. A Hybrid Attribute–Based Recommender System for E–learning Material Recommendation. *IERI Procedia* 2: 565–570. https://doi.org/10.1016/j.ieri.2012.06.135.

Schafer, J.B., D. Frankowski, J. Herlocker, and Shilad Sen. 2007. Collaborative Filtering Recommender Systems. In *The Adaptive Web*, ed. Peter Brusilovsky, Alfred Kobsa, and Wolfgang Nejdl, 291–324. Berlin: Springer. https://doi.org/10.1007/978-3-540-72079-9_9.

Schoenfeld, A.H. 1998. Towards a theory of teaching-in-context. *Issues in Education* 4 (1): 1–94. https://doi.org/10.1016/S1080-9724(99)80076-7.

Segal, A., Z. Katzir, Y. Gal, G. Shani, and B. Shapira. 2014. Edurank: A Collaborative Filtering Approach to Personalization in E-Learning. In *Proceedings of the Seventh International Conference on Educational Data Mining*, 68–75. London: International Educational Data Mining Society.

Selwyn, N. 2019. *Should Robots Replace Teachers?: AI and the Future of Education*. Cambridge: Polity Press.

Sevarac, Z., V. Devedzic, and J. Jovanovic. 2012. Adaptive Neuro-fuzzy Pedagogical Recommender. *Expert Systems with Applications* 39 (10): 9797–9806. https://doi.org/10.1016/j.eswa.2012.02.174.

Shi, D., T. Wanga, H. Xinga, and H. Xu. 2020. A Learning Path Recommendation Model Based on a Multidimensional Knowledge Graph Framework for E-Learning. *Knowledge-Based Systems* 195: 105618. https://doi.org/10.1016/j.knosys.2020.105618.

Shu, J., X. Shen, H. Liu, B. Yi, and Z. Zhang. 2018. A Content-Based Recommendation Algorithm for Learning Resources. *Multimedia Systems* 24 (2): 163–173. https://doi.org/10.1007/s00530-017-0539-8.

Skinner, B.F. 1974. *About Behaviorism*. New York: Random House.

Tuomi, I. 2018. *The Impact of Artificial Intelligence on Learning, Teaching, and Education*. Luxembourg: Publications Office of the European Union.

Van Meteren, R., and M. van Someren. 2000. Using Content-Based Filtering for Recommendation. In Proceedings of the Machine Learning in the New Information Age MLnet/ECML2000 Workshop, 47–56.

Wu, D., J. Lu, and G. Zhang. 2015. A Fuzzy Tree Matching-Based Personalized E-Learning Recommender System. *IEEE Transactions Fuzzy Systems* 23 (6): 2412–2426. https://doi.org/10.1109/TFUZZ.2015.2426201.

Ye, M., Z. Tang, J. Xu, and L. Jin. 2015. Recommender System for E-Learning Based on Semantic Relatedness of Concepts. *Information* 6 (3): 443–453. https://doi.org/10.3390/info6030443.

Yin, M., J. Wortman Vaughan, and H. Wallach. 2019. Under-Standing the Effect of Accuracy on Trust in Machine Learning Models. In *Proceedings of the 2019 chi Conference on Human Factors in Computing Systems*, 1–12. Glasgow: Association for Computing Machinery.

Zaiane, O.R. 2002. Building a Recommender Agent for E-Learning Systems. In *International Conference on Computers in Education*, 55–59. Washington, D.C.: IEEE Computer Society. https://doi.org/10.1109/CIE.2002.