Információs Társadalom

[Information Society]

A SOCIAL SCIENCE JOURNAL

Founded in 2001

English Issue Vol. XXI, No. 2 (2021)

Published by INFONIA Foundation

Principal sponsor: Budapest University of Technology and Economics, Faculty of Economic and Social Sciences

Technical partner: SZTAKI

ISSN 1578-8694

Produced by Server Line Print & Design, Budapest

The journal Information Society (In Hungarian: Információs Társadalom, abbreviated as InfTars) aims to provide a platform for research and discussion of the theories and applications of Information Society Studies. Currently every second issue is published in English, the rest are in Hungarian.

The journal is fully Open Access and freely available at https://inftars.infonia.hu/

InfTars is indexed in both the *Web of Science Social Sciences Citation Index and Scopus*, and all titles are automatically submitted to *Crossref*.

Since Vol. VIII, No. 1 (2008) the journal has been included in the Thomson Reuters index (Social Sciences Citation Index, Social Scisearch, Journal Citation Reports/Social/Sciences Edition)

E-mail: inftars-lapman@ponens.org

CONTENTS

MIHÁLY HÉDER

AI and the resurrection of Technological Determinism

This paper elaborates on the connection between the AI regulation fever and the generic concept of Social Control of Technology. According to this analysis, the amplitude of the regulatory efforts may reflect the lock-in potential of the technology in question. Technological lock-in refers to the ability of a limited set of actors to force subsequent generations onto a certain technological trajectory, hence evoking a new interpretation of Technological Determinism. The nature of digital machines amplifies their lock-in potential as the multiplication and reuse of such technology is typically almost cost-free. I sketch out how AI takes this to a new level because it can be software and an autonomous agent simultaneously. The research was supported by the János Bolyai Research Scholarship of the Hungarian Academy of Sciences

LECTORI SALUTEM!

We are proud to present this Special Issue on the theoretical and practical issues around the Ethics of Artificial Intelligence. The issue was supported by the BME Faculty of Economic and Social Sciences (as always), ELKH SZTAKI and the The Artificial Intelligence National Laboratory of Hungary (MILAB).

The foreword of the issue is presented by our distinguished guest Dr George Tilesch, author of *BetweenBrains: Taking Back Our AI Future,* and founder and president of the Phi Institute.

Then, we find ourselves right in the middle of it all. Aron Dombrovszki investigates some ethical problems around Autonomous Weapons Systems, in particular whether there is an unjust bias against them. Appealing for moral consistency, Dombrovszki approaches the problem through the conceptual framework of the Just War Theory, before expanding his coverage to other AI technologies and to possible resolutions of the apparent double standards within the field of AI robotics.

Attila Gyulai and Anna Ujlaki adopt a political theoretical perspective to address the problem of AI regulation. By disregarding the political problem of enforceability, they argue that the applied ethics approach dominant in the discussions on AI regulation is incomplete. Applying realist political theory, the article demonstrates how prescriptive accounts of the development, use and functioning of AI are necessarily political.

Karoline Reinhardt elaborates the questions a diversity-sensitive social platform raises with regard to responsibility, and proposes a tentative ethical framework of responsibility for these platforms. This framework could help balance between the need for respecting human diversity and avoiding unjust discrimination, pitfalls and privacy concerns arising from the collection, processing and storage of diversity-related data.

Kinga Sorbán, in her paper on recommendation engines, introduces the relevant rules present in the existing ethical guidelines and elaborates on how they are to be applied within the streaming industry. The paper strives to adopt a critical standpoint towards the provisions of the ethical guidelines in place, arguing that adopting a one-size-fits-all approach is not effective due to the specificities of the content distribution industry.

Constantin Vică, Cristina Voinea and Radu Uszkai infuse another aspect of realism – not unlike Gyulai and Ujlaki – into the AI ethics debate. Their concern is that, instead of a genuine ethical interest for AI, we are witnessing moral diplomacies resulting in moral bureaucracies battling for moral supremacy and political domination. They provide a short overview of what they term 'ethics washing' in the AI industry, analyze a case study, then conclude with some general considerations regarding the future of the ethics of AI.

This issue also contains a regular paper, not connected to the special topic. In our penultimate article, Peter Konhäusner, Maria Margarita Cabrera Frias and Dan-Cristian Dabija investigate the methods being implemented by platforms to use monetary incentives on their crowd. Their paper uses an inductive empirical meth-

od based on grounded theory, with data gathered from 15 different platforms that are known to be using monetary incentivization methods, to analyze and categorize the different strategies used by the platforms and their marketing objectives. The authors present useful information to assist managers to make the right decisions regarding monetary incentives and for fostering the potential of their crowd.

Finally, back in the AI regulation topic, in the last paper of this issue Héder elaborates the connection he sees between the AI regulations fever and the generic concept of Social Control of Technology. According to his analysis, the amplitude of the regulatory efforts may reflect the lock-in potential of the technology in question. This refers to the ability of a limited set of actors to force subsequent generations onto a certain technological trajectory, hence evoking a new interpretation of Technological Determinism for Artificial Intelligence.

PRELUDE

Dear Reader,

As both the volume and range of the studies appearing in this fine bouquet will clearly demonstrate, in the last few years AI Ethics became a burgeoning field that increasingly permeates business and policy, beyond its academic roots. But does the field of AI ethics and governance as well as its sizeable global community of practitioners pull their weight? Does this domain and its fervent advocates fulfill their purpose as embodying adequate social control over the spreading of a technology of unprecedented power, exponentiality and fickleness: AI?

We belong to the school of thought that treats AI Ethics very holistically - and equally pragmatically. Our compeers also believe that in its broadest definition, ownership of Ethical AI belongs to an informed society, composed of responsible digital citizens who drive related social movements - not just to professional practitioners.

But as this question is being often approached in the present series of studies, is control of AI technology possible? Can regulation solely aim at that? We are of a belief that even if *control* is and will stay elusive, *steering* or *stewardship* should be set as the prime civilizational objective - and that with carefully selected and crafted combination of methods.

However, the farther we find ourselves from the comfort of our studies and enter the labyrinth of fieldwork, the more we encounter distortions, misrepresentations and reductionisms that jeopardize the success of the ethical and responsible AI mission. The novelty and cross-disciplinary complexity of this realm vividly showcases the shortcomings of trying to solve new problems with old tools and mindsets, as well as exposes the unsolved burdens we still carry from previous technological waves. For many stakeholders who by and large use “digital”, “data” and “AI” interchangeably (especially in policy), the specifics of AI still need to sink in: namely its distinctive capabilities of autonomous decision-making, learning capability and the high level of potential opacity (Héder 2020b). On the other end of the spectrum, many technology leaders in charge of AI governance who understand its fundamentals would prefer to reduce the intricacies of the AI Ethics problem set to just data and model bias - and solve it with a dedicated tool that merely checks boxes.

Waking up in an AI world caused us to try to wrap our heads around this set of novel phenomena. These first attempts led to the creation and proliferation of AI Ethics guidelines, numbered in the hundreds of manifestoes by now. While they have been and will be criticized, - sometimes reasonably so -, for being too numerous and obscure, too much overlapping but not too useful, and too self-important, we believe being principled is an unavoidable, highly necessary step - but not the destination. Mostly originating from organizations trying to fill a trust gap in the present state of global governance, one can explain many of their shortcomings to be mainly derived from the factors of a global international order being constantly battered by dissenting incumbent and aspiring hegemons - and consequentially losing significance.

Since many of the studies in this collection come from the CEE region that should be better known for its own, ingrained version of critical thinking, - probably rooted in its grim history overloaded with enforced dogmatism -, one cannot overlook the pattern of wake-up calls that define themselves as "realist"-see for instance Gyulai and Ujlaki (2021) in this issue. In this world becoming multipolar and with authoritarian instincts on the rise, one can even argue a temporary advantage in the realm of AI that benefit those who primarily grasp AI from the perspective of power and dismiss the ethical side as a nice-to-have or just noise. A ruthless AI race mindset permeates both the realms of geopolitics and that of Big Tech: never before was "the winner takes it all" taken so literally. Moreover, as realists are painfully aware, reining in one of the most concerning AI application fields, Autonomous Weapons Systems (AWS) seems distant. Most the UN Security Council Members are opposed to a binding global ban based on New Cold War reflexes, while many smaller states would certainly be in favor as well as the general public. (In one study of this issue, Aron Dombrovszki (2021) is offering a nuanced perspective as the advocate against oversimplified AWS bias.)

China is set out to become a hyperpower built on an AI engine and no stake is too high for them. New Zealand treats data as a tribal heirloom that many generations curate for AI to solely serve citizen well-being. Dubai leaps ahead with its own version of techno-absolutism and deploys vast swaths of government AI services that are measured against the goal of raising citizen happiness levels. The Pope calls for multi-stakeholder global work on human-centric, ethical and responsible AI to preserve Creation. These examples clearly raise the questions: how to be values-based and human-centered in AI with global applicability, while also being mindful of cultural differences, sectoral interests and societal priorities around the globe? How to rethink and redesign our global institutional frameworks and fill them with new meaning to successfully bridge the trustworthiness gap, the most ominous social plague of our civilizations(s)? How to convince Big Tech (equals Big AI) to internalize ethical and responsible AI as a strategic imperative that is key to attract and retain 21st century conscious customers - and not a nice-to-have, borderline ethics-washing parlor trick, a dark possibility Vică, Voinea and Uszkai (2021) elucidate in the current issue?

The emerging AI world – especially in the West – currently has an AI-ready vision and societal model deficit, and that needs remedying first. Our biggest bet is on an informed and responsible society: the emerging class of digital citizens and consumers, professionals and thought leaders who increasingly demand being in charge (Heder 2020a) of their privacy and choices, judge the ethical decision of their employers (and move on if need be), and are ready to stand up for an AI Future that prioritizes human well-being as the *ultima ratio*. Our world could benefit from *AI as augmented intelligence*, a machine-assisted extension of what makes us human -and not the path of artificiality which inevitably dehumanizes. The task of researchers and the AI-savvy is to speak the truth and do their best to make AI understood for conscious citizens worldwide *sans* sensationalism and obfuscation, so that societies could understand what is at stake, what are the new rules, and convert technospeak to challenges and solutions that impact their very lives. This

issue of studies makes a great contribution to this mission of great significance and therefore deserves your kind reading - which you will hopefully find equally profound and enjoyable.

Author Information

Dr. George A. Tilesch, Founding president of the PHI Institute for Augmented Intelligence
Co-Author of BetweenBrains: Taking Back Our AI Future
Member of the Board of Advisors, Experfy – Harvard Innovation Labs
AI Ambassador, John von Neumann Computer Society

References

Dombrovszki Áron. "The Unfounded Bias Against Autonomous Weapons Systems." *Információs Társadalom* 21, no. 2 (2021). https://doi.org/10.22503/inftars.XXI.2021.2.2.

Héder Mihály. "A Criticism of AI Ethics Guidelines." *Információs Társadalom* 20, no. 4 (2020a): 57. https://doi.org/10.22503/inftars.XX.2020.4.5.

Héder Mihály. "The Epistemic Opacity of Autonomous Systems and the Ethical Consequences." *AI & SOCIETY*, (July 30, 2020b). https://doi.org/10.1007/s00146-020-01024-9.

Gyulai Attila, and Ujlaki Anna. "The political AI: a realist account of AI regulation." *Információs Társadalom* 21, no. 2 (2021). https://doi.org/10.22503/inftars.XXI.2021.2.3.

Constantin Vică, Cristina Voinea, and Radu Uszkai. "The emperor is naked: moral diplomacies and the ethics of AI." *Információs Társadalom* 21, no. 2 (2021). https://doi.org/10.22503/inftars.XXI.2021.2.6.

The Unfounded Bias Against Autonomous Weapons Systems

Autonomous Weapons Systems (AWS) have not gained a good reputation in the past. This attitude is odd if we look at the discussion of other – usually highly anticipated – AI-technologies, like autonomous vehicles (AVs); whereby even though these machines evoke very similar ethical issues, philosophers' attitudes towards them are constructive. In this article, I try to prove that there is an unjust bias against AWS because almost every argument against them is effective against AVs too. I start with the definition of "AWS." Then, I arrange my arguments by the Just War Theory (JWT), covering *jus ad bellum*, *jus in bello* and *jus post bellum* problems. Meanwhile, I draw attention to similar problems against other AI-technologies outside the JWT framework. Finally, I address an exception, as addressed by Duncan Purves, Ryan Jenkins and Bradley Strawser, who realized the unjustified double standard, and deliberately tried to construct a special argument which rules out only AWS.

Keywords: *artificial intelligence; autonomous weapons systems; self-driving cars; just war theory; AI ethics; military ethics*

Acknowledgements

This research was supported by the MTA Lendület Values and Science Research Group; and the UNKP-20-3 New National Excellence Program of the Ministry for Innovation and Technology from the source of the National Research, Development and Innovation Fund.

Author Information

Aron Dombrovszki, Eötvös Loránd University (ELTE); MTA Lendület Values and Science Research Group
https://elte.academia.edu/AronDombrovszki

How to cite this article:

Dombrovszki, Aron. "The Unfounded Bias Against Autonomous Weapons Systems." *Információs Társadalom* XXI, no. 2 (2021): 13–28.
https://dx.doi.org/10.22503/inftars.XXI.2021.2.2

Introduction

Autonomous Weapons Systems (AWS) have not gained a good reputation in the past: in 2012, Human Rights Watch suggested a pre-emptive ban, and The Campaign to Stop Killer Robots also raised concerns in the media with an aim to influence public opinion against AWS. The press usually refers to AWS pejoratively as "killer robots," accompanied by a picture of a frightening Terminator-like sci-fi machine (e.g. Kahn 2020; Kessel 2019; Scharre 2020). The academic debate on the topic, however, tends to be more balanced than the general discussion; nevertheless, the majority approach towards AWS is still negative (Rosert and Sauer 2018).

This prejudice differs markedly from the general attitude towards other artificial intelligence (AI) technologies, for example, autonomous vehicles (AVs). AVs are highly anticipated, and even though similar worries can be raised against them, philosophers' attitude is far more accepting.

Purves, Jenkins and Strawser (2015) also mention this kind of a double standard between AWS and other AI-technologies:

> Any account of the permissibility of autonomous weapons systems will risk prohibiting the use of autonomous decision-making technologies that most people view as neutral or morally good. While many of us tend to have a significant moral aversion to the thought of autonomous weapon systems, most have no such similar moral aversion to non-weaponized autonomous systems, such as driverless cars. In fact, for many people, the opposite is true: many of us hold that non-weaponized future autonomous technology holds the potential for great good in the world.

In this article, I aim to prove that Purves, Jenkins and Straswer's observation is accurate, and almost every argument against AWS is effective against AV. To achieve this goal, I start with the definition of "AWS". I then arrange my arguments by the Just War Theory (JWT), covering *jus ad bellum, jus in bello* and *jus post bellum* problems against AWS. Meanwhile, I draw attention to similar problems against other AI-technologies outside the JWT framework. The aforementioned Purves, Jenkins and Strawser article is an interesting exception, because they realized the unjustified double standard, and deliberately tried to construct a special argument which rules out only AWS. I conclude by addressing their arguments and show their vulnerabilities.

1. What are AWS?

Before discussing what are AWS, it is worth noting that the prevalence of misconceptions, often spread by the media, put obstacles in the way of a balanced discussion about AWS. Stigmatized as "killer robots" or referred misleadingly as "lethal autonomous weapons", people may imagine biped androids with guns in their hands. Such terms are misleading since they emphasize the lethal aspect of the system,

which is not a necessary feature of an AWS. For this reason, I prefer to use the term "autonomous weapons systems," as it highlights their autonomous nature, which is more relevant to the topic at hand. In reality, autonomy in weapon systems is a platform-independent functionality: almost every weapon system can be made autonomous, leading to a great variety of devices (Rosert and Sauer 2018). Thus, it is essential to clarify what I mean by "AWS" in this article.

According to a broad definition, an AWS is "a weapon system that, once activated, can select and engage targets without further intervention by a human operator. This includes human-supervised autonomous weapon systems that are designed to allow human operators to override operation of the weapon system, but can select and engage targets without further human input after activation" (U.S. Department of Defense 2012). This definition is a good starting point, but in order to get a more differentiated view – suitable for philosophical discussion – I introduce two distinctions, which hopefully make clear what notion of AWS I am going to discuss herein.

It is necessary here to invoke the distinction from robotics between *general AI* and *modular AI*, as these categories can determine the scope of the relevant issues concerning AWS. Machines equipped with general AI can apply their software to solve, in principle, any problem; essentially, they are general-purpose problem-solvers, like the human brain. Machines like Alphabet's DeepMind are designed to have general AI. By contrast, modular AI is created to excel at a specific task. For example, IBM's Deep Blue is an intelligent chess machine capable of beating human chess-masters, but it cannot be used to accomplish other tasks (Sánchez and Herrero 2018).

Applying this distinction to AWS, a weapon system equipped with general AI could fulfil the public android-like image of a killer-machine, including moving, targeting, and deciding on operations autonomously. For example, the Terminator in the movies has general AI, and as such, these machines would rightfully raise the Hollywoodian fear that they could rebel and turn against humanity (Asaro 2008).

However, I think this worry is ill-founded. Philosophers like Robert Sparrow (2007) take the possibility of these machines seriously, counting on them as intelligent creatures, but the truth is that even if in the distant future AWS with general AI were possible, as far as is known, militaries are not aiming to develop them. Armies anticipating AWS might do so because these machines would lack emotions and other irrelevant considerations during operation, as they would not need skills that are not directly relevant for them to carry out the order given by the commander. Indeed, an AWS with general AI would rather be a drawback than an advantage (Schulzke 2012).

Besides the general categorization of AI, there is a taxonomy of weapons systems concerning their level of autonomy. Marra and McNeil (2013) introduced the distinction between three different kinds of weapons systems by the requirement of the human operator in the decision process. They refer to this process as the "loop". First, those systems in which there is a human "in the loop", i.e. an operator to execute at least one of the processes that is needed; for example, to launch a bomb from an airplane. Second, a system where a human operator is "on the loop", meaning that the system is capable of executing all of the tasks alone but with a human supervisor having veto power; for example, some current on-the-loop systems are the

Phalanx CIWS and the IAI Harop. Third, weapon systems that are fully autonomous, where all human operators are "out-of-the-loop". Throughout this paper, the term "AWS" will refer to out-of-the-loop systems, which are yet to be developed.

2. Jus ad bellum

In JWT, *jus ad bellum* contains a set of principles, prescribing the just use of force and legitimate reasons for going to war. According to Walzer's original account, only self-defensive wars can have a just cause, such as to protect the sovereignty of the state, innocent life and basic human rights (Walzer 1977). Humanitarian intervention – for example, to prevent a genocide – might also be considered as a legitimate initiative. Once a state has entered a war, its ultimate aim has to be the restoration of peace. While the belligerents[1] carry out this aim, the soldiers, as well as their commanders, have to have the right intention. This last rule, even though it is debatable, will be crucial as the source of one central objection to AWS that I discuss at the end of this paper.

2.1. The lower threshold to engage war

The main objection against the development of AWS ad bellum is that their deployment will lead to radical asymmetry in warfare: a military with AWS gains a significant advantage over potential enemies. It is then more likely that such advantaged armies will engage in wars, and hence there will be more – probably unjust – wars in the future. This willingness to go to war can have many reasons, the lower costs of warfare, the reduced risk of losing human lives, and, of course, the higher chance to overcome an enemy (Johansson 2011).

Earlier, Bradley Jay Strawser defended uninhabited aerial vehicle (UAV) technologies against this objection, and I think his considerations can also be used to defend AWS. According to Strawser (2010, 359), "the scope of this issue far exceeds UAVs [...], of course, but strikes at any asymmetry in military-technological development whatsoever." Recall history's great military-technological revolutions: the discovery of gunpowder and the cannon in the 15th century, the steel and steam revolution in the 19th century, and the appearance of nuclear weapons in the mid-20th century (Sánchez and Herrero 2018; Krause 1992). All of these changes resulted in temporary radical asymmetry between belligerents, which could easily cause a series of unnecessary and unjust wars.

If we think that this objection stands, and asymmetric warfare almost always caused unjust wars in the past, which they will also cause in the future, then it is reasonable to extend the objection. Those who reject AWS on this ground should argue that states have to stop developing *any* kinds of new weapons and military technologies – including defensive ones – except for those explicitly developed for

[1] A belligerent is a nation or person engaged in war or conflict, as recognized by international law.

medical purposes. So, UAVs or AWS are not unique in this respect, and one should either argue against all military technologies or anticipate with a neutral or positive outlook all new developments, including AWS.

3. Jus in bello

JWT is more than an ethical theory, as its jus in bello principles are the foundations of International Humanitarian Law. It has two main principles that are essential for the present discussion. The first one is the principle of discrimination: civilians and persons who are *hors de combat* should always be protected by making a clear distinction between combatants and non-combatants. The former group can be legitimately targeted and engaged by enemies using lethal force, while the latter have immunity until they engage in hostile activities. The second principle is the principle of proportionality: while a military accomplishes an objective, it has to resort only to the force necessary in order to limit unnecessary suffering in war.

3.1. Malfunctioning, hacking, stealing

Every machine can potentially be exposed to malfunctioning, stealing, hacking or other security issues during operation. I will treat these different threats as one because their consequences are similar. While developers are usually willing to admit that nothing is flawless, when it comes to AWS, the potential damage can be extensive. Also, even carefully written software cannot prevent the stealing of the physical device (Lucas 2013).

I want to address these issues by pointing out that these risks are present quite generally. Such accidental harm could also affect automated medical diagnosis systems, smart cities, Internet of Things (IoT) devices[2] (…), or autonomous vehicles (AV), but few would argue that by virtue of these dangers the development of these useful devices should be suspended.

First, note that AWS would operate in militarized zones only; hence, it is highly probable that they would not endanger civilians. Second, despite the theoretical possibility, AWS would not be equipped with weapons of mass destruction. These considerations point towards the conclusion that if a problem were to occur, the AWS would cause damage only to the combatants in a battlefield.

Without clear benefits[3], this damage would also be unacceptable, but compared with the other technologies mentioned above, one might say that AWS is in a better position, as a failure in an AV or in a highly complex smart city might also threat-

[2] This includes a wide variety of things, including "a person with a heart monitor implant, a farm animal with a biochip transponder, an automobile that has built-in sensors to alert the driver when the tyre pressure is low, or any other natural or man-made object that can be assigned an IP address and is able to transfer data over a network" (Rouse 2020).

[3] As Strawser (2010) points out, AWS have several benefits for the militaries and can actually prevent the death of human combatants.

en the lives of civilians. A malfunctioning AV can kill innocent people, but this fact rarely leads to arguments against such vehicles. Philosophers instead try to settle the issue of responsibility after the accident has happened.

One could argue that the comparison between AWS and AVs is misleading because the former can cause more significant harm, while an AV would kill fewer people. Note that this utilitarian consideration misses the possible scenario where the car hits a crowd, injuring possibly dozens of people. However, more importantly, it ignores the AV's planned integration with smart cities, which guarantees vulnerabilities, and new, unforeseeable threats.

Every new technology brings up new security issues. However, the trade-off between the anticipated risks and benefits is a more important question. AVs promise to make everyday life so much easier that most of us are willing to view the risks as part of an acceptable trade-off. What has not been acknowledged yet are the similarly significant benefits of AWS for militaries during armed conflicts.

3.2. Discrimination and proportionality

Besides accidents and contingent issues, like malfunctioning or hacking, Human Rights Watch have argued that AWS could not observe JWT's two central *in bello* principles (Human Rights Watch 2012), and they proposed a pre-emptive ban on such machines.

According to their report, AWS will lack the sensory, computational and interpretative abilities to distinguish between combatants and civilians, which is sometimes also a difficult task for a human being. This is because even if a system can distinguish armed soldiers in uniforms from regular citizens, there are certain circumstances where this is not enough. AWS have to recognize illegitimate targets; for example, persons who are *hors de combat*, surrenderers, guerrilla soldiers – who do not wear uniforms and attack in unusual fashion and places – and unusual targets, like civilians, who directly participate in hostilities, and therefore have lost their immunity. Applying the principle of discrimination can be extremely difficult in these unusual situations.

The principle of proportionality has raised similar objections. Human Rights Watch has expressed doubts that AWS will be able to evaluate the exact measures of necessary force in such complex environments as a modern battlefield. They emphasize the necessarily subjective and context-dependent nature of the skills that are needed to observe the principle of proportionality. So, according to Human Rights Watch, AWS theoretically could not have the capabilities to measure what is the proportionate use of force, and therefore, what is legally allowed under the Law of Armed Conflicts.

Even though AVs and other AI-machines have similar capabilities as AWS, a different approach towards them can be spotted in the literature. It seems that scholars have very different expectations regarding AVs: they assume that these machines will reduce reaction time and will also overcome the weaknesses in human judgment. Since human error causes over 90% of traffic accidents, a significant increase in road safety could be expected (Friedrich 2016).

This optimism is strange as I think it is not easier to write an algorithm that can navigate in an urban environment, in particular, considering that human drivers will still be on roads, than to develop an AI system that is able to appropriately distinguish between combatants and non-combatants. Traffic in a metropolis is highly dynamic, with many unwritten rules and unexpected situations, where human judgment seems indispensable. For example, if there is a police traffic control in place instead of traffic lights, AVs have to be able to recognize the signs given by the officer. Also, human drivers often do not drive precisely by the rules, so AVs have to take into consideration the rule-breakings of the drivers too. It is also to be noted that strictly observing the rules in traffic sometimes can be inefficient and annoying. Can AVs decide that breaking the speed limit by 10% is appropriate or not? Besides, pedestrians might also appear at the most unexpected places where they should not, and the software has to handle these situations. Road constructions, small roads without signs, and the signs of other agents in traffic also have to be considered.

These challenges do not seem more straightforward than to decide who is a civilian and who is a uniformed, armed combatant.[4] Various strategies already exist to enhance the recognition of legitimate targets, e.g. transponders, behaviour analysis, location- and time-limited actions, or multi-sensor analysis. Combining these already existing solutions can offer considerable protection to civilians (Hughes 2014). With respect to the proportionality issue, the chances are bigger for AI to be better than humans. Also, evaluating a proportionate attack is a computational task, where machines abilities are better than humans.

Nevertheless, I admit that these are just speculations. It would be an empirical question whether the introduction of AVs to the roads will lead to increased safety or not. The very same is true though for AWS: their use could reduce unnecessary harm and death in just wars, or could lead to the opposite. What is important to see is the unjustified double standard in favour of AVs and against AWS.

3.3. An allegedly AWS-specific issue presented by Purves, Jenkins and Strawser

All the previous objections have the problem that they are too broad, and accepting them leads to the unacceptability of developing automated technologies across a broad spectrum. Purves, Jenkins and Strawser (2015) recognized this issue, and they tried to construct an argument targeting AWS specifically. Their point is that applying AWS would be similar to deploying psychopaths in the battlefield. They based this claim on the assumption that the decisions of the AWS

[4] I do not deny that there is an epistemological difference between friendly and hostile environments (Sterelny 2003). Participants of a friendly situation are aimed at unambiguity and directness, while the uncertainty of signals, unpredictability, mimicry, and camouflage are common elements of strategy in a battlefield. However, this circumstance does not affect the relevant distinction of combatants and non-combatants as everyone is interested in the clear-cut differentiation in this particular case.

cannot be made for the right reasons.[5] Hereafter, I examine this argument in more detail.

Purves, Jenkins and Strawser's strategy relies on a principle in JWT: "there is a positive requirement to act for the right reasons in deciding matters of life and death" (2015). *Ad bellum*, it is not enough to have a just cause to go to war, but the belligerents have to act correctly for this very reason (see Section 2). AWS do not fall under this rule because they do not have the right to enter into wars: politicians and military advisors play the decisive role in these matters.

Purves, Jenkins and Strawser (2015) were not satisfied with this narrow requirement. The authors claim that it can be extended to the whole time of the war, creating an *in bello* obligation for the soldiers on the battlefield to act for the right reasons in their every act. Even though this is not part of the traditional JWT, they refer to authorities who support this idea, like Thomas Nagel (1972) and Peter Asaro (2012).

AWS are, by their very design, not able to act for the right reasons, so they cannot be applied in a just war. To illustrate their point, the authors draw a scenario with a sociopathic soldier, taken to be similar to the application of an AWS:

> Imagine a sociopath who is completely unmoved by the harm he causes to other people. He is not a sadist; he does not derive pleasure from harming others. He simply does not take the fact that an act would harm someone as a reason against performing the act. In other words, he is incapable of acting for moral reasons. It then comes about that the nation-state of which this man is a citizen has a just cause for war: they are defending themselves from invasion by an aggressive, neighbouring state. It so happens that the man joins the army (perhaps due to a love of following orders) and eagerly goes to war, where he proceeds to kill scores of enemy soldiers without any recognition that their suffering is morally bad. He is effective precisely because he is unmoved by the harm that he causes and because he is good at following direct orders. Assume that he abides by the classic jus in bello rules of combatant distinction and proportionality, yet not for moral reasons. No, the sociopathic soldier is able to operate effectively in combat precisely because of his inability to act for moral reasons.

Purves, Jenkins and Strawser (2015) conclude that anyone who thinks that it is problematic that a sociopath soldier would be involved in waging war, has to accept that applying AWS would be just as problematic.

Even though the authors tried to present an argument that is specific to AWS, one can easily make an analogy with AVs in this case: these machines sometimes also have to make decisions about matters of life and death. Consider the widely discussed trolley-type dilemmas arising in the literature (Lin 2016). Purves, Jenkins and Strawser (2015) are aware of this and present a twofold answer against these doubts.

[5] Moreover, the supposedly unfulfilled requirement of acting for the right reasons – or the lack of moral reasoning in general – can be the basis of other deontic objections against AWS; for example, the problem of human dignity (Sharkey 2019) or respect (Skerker et al. 2020). In this paper, I restrict myself only to analyzing the arguments presented in Purves, Jenkins and Strawser's (2015) work.

On the one hand, their response takes into consideration the *raison d'être* of these machines. AVs are created for peaceful purposes in order to make traffic safer, and despite the fact that they sometimes have to make decisions with potentially lethal outcomes, these are not part of their everyday operation. On the other hand, they retort that there is a distinction between the frequencies of the decision-making: while AWS will continuously make lethal decisions as they are supposed to do, AVs will only do so on rare occasions. They consider these two arguments together satisfying enough to differentiate between weaponized and non-weaponized autonomous systems.

It is not easy to argue against Purves and his colleagues' position because they keep it smooth and sophisticated without being too ambitious. At the end of their paper, they position their stance with many qualifications:

> Even if the responses fail to maintain a hard moral distinction between weaponized and non-weaponized AWS, however, we are not ultimately concerned about our argument ruling out driverless cars and other autonomous systems. We ought to meet a high bar before deploying artificial intelligences of any kind that could make morally serious decisions—especially those concerning life and death. It is plausible that no autonomous system could meet this bar (Purves, Jenkins and Strawser 2015).

I agree that we have to be extremely careful before we start to use AVs or wage wars with AWS. However, I maintain that those authors failed to provide a persuasive moral distinction between the two technologies. In the following, I raise three points that do not necessarily falsify their arguments but weaken them considerably.

First, I would like to point out that the conjunction they used to underpin the difference between AVs and AWS contains two different types of conjuncts. The second one – that AVs will make fewer lethal decisions than AWS – is a quantitative argument; however, I think it misses the point. Note that, according to JWT, combatants are legitimate targets, so only the accidental non-combatant killings should count in this comparison. This fact considerably changes the intuitive appeal of the argument, because the difference in this respect is slight. I acknowledge that AWS are still in a worse position, but the boundaries are vague. For this reason, it is difficult to build solid grounds for this objection without telling how many lethal decisions are acceptable in an autonomous machine's life. This task is still ahead of the authors.

The first conjunct – which points out the reason why people make a machine – is a qualitative difference between AVs and AWS, and it is more interesting as I think this embodies the real reason why people are so hostile towards AWS technologies. We do not like things that are designed to take someone's life and, in addition, war and weapons have extremely negative connotations in western culture. These are sometimes legitimate concerns, but given the correct understanding of JWT, most of them are questionable. In a just war, the defensive state has the right to use weapons

no matter what technologies are involved.[6] However, in an unjust war, it does not matter if we use guns, drones or AWS, our actions will not be legitimate. Moreover, in JWT, as one can argue, the purpose of deploying AWS in the battlefield is to protect innocent civilians and to achieve peace faster and with reduced loss, which seem to be desirable goals.

Furthermore, I would like to add that the military chain of command creates special circumstances in moral responsibility, and endows agents with various kinds of ethical status. Due to this fact, we can expect very different things from commanders and soldiers. To highlight the importance of right intention, the authors use Jaworska and Tannenbaum's (2014, 245) example from everyday life:

Consider, first, giving flowers to Mary only in order to cheer her up, as opposed to doing so merely to make Mary's boyfriend jealous. Although the two actions are alike in one respect—both involve giving Mary a gift— the different ends make for a difference in the actions' nature and value. Only the former is acting generously, while the latter is acting spitefully. In one sense, the intended end is extrinsic to the action: one can have and intend an end independently of, and prior to, performing the action, and the action can be described without any reference to the intended end. And yet something extrinsic to an act can nevertheless transform the act from merely giving flowers into the realization of acting generously (or spitefully), which has a distinctive value (or disvalue).

It would be unnecessary to deny that the intentions have a crucial role in our *everyday* moral – and also legal – judgment, but in the military, commanders have the responsibility – and the burden of punishment – instead of their soldiers due to the chain of command (Schulzke 2012). For this reason, it is questionable whether a soldier on the battlefield has to have the right intention in order to morally justify his or actions. This only applies in a special version of JWT, and it is plausible to suppose that it is enough if the commanders give their orders with the right intention.

Finally, I would like to point out that the objection of Purves, Jenkins and Strawser (2015) seems to disregard the intentions of the developers of these machines. We should not be surprised by their analogy between a sociopathic soldier and AWS, because this similarity is intended. As I mentioned in Section 1, military robots will only have modular AI, while lacking moral sense, in order to follow directions precisely. This is how militaries want them to be, so the problematized similarity with a sociopath soldier is not a bug, but a feature.

4. Jus post bellum

Jus post bellum principles aim for a trouble-free transition from war to peace. A significant part of this progress is the accountability of potential war criminals in

[6] Weapons that are below the legally required line of distinction and proportionality, or considered unethical – e.g. weapons of mass destruction, landmines, blinding lasers, expanding bullets – are exceptions. Keep in mind that certain armaments can be made autonomous, but AWS are not weapons themselves (see Section 1).

international courts. Therefore, it is an essential requirement in JWT that every belligerent has to be a liable moral agent. AWS seem to challenge this principle.

4.1. Responsibility gap

According to Robert Sparrow (2007), the main challenge in JWT posed by the deployment of AWS is the so-called responsibility gap issue (cf. Matthias 2004). Suppose that – for some reason – an AWS made a mistake and destroyed a village with civilians only. Note that AWS – unlike a remotely controlled uninhabited aerial vehicle – is targeted and engaged automatically without a human in the loop. Who is responsible for this war crime?[7]

Several possible candidates can bear the responsibility: the most obvious is the AWS itself. Nevertheless, it seems that AWS is just not the right type of entity to bear moral responsibility. It lacks the general AI that would be needed to comprehend the consequences of its actions. Moreover, an AWS cannot be punished in a meaningful way (Sparrow 2007).

Apart from the machine, the programmer might also be liable (Kuflik 1999). This suggestion seems more plausible than the previous one, but still faces serious difficulties. I will mention three of them. First, the codes used in AWS and other highly independent automata are learning algorithms, so the output of a given input is unknown, even to the programmer. Second, programmers usually work in teams, so it would be difficult to identify the one person who wrote the code that led to the tragedy. Additionally, the individual members of the team rarely understand how the full software works in practice. Finally, due to the modular architecture of the system, situations may occur in which developers buy the software from another project to use it in the AWS. For instance, software developed for AVs can be used to automatize infantry fighting vehicles. In this case, the programmers may not know that their codes were used in an AWS; therefore, it would be unjust to blame them.

Finally, one can blame the operation commander who ordered the deployment of the AWS to the battlefield (Lazarski 2002). *Prima facie*, this seems a fair solution, but a similar issue to the previous objection can be raised here. The commander may even have less knowledge than the programmers about how an AWS would respond in different situations. Therefore, it would not be fair to punish him or her for the malfunction of the machine.

Those who object based on the responsibility gap issue argue that – mostly because of the reasons introduced above – nobody can be blamed for the wrongdoings of an AWS. So, according to the *jus post bellum* principles in JWT, their deployment is morally wrong (Sparrow 2007).

Some philosophers think that the existence of the responsibility gap is the ultimate objection against AWS. However, regardless of what improvements can be

[7] Nevertheless, not everyone acknowledges the existence of the so-called responsibility gap issue posed by autonomous technologies (Tigard 2020), but I do not aim to discuss this issue any further in this article. Instead, I focus on the arguments and the debate between those who agree that this is a severe problem.

delivered by AVs and how safe they will be, it is highly probable that they will also make mistakes, even lethal ones – granted though that the number of these events will be insignificant. But, when it happens, the very same responsibility gap is present. According to the literature on the subject, a solution to this issue is not easy.

Parallels with AWS can be made concerning the possible candidates who should bear the responsibility in the case of an accident. These are the vehicle itself, the manufacturer (Gurney 2013) or the owner of the car (Hevelke and Nida-Rümelin 2015). The first option is usually not even considered due to the lack of agency – similarly to the case of AWS. However, the debate between the last two options is lively, because both positions have strong arguments in their favour.

The critical point for the present discussion is the lack of consensus about the responsibility gap issue in the ethics of AVs as well as AWS and the different conclusions that it usually evokes. In the former area of research, scholars appear hopeful about solving the issue, while in the latter case, usually the contrary conclusion is drawn, namely that AWS should be banned as the responsibility gap issue is unresolvable. This double standard, as I have shown, is untenable. We are therefore left without AWS-specific arguments in favour of banning AWS.

5. Summary

Before I summarize this paper's conclusions, I would like to address a general objection against the framework applied through the whole investigation. Some will argue: I should have taken into account that not every war is a just war, or at least it may be difficult to argue beyond any reasonable doubt that it is just. One should not argue for or against AWS on an abstract, idealistic theoretical ground. In light of this, the approach of this article is naive, unrealistic or even unethical. I am aware of this issue, but I would like to add three considerations that may weaken its strength.

First, this kind of objection seems to support my thesis about the unbalanced approach towards the topic. When one starts to read the literature about AWS, one rarely sees these arguments explicitly discussed, neither pro nor contra. This does not mean that concrete, contextual arguments are invalid or unintuitive, just that they are usually not frequently mentioned problems against AWS. In fact, most scholars who oppose AWS construct their arguments in an abstract theoretical space, similar to the one I have used here. This is why I think that it is rational to examine the discussion in an allegedly naive or idealistic approach. I would like to emphasize that, apparently, no one is bothered by this methodology as long as it supports the arguments against AWS, but when the very same approach leads to the contrary, somehow the framework becomes "idealistic".

Second, it is worth reconsidering the goals of developing an ethical theory or policy-making. Generally speaking, in these investigations, one aims to create certain imperatives that people should follow to act virtuously or at least lawfully. Anyone who has the ambition to work out normative theories should suppose that people will follow the rules voluntarily, or the state will have the necessary resource to

force its citizens to follow them. For example, it would be a strange line of thought to legalize thievery because we live in a world where people often break private property laws. In the same way, it is questionable to argue against the JWT framework on the basis that there have been numerous unjust wars in history. In normative projects like considering the ethical status of AWS, we do not aim to describe the facts of the world; instead, we propose certain principles or rules presupposing that these will be observed.

Third, to argue for the ban of AWS on the basis that states are not going to observe the relevant regulations during the deployment of the machines is in a way self-defeating. What guarantees that any law that prohibits AWS will be followed? I think nobody can warrant it, and there are cases indeed when outlawed weapons of mass destruction – like sarin or the VX nerve agent – were used despite their international ban (Murphy 2013; Zurer 1998). Nevertheless, no one would use these unfortunate incidents as evidence that the ban was a wrong decision. Similarly, the potential threat of disobeying the laws and regulations on the use of AWS can hardly be an argument against their deployment. Recalling the previous point, in normative investigations like the discussion of AWS, we have to presuppose that people are generally rule-following; otherwise, all our efforts will be futile.

Despite the above considerations, this paper aimed not to argue in favour of AWS, only to provide a meta-analysis of the debate by pointing out specific biases. By being aware of these partialities, scholars can develop better arguments for or against AWS in the future. Throughout this article, I examined five objections against the deployment of AWS and tried to show that most objections are general to AI-technologies altogether, and Purves, Jenkins and Strawser's deliberately specific argument faces problems. Therefore, without any morally relevant distinction, we should either anticipate AWS with other potentially lethal AI-technologies or rule out all of them.

The first objection was the worry that a military that owns AWS will likely go to war more frequently, because of the reduced costs, and its guaranteed advantage over other militaries. I argued that this applies to every improvement in military technology. So, we either accept that every new development should be banned or accept AWS as just another step in the evolution of weapons systems.

According to the second, contingent objection, AWS will not be able to discriminate between combatants and non-combatants properly and will lack the capabilities to measure the proportionate means of attack accurately. I argued that parallel doubts against AVs could be raised – but in fact, are rarely addressed in the philosophical literature. Therefore, we either ban the development of AWS along with AVs or accept the fact that these technologies can potentially overcome human weaknesses, and thus we should anticipate their development.

The third objection was the worry that AWS could be stolen, could malfunction or could be hacked and these outcomes could lead to disastrous events. I argued that AVs and their supporting technologies, like smart cities or power plants connected to the network, are similarly vulnerable in this respect; in fact, as civilian systems – where security is not always the priority –, perhaps even more so. A smart city has as much vulnerability as an AWS and the consequences of a possible attack or malfunction is also catastrophic.

Perhaps the most challenging issue concerning AI-technologies is the responsibility gap: from intelligent elevator systems through AVs to AWS, many technologies are affected by this (Matthias 2004). However, depending on the type of technology in question, ethicists assess the problem differently. In the field of war ethics, the responsibility gap is usually an ultimate reason to ban AWS in the future. But when it comes to AV, the attitude of philosophers is markedly different: they try to resolve the problem in order to remove the barrier to AV application. I argued that this double standard is mistaken because such an objection rules out a broad spectrum of AI-technologies, AVs among them.

Finally, Purves, Jenkins and Strawser's objection was created specifically against AWS, but its success is debatable. I called attention to two points regarding their insights. First, the military chain of command creates a special context concerning responsibility attribution, so the major purpose of creating AWS is to eliminate unnecessary human emotions and intentions, but the authors have not taken into consideration this fact.

Those who would like to argue against the deployment of AWS have to emphasize its distinguishing characteristic that other AI-technologies or weapons lack. This characteristic can be the basis of a forthcoming argument against them. However, in most of the objections, this characteristic is omitted, which makes the argument too broad to be effective. Purves, Jenkins and Strawser (2015) point out this mistake and attempt to create a specific objection. They succeeded in outlining the distinguishing characteristic – the lack of right intention – but their argument can be challenged because right intention *in bello* for soldiers is not necessary to wage a just war – this requirement only applies to politicians and military leaders *ad bellum*.

References

Asaro, Peter M. "How Just a Robot War Could Be?" In *Current Issues in Computing and Philosophy*, edited by Adam Briggle, Katinka Waelbers, and Philip Brey, 50–64. Amsterdam: IOS Press, 2008.

Asaro, Peter M. "On Banning Autonomous Weapon Systems: Human Rights, Automation, and the Dehumanization of Lethal Decision-making." *International Review* of the Red Cross 94, no. 886 (2012): 687–709. https://doi.org/10.1017/S1816383112000768

Friedrich, Bernhard. "The Effect of Autonomous Vehicles on Traffic." In *Autonomous Driving*, edited by Barbara Lenz, Markus Mauer, and J Christian Gerdes, 317–34. Berlin: Springer, 2016.

Hevelke, Alexander, and Julian Nida-Rümelin. "Responsibility for Crashes of Autonomous Vehicles: An Ethical Analysis." *Science and Engineering Ethics* 21, no. 3 (2015): 619–630. https://doi.org/10.1007/s11948-014-9565-5

Hughes, Joshua. "Could Autonomous Weapons Systems Be Used Legally Under the Law of Armed Conflict?" Last modified 2014. https://www.academia.edu/8193381/Could_autonomous_weapons_systems_be_used_legally_under_the_Law_of_Armed_Conflict

Human Rights Watch. *Losing Humanity: The Case against Killer Robots*. New York: *Human Rights Watch and International Human Rights Clinic,* 2012.

Jaworska, Agnieszka, and Julie Tannenbaum. "Person-rearing relationships as a key to higher moral status." *Ethics* 124, no. 2 (2014): 242–271. https://doi.org/10.1086/673431

Johansson, Linda. "Is it morally right to use unmanned aerial vehicles (UAVs) in war?" *Philosophy & Technology* 24, no. 3 (2011): 279–291. https://doi.org/10.1007/s13347-011-0033-8

Kahn, Jeremy. "Air Force A.I. Test Raises Concerns Over Killer Robots." *Fortune.* Last modified: December 2020. https://fortune.com/2020/12/21/killer-robots-ai-us-air-force-experiment-u2-spy-plane-artumu/

Kessel, Jonah M. "Killer Robots Aren't Regulated. Yet." *The New York Times.* Last modified: December 2019. https://www.nytimes.com/2019/12/13/technology/autonomous-weapons-video.html

Kuflik, Arthur. "Computers in control: Rational transfer of authority or irresponsible abdication of autonomy?" *Ethics and Information Technology* 1, no. 3 (1999): 173–184. https://doi.org/10.1023/A:1010087500508

Lazarski, Anthony J. "Legal Implications of the Uninhabited Combat Aerial Vehicle." *Aerospace Power Journal* 16, no. 2 (2002): 74–83.

Lin, Patrick. "Why Ethics Matters for Autonomous Cars." In *Autonomous Driving*, edited by Barbara Lenz, Markus Mauer, and J Christian Gerdes, 69–85. Berlin: Springer, 2016.

Lucas, George Jr. "Engineering, Ethics & Industry: The Moral Challenges of Lethal Autonomy." In *Killing by Remote Control: The Ethics of an Unmanned Military*, edited by Bradley-Jay Strawser, 221–29. New York: Oxford University Press, 2013.

Marra, William, and Sonia McNeil. "Understanding 'The Loop': Regulating the Next Generation of War Machines." *Harvard Journal of Law and Public Policy* 36, no. 3 (2013): 1139–1185. https://dx.doi.org/10.2139/ssrn.2043131

Matthias, Andreas. "The Responsibility Gap: Ascribing Responsibility for the Actions of Learning Automata." *Ethics and Information Technology* 6 (2004): 175–183. https://doi.org/10.1007/s10676-004-3422-1

Murphy, Joe. "Cameron: British scientists have proof deadly sarin gas was used in chemical weapons attack." *Independent.* Last modified: September 2013. https://www.independent.co.uk/news/world/middle-east/cameron-british-scientists-have-proof-deadly-sarin-gas-was-used-chemical-weapons-attack-8800528.html

Nagel, Thomas. "War and Massacre." *Philosophy of Public Affairs* 1, no. 2 (1972): 123–144. https://doi.org/10.1177/000276427201500678

Purves, Duncan, Ryan Jenkins, and Bradley J. Strawser. "Autonomous Machines, Moral Judgement, and Acting for the Right Reasons." *Ethical Theory and Moral Practice* 18, no. 4 (2015): 851–872. https://doi.org/10.1007/s10677-015-9563-y

Rosert, Elvira, and Frank Sauer. “Perspectives for Regulating Lethal Autonomous Weapons at the CCW: A Comparative Analysis of Blinding Lasers, Landmines, and LAWS.” Last modified: 2018. https://www.academia.edu/36768452/Perspectives_for_Regulating_Lethal_Autonomous_WeapoWe_at_the_CCW_A_Comparative_Analysis_of_Blinding_Lasers_Landmines_and_LAWS

Rouse, Margaret. “Internet of Things.” *IoT Agenda.* Last modified: February 2020. https://internetofthingsagenda.techtarget.com/definition/Internet-of-Things-IoT

Sánchez-Herrero, Virginia Romero. “The Ethics of Strategic Artificial Intelligence: An Assessment of Autonomous Weapons Systems through the Just War Tradition.” Last modified 2018. https://www.academia.edu/36382896/The_Ethics_of_Strategic_Artificial_Intelligence_An_Assessment_of_Autonomous_Weapons_Systems_through_the_Just_War_Tradition

Scharre, Paul. “Are AI-Powered Killer Robots Inevitable?” *Wired.* Last modified: May 2020. https://www.wired.com/story/artificial-intelligence-military-robots/

Schulzke, Marcus. “Autonomous Weapons and Distributed Responsibility.” *Philosophy & Technology* 26, no. 2 (2012): 203–219. https://doi.org/10.1007/s13347-012-0089-0

Sharkey, Amanda. “Autonomous Weapons Systems, Killer Robots and Human Dignity.” *Ethics and Information Technology* 21 (2019): 75–87. https://doi.org/10.1007/s10676-018-9494-0

Skerker, Michael, Duncan Purves, and Ryan Jenkins. “Autonomous Weapons Systems and the Moral Equality of Combatants.” Ethics and Information Technology 22 (2020): 197–209. https://doi.org/10.1007/s10676-020-09528-0

Sparrow, Robert. “Killer Robots.” *Journal of Applied Philosophy* 24, no. 1 (2007): 62–77. https://doi.org/10.1111/j.1468-5930.2007.00346.x

Sterelny, Kim. *Thought in a Hostile World: The Evolution of Human Cognition.* Oxford: Blackwell, 2003.

Strawser, Bradley Jay. “Moral Predators: The Duty to Employ Uninhabited Aerial Vehicles.” *Journal of Military Ethics* 9, no. 4 (2010): 342–368. https://doi.org/10.1080/15027570.2010.536403

Tigard, Daniel W. “There Is No Techno-Responsibility Gap.” *Philosophy & Technology* (2020). https://doi.org/10.1007/s13347-020-00414-7

U.S. Department of Defense. DIRECTIVE NUMBER 3000.09, November 21, 2012.

Walzer, Michael. *Just and Unjust Wars.* New York: Basic Books, 1977.

Zurer, Pamela. “Japanese cult used VX to slay member.” *Chemical & Engineering News* 76, no. 35 (1998): 7. https://doi.org/10.1021/cen-v076n035.p007

The political AI: A realist account of AI regulation

This article adopts a political theoretical perspective to address the problem of AI regulation. By disregarding the political problem of enforceability, it is argued that the applied ethics approach dominant in the discussions on AI regulation is incomplete. Applying realist political theory, the article demonstrates how prescriptive accounts of the development, use, and functioning of AI are necessarily political. First, the political nature of the problem is investigated by focusing on the use of AI in politics on the one hand and the political nature of the AI regulation problem on the other. Second, the article claims that by revisiting some of the oldest political and theoretical questions, the discourse on guidelines and regulation can be enriched through the adoption of AGI and superintelligence as tools for political theoretical inquiry.

Keywords: *Artificial Intelligence, Political Theory, Political Realism, Applied Ethics, Enforceability*

Author Information

Attila Gyulai, Centre for Social Sciences - Institute for Political Science / University of Public Service
https://orcid.org/0000-0003-2471-6049

Anna Ujlaki, Centre for Social Sciences – Institute for Political Science / Corvinus University of Budapest
https://orcid.org/0000-0002-8030-624X

How to cite this article:
Gyulai, Attila, Anna Ujlaki. "The political AI: A realist account of AI regulation."
Információs Társadalom XXI, no. 2 (2021): 29–42.
https://dx.doi.org/10.22503/inftars.XXI.2021.2.3

1. Introduction

With the widespread emergence of artificial intelligence in the modern world, decision-makers have become increasingly aware that the development, use and functioning of AI require some level of regulation. From AI companies and research institutions to governments and supranational organizations, several actors in the field recognized that despite the proliferation of self-imposed value-systems and guidelines (Hagendorff 2020; Héder 2020), and even if the value-alignment and value-loading problems (Christian 2020; Bostrom 2014) become solved, the question of how the desired values and norms could be realized remains open. Although some researchers think that the race for AI will result in a race for regulation (Smuha 2021), or that the ethical frameworks and the users can be brought closer to each other (Hatamleh and Tilesch 2020), thus allowing the guidance problem to be settled by taking the values and norms into account on the one hand and the developers and users on the other, we argue that a third factor, namely politics should be included in the discussion.

In this article, we argue that the problem of implementing AI guidelines, whether they concern narrow AI, expert systems or superintelligence, will be necessarily focused on the question of enforceability. That is, whereas the discussion on guidelines concerns mostly their value content and the norms they are aimed at applying, our claim is that the issue of how they can become binding should be taken into account more seriously. By its nature, it is a political problem, inasmuch as it is not merely political on a thematic level of how AI is used within our political institutions but in its very logic that is defined by putting forward values and seeking their realization.

To understand the practitioner's view regarding values, ethical guidelines and their realization requires taking the political nature of the context of the complex field of AI more seriously into account. Our claim is that political theory, precisely because it is focused on the links between norms and action, prescription and realization and justification and enforceability can help with understanding a further level of the emerging AI problem (Damnjanovic 2015, 76; Schippers 2020, 35). In other words, to understand on a practical level how to regulate AI one needs to consider the political context of regulation.

Therefore, the article aims to sketch a twofold problem that emerges at the intersection of AI and political theory. First, we demonstrate that the implications of approaching AI from a political theoretical perspective require that AI should be considered with a particular focus on the special nature of the political sphere. Here, we challenge the 'applied ethics' approach of those authors who regard AI as a mere regulatory problem. We claim that in discussing the regulation of weak AI, the context of political action and the specific conflictual nature of politics must always be taken into consideration either on the level of the emergence of AI within political practice or in the more abstract understanding of the general political context of AI. We address this as a problem of political realism: building on an analogy, we claim that just as the current mainstream in AI research conceives of guidelines as 'applied ethics', problems of relevance and efficacy should be reconsidered, just as realist political theory did to the relationship between normativity and political

action. Second, we claim that the implications of AI for political theory could be beneficial for addressing some of the oldest political theoretical questions; for example, issues of peaceful coexistence, sovereignty or authority. Here, we claim that AGI and superintelligence are relevant tools for political theoretical inquiry.

The authors of this article are political theorists with no formal training in AI. But we are enthusiastic about both the nature of the political world and the possibility of political theorizing on the practical guidance of human conduct. We believe that our political theoretical contribution may help AI experts to understand the inherent limitations of regulating AI. At the same time, we see the current interest in AI as a great opportunity to understand the ordinary operation of the political world.

2. The realist approach

We take political realism as a framework for discussing the political context of AI guidelines. In this section of the paper, we briefly define political realism and highlight what we believe is an intrinsic link between the main concern of realist political theory and contemporary efforts to prescribe the development and functioning of AI.

Realism has always been a main general understanding of politics, even if during the latter part of the 20th century it became predominantly known as a paradigm in IR where it was described as politics focusing on power, interest or the use of force (Morgenthau 2005). Whereas these terms are not alien to contemporary realist theory, the focus of the authors belonging to this loose and heterogeneous 'movement' is broader. Recent realist resurgence was triggered by political philosophies which conceived of politics as the application of pre-existing ethical principles elaborated via abstract reasoning. Contemporary political realism holds that any normativity in politics, provided we want it to be effectual, must conform to the specificities of politics. That is, realist political theory is a countermovement against political philosophies that disregard the autonomy of politics and see it as nothing more than applied ethics.

Political realists do not deny that normativity is an ineliminable part of politics. In fact, contemporary realist theorists maintain that politics deserves its name as a form of legitimate rule due to its contrast to mere domination. Unlike mainstream political philosophy, however, realists do not think that abstract values such as justice could or should be considered before answering the 'first political question' of securing order as the condition of cooperation (Williams 2005, 3; Sleat 2018, 5–6). Order, security or legitimacy are values intrinsic to politics, therefore their realization conforms to the specificities of politics thereby providing actual normative standards for political action. By contrast, proponents of justice, equality and fairness fail to understand politics properly when they begin by elaborating abstract ideals and expect them to be applied to the sphere of politics. The same applies to ideals of transparency or explainability, along with some other values often discussed in the context of AI ethics. Moreover, the applied ethics approach can be regarded as an attempt 'to evade, displace or escape from politics' (Galston 2010, 386) against which realists, to become politically more relevant, demand the autonomy of politics and political thinking (Galston 2010, 408; Geuss 2008, 23; Sleat 2011, 471; Williams 2005, 3). Applied mo-

rality, as Bernard Williams calls it, is a mistaken way of conceiving normativity in politics (Williams 2005, 2).

Mainstream political philosophers, advocates of the applied ethics view, maintain that politics is just one of the spheres that has its own set of rules, yet the source of normativity for them remains abstract morality (Leader-Maynard and Worsnip 2018, 761). Discussions of several areas which have quite specific codes of conduct, ethical guidelines and regulations (such as medicine) are predominated with this approach. We argue that this approach fails to answer a crucial question: how to expect abstract values to regulate, justify and measure (political) behaviour without having the necessary means (enforceability) to become effectual in politics? Self-imposed guidelines fulfil their task only up to the point where they do not undermine the goals of their authors, or where the values incorporated into them do not conflict with each other. AI guidelines just because they are normatively attractive will not be binding without being enforceable. As realists tend to say, it is politics that provides collectively binding – and legitimate – decisions (Burelli 2020).

This leads us to a further insight about guidance conceived as a link between theory and practice. Realists criticize mainstream political philosophers due to the political irrelevance of their abstract morality. Realists claim that their approach is epistemologically better since they acknowledge that politics has its own set of rules and contextually embedded norms. From political relevance, realists proceed in two different directions. Some theorists maintain that political theory should elaborate measures, justification and critical standards for the political sphere. Realist political theory can give guidance to politics precisely because by acknowledging the autonomy of the political realm, political theory can become more relevant (Rossi and Sleat 2014, 689). Others, however, suggest that more relevant – realist – descriptions of politics mean that political theory should give up any aspirations to become action-guiding; it should remain a descriptive effort, and focus on what 'thinking politically' means (Horton 2017; Freeden 2018). It might be a reasonable compromise to acknowledge that theory can become action-guiding, but what practitioners actually 'follow' is 'contingent and, crucially, incidental' to the purposes of political theory (Horton 2017, 498). That is, it is unreasonable and unrealistic to expect that only because a guideline meets the abstract standards it becomes binding and relevant for practice.

As we have seen, realists do not claim that politics should be devoid of normativity. At the same time, realism is considered as a methodological innovation in political theory which, instead of putting forward specific proposals for political action, argues for a closer look on politics before defining the values to be realized (Rossi 2016; Jubb 2017). Although in what follows, we will make use of some of the substantive claims offered by realism, on the whole, our aim is to draw attention to politics both as the actual context of guiding and regulating AI and as the nature of these efforts that aim at bridging abstract norms and practice. That is, we do not suggest that the AI guidelines problem can only be dealt with by adopting realism. Realism, however, is a good way of showing the limits of guidance in a context where the need for binding decisions appears on the horizon.

3. The politics of AI: Enforceability and collectively binding decisions

In the context of current debates on regulating AI, politics becomes meaningful on two levels. First, the *AI in politics* problem concerns the use of artificial intelligence in politics. According to an instrumental understanding, AI needs to be regulated due to its profound (often worrying) impact on political practice from campaigning to administrative decision-making. Second, the *politics of AI* needs to be discussed to delineate the political nature of the problem of regulation, by which we mean that any attempt to achieve a normatively binding prescription for developing and applying AI is necessarily a political action. As for the former level, we take concerns about the proliferation of AI and the hopes regarding its regulation to be based on an idealistic understanding of politics. Taking the much-debated issue of interfering with voter behaviour as a brief example, regulation often comes down to attempts to exclude manipulation from political practice. As the Cambridge Analytica scandal or the Brexit referendum (Schippers 2020) exposed, political professionals rely heavily on machine learning hoping that the vast amount of voter information they gathered can be turned into meaningful data for segmentation, targeting and messaging at unprecedented levels of precision. Although it is worth noting that, especially in US electoral politics, computational tools are far from new and have been present since the 1950s (McKelvey 2021; Issenberg 2012), recent advances in AI precipitated worries about how technology undermines democracy. AI is seen to be weakening democracy by increasingly centralizing and controlling information and communication, creating fake identities, support and messages, as well as by altering the perceived political reality by reinforcing filter bubbles (Savaget 2019). In sum, the problem with AI and the reason why it should be regulated is that it provides politicians with more efficient means of manipulation whereas voters become more disempowered, and thereby accountability as a fundamental feature of democratic governance becomes void. However, more realistic authors have always been more cautious with the idea that voters are autonomous agents and that electoral politics without manipulation does exist (Schumpeter 1987; Körösényi 2010; Achen and Bartels 2016). Separating acceptable and unacceptable forms of influencing voter behaviour is a debate that is already political and cannot be solved by abstract ethical principles or self-regulation however carefully elaborated. Limiting political manipulation is not something beyond or before politics but a part of it, therefore politics seriously constrains the possibilities of how and to what effect the use of AI can be constrained. It is here where the AI in politics problem turns into the more abstract question of the politics of AI to which we turn now.

3.1. The role of the state

The problem of the state serves as a link between concerns about the use of AI in politics and the more abstract problem of the politics of AI. On the one hand, the state is evidently implied in the problems described above, yet it is hard to find in the discussion about AI ethics and guidelines. On the other hand, the general political nature of AI and AI regulation might be approached through the concept of the state, even if it is

far from being the sole route a political theorist might follow. The absence of the concept of the state from the AI ethics discourse is somewhat understandable inasmuch as the topic of regulation through guidelines is focused on self-governance. As Hagendorff (2020, 100) remarks, ethical guidelines developed by companies and research institutions serve to discourage the creation of a 'truly binding legal framework'. Even if such strategic implications might not be general in all systems of AI ethics, enforceability is an open question that cannot be answered by completely disregarding states as key players in the field. The role of the state emerges on two levels.

First, as Hagendorff rightly observes, ethics – AI ethics included – cannot reinforce itself as it lacks the necessary – coercive, we might add – means (Hagendorff 2020, 99). The state is obviously the primary existing candidate for this role. However, considering the discussion on AI guidelines it seems that AI ethics – as a kind of applied ethics – expects that its principles will be followed merely due to their rightness, epistemological soundness, and normative attractivity. This, however, does not answer enforceability or a situation in which competing values are present. Understandably, any self-regulating effort will reflect the particular position of its author even if there are attempts to put forward universal norms as well. Although not necessarily motivated by selfishness, the proliferation of ethical guidelines results in a proliferation of particular positions and the possibility of conflicting values as well. Recent studies (Hagendorff 2020; Héder 2020) have revealed several overlaps but also differences between these documents. From this, it follows that beyond lacking the means to enforce a guideline, a further problem results from there being many possibly conflicting guidelines. Just as with ethical beliefs in a society, in the absence of an arbitrator, not only does implementation remain unsolved but conflict resolution as well. Whereas the first problem implies inefficacy, the plurality of guidelines might result in a disorder of unaligned particularities.

The other problem concerns the supranational level of enforceability. Evidently, even if binding guidelines exist on the national level enforced by states, most of the problems that required regulation in the first place remain unsolved between and above states. However, it is unlikely that a global solution might be adopted in the form of a supranational regulatory agency (Erdélyi and Goldsmith 2018). Any global attempt to regulate the development and functioning of AI will be just as efficient as any previous effort to make binding decisions on, for instance, climate change or global tech companies. On a global level, states are as inevitable as they are insufficient when it comes to regulating AI at least until the establishment of a global government which, from a realist point of view, seems to be highly unlikely.

3.2. Value transfer and enforceability

As the condition of transferring values, the problem of enforceability emerges on a more abstract level, revealing how AI guidance is political by nature regardless of any actual value or political content. The applied ethics approach in the field of AI guidelines has become increasingly nuanced as more and more regulative levels and methods have been differentiated. Christian's (2020) alignment problem, for ex-

ample, addresses the questions of the methods and contents of value transfer from humans to (narrow) AI. Considering the emergence of superintelligence, Bostrom's value-loading problem also differentiates various ways of transferring human values into AI (Bostrom 2014). The discussion however remains focused on the interaction between normativity and technology while politics, if it appears at all, seems merely to be a disturbing factor to be neutralized. Criticizing Bostrom's account, however, Totschnig (2019, 916) emphasizes that the predominantly technological approach misreads the nature of control over a future superintelligence inasmuch as it should be considered as driven by the political dimension of self-interest. To avoid a warlike situation between a superintelligence and humans mutually considering each other as an existential threat, and achieving a peaceful coexistence, AI must not be antagonized by treating it as a tool or servant (Totschnig 2019). While Totschnig's realist implications are promising, in the end, the described relation between humans and AI becomes idealistically depoliticized and the seemingly political model fails to address the relationship between value transfer and enforceability.

Totschnig notes that the mutual existential threat that characterizes the warlike situation between AI and human agents lasts only until the AI begins to develop into superintelligence (Totschnig 2019). From that point, AI will have control over all the necessary means to transform the mutual threat into a one-sided vulnerability of humans. Certainly, if the superintelligence decides to get rid of humans, their political situation dissolves. The other option, however, namely the peaceful coexistence achieved by recognizing the self-interest of AI, surprisingly fails to develop an adequate account of politics and the political context of regulation. Contrary to what Totschnig says here, peaceful coexistence through recognition does not end the Hobbesian warlike situation but actually extends it towards politics. Let us remind ourselves that under the realist framework politics is more than successful domination; it is a legitimate form of rule that is not merely acknowledged just because there is no other viable option but understood as acceptable. A threatening superintelligence and acquiescent humans cannot have a political relationship. Both the value-alignment and the value-loading problem should be raised at this point. Considering that instead of mutual vulnerability humans are now disproportionally weaker, no guideline can be transferred to AIs based merely on the attractivity of its principles. The situation becomes political when coexistence with superintelligence exceeds mere acquiescence and resignation. That is, to be called properly political, any cooperation or order needs to be justified, however, justification emerges from within the very context of coexistence and the shifting power-relations between human and AI agents.

This extension of Totschnig's reading of Bostrom's approach is meant to be a model of how AI guidance can be conceived. Analogous to what we said about how the realist position reveals the boundaries within which actions and relations can be called political, superintelligence in the extended example above serves as the extreme case of guiding AI. Given that the wider AI problem is about the externalization and delegation of more and more human decisions to artificial intelligence (Chiodo 2020), no paradigmatic difference exists between narrow AI and strong AI or superintelligence, when human decision-makers are expected to develop guide-

lines which, present-day at least, relate to the development, use and functioning of AI as well. Following from our 'politicized' account of the context of AI guidelines, it might be concluded that the political relation is multi-layered, and encompasses human versus human, human versus AI, and possibly AI versus AI relationships, given the condition of enforceability and binding prescription emerging on the horizon.

4. Implications of AI to Political Theory and the concept of regulation

In the previous section, we demonstrated how a political theoretical perspective on artificial intelligence reveals its particular political nature, and we showed why it is mistaken to regard AI as a regulatory rather than a political problem. In this section, we flip the perspective and unpack the implications of addressing AI as a political problem, that is, we show what political theory can learn from discussions about the regulation of artificial intelligence. We argue that two fundamental implications follow from this perspective. One implication is that AI and claims to its regulation embody a compilation among the traditional problems of political theory. Therefore, the seemingly new fears around AI and its regulation are, in fact, well-known problems for political thought. The other implication involves claiming that AI – even in its strong form, such as artificial general intelligence or superintelligence – is a valuable tool for understanding the political realm if applied as a particular political theoretical methodology, precisely, in a thought experiment.

4.1. The impossibility of regulation: A classical problem

The increasingly popular topic of artificial intelligence may seem marginal from a political theoretical perspective. However, contemporary debates about some dimensions of AI are, in effect, political in the sense that they revive some classical dilemmas of political thought. The current discussion about regulating AI and the need for ethical guidelines – perhaps because of the urgency of such claims – is surrounded by an unsettling atmosphere, whereas regulation is a longstanding issue for political theory. Roughly speaking, politics is exactly about the problem of control. That is, now we show how current debates on AI guidelines reiterate earlier concerns of political theory.

The idea of an artificial entity and its inherent dangers to humanity has engaged ancient imagination, for example, in Greek mythology, in the form of Pandora, an artificial person created by Hephaestus (Pereira 2021). However, artificial intelligence has appeared in modern political thought as a distinctively political idea. Hobbes is rightfully known as 'the grandfather of AI' (Haugeland 1985, 23) for two reasons. First, he invented the idea of reasoning as computation, and second, he elaborated the idea of the application of an artificial person for politics (Hobbes 1651, 1655). In his reasoning, to overcome the brute reality of the state of nature, in which conflict is permanent, the 'unity of the multitude' brings into being the 'state,' the Leviathan, which is a 'fictional character', by authorizing a representative, who represents the

state by acting in the name of it. This representative, the 'sovereign', is another artificial person (embodied by a natural person or an assembly) who also lacked any existence before the act of covenanting (Skinner 2018, 358). Therefore, the Hobbesian theory of social contract aims to not only argue for the desirability of political order, and thus, for the need for government, but also to offer a tool for justifying the legitimacy of rule (ibid. 360–361). While the Hobbesian conception of war of all against all in the state of nature refers to the dangers inherent in the absence of political order (as it was seen above concerning our expansion of Totschnig's reading of Bostrom's superintelligence), Hobbes's concerns for legitimate authority imply a different type of danger inherent in political life. The conditions for the right to rule – one of which states that only the sovereign is authorized by governance and the other that governance must aim for the preservation of life and health of the members of the commonwealth (ibid.) – indicate that these artificially created entities always include the potential to exceed the constraints set for them. In the Hobbesian framework, this means that for the artificial person of the state, for which the metaphor of Leviathan seems particularly apt in this regard, there is a permanent threat of seizure by someone without proper authorization. At the same time, there is also an indispensable danger that the artificial person of the sovereign is a counter to the common good. The fear of the inherent potential of overreaching the scope of the authority is more explicit in Locke's social contract theory that permits the withdrawal of obedience to the sovereign in case of abuses of power (Locke 1689).

It is clear, therefore, that the aspiration to restrict artificial entities emerged at the very moment when the idea of artificial intelligence emerged. Nonetheless, and more importantly, artificial intelligence as a political concept is not only interconnected with attempts to specify its limits but can be regarded as a mechanism for balancing two persistent dangers of the political realm: the extremes of the disorder of the state of nature and the tyrannical, illegitimate use of force or even terror. Therefore, the original idea of artificial intelligence as a political conception highlights the inherent fragility of *the political*. In sum, the state and the sovereign, as archetypes of artificial political entities can at the same time offer desirable solutions and severe challenges for political life.

Recalling the realist viewpoint of the political sphere, it seems that the only attainable goal is a *modus vivendi*, which resonates with the idea that an inherent characteristic of the political world is balancing the possibilities of two extremes. History of politics supports this more pessimistic view: occasionally, eruptions of civil war and failed states still embody the brute reality of the Hobbesian state of nature, while the existence of authoritarian and totalitarian dictatures altogether with hybrid regimes are eternal reminders of the impossibility to limit power in a once-and-for-all manner.

In light of this reality of the political world, new claims for the regulation of artificial intelligence, more specifically on weak AI, are less promising. Debates on the regulation of AI concentrate on the need to connect principles such as fairness, accountability, safety, sustainability, and social inclusion, among others, to AI governance (for a more exhaustive list, see Hagendorff 2020). Nevertheless, the most discussed issue is transparency, which is among the primary claims for several AI

ethics guidelines released by different institutions and companies in the past few years.

The current boom in ethical guidelines for AI involves several criticisms concerning the effectiveness of such guidelines based on their potential to implement transparency and other claims effectively. This line of criticism can be divided into three types of argument. The first type challenges the AI guidelines on their extensive list of ethical claims based on their ineffectiveness. This type, which can be called 'tick-box criticism,' can be coupled with a proposal of some different approach, for example, virtue ethics (see Hagendorff 2020). The second type, which can be called 'double standard criticism', is more sceptical about the possibility of guiding AI and whether full transparency can be achieved at all. This criticism builds on the argument that it would be a double standard to call for higher transparency in AI compared to human decision tools and human reasoning (see Zerilli et al. 2019). The third type of criticism is more focused, what we call 'specificity criticism', and argues that current Artificial Intelligence Guidelines (AIGUs) are not specific to AI, but they are simply attempting to gain social control over technology. This criticism also demonstrates that transparency and explainability are claims that specifically concern AI because in such cases there is a possibility of the autonomy of AI. In that case, though, the double standard problem arises (see Héder 2020).

These criticisms imply that there is a profoundly political characteristic of AI. On the one hand, there is a relative autonomy inherent in AI that can be understood in a broader sense. It is impossible to regulate in every detail, something that can develop by itself. On the other hand, concerning the expert systems of weak AI, the double standard criticism and specificity criticism correctly acknowledged that it would be an unfair expectation to regulate the decision-making of artificial intelligence in domains where human decision-making cannot be entirely regulated likewise. However, contrary to the double standard criticism, we do not base our argument on the similarity between the obscurity of artificial decision tools and human cognitive processes. Instead, we build our argument on the political characteristic of AI. Using AI as a tool is similar to political authorization: although accountability is the main virtue in politics, it would be unrealistic to expect legislative, executive, or judicial officials to act 'perfectly'. We can only hope that they behave to the best of their knowledge, and while we usually hold them to account for significant breaches of their power, mostly, we authorize them because authorization is the only legitimate way to create order without slipping into a Hobbesian state of nature or a tyrannical regime.

4.2. Using AI as assistance for practical thinking

From a political theoretical perspective, the differences between weak and strong AI and the differences between claims for their regulation are not striking. Hence AI resonates with an inherent problem of political theory, what we called balancing between extremes; our scepticism towards full transparency concerns both weak and strong AI. Moreover, the more 'fictional' ideas of artificial general intelligence

(AGI) and superintelligence are also highly relevant for political theorizing. Political theoretical methodology frequently employs some fiction in the form of thought experiments and intuition pumps. These methods help us test our reasoning, build and destroy arguments, or explore our intuitions. Therefore, they can be used for different aims, and for this reason, they can lead to entirely different conclusions. In fact, the idea of a state of nature is a typical thought experiment, a mental visualization used by political theorists to justify their arguments on particular issues. Nevertheless, there are other methods involving fiction in political theory, such as idealistic or even utopian ideas about a just society and a just world or assumptions about perfectly reasonable individuals in situations of complete information. Also, there are dystopian ideas in political theory about the absence of order or a surveillance state.

From the perspective of political theory, therefore, the possibility of the emergence of AGI or superintelligence is not as marginal as for more technical discussions. AI has the potential to reveal the complexity and unpredictability of the political world and the role of human conduct in shaping the political world. Using AI as a thought experiment as Bostrom and other authors who engage themselves with the idea of strong AI sometimes do, reveals that claims for transparency and *a priori* determined rules can never be entirely enforced. Armstrong (2007), for example, elaborates a relatively universal solution to guide future superintelligence, still addresses several cases in which regulation may fail, and admits that in some of these scenarios, we are 'screwed'. Bostrom also addresses a broad range of potential problems concerning the transparency and regulation of a future superintelligence.

The idea of the emergence of multipolar general artificial intelligence is also a helpful tool to understand politics. The AI race (even in its weak form) is similar to other technological races in human history, involving crucial political challenges, such as the race for the fission bomb, fusion bomb, satellite launch, or the ICBM (Bostrom 2014). As Bostrom demonstrates, governments have always been seeking to gain control over such projects, which may provide them with a decisive strategic advantage (Bostrom 2014, chapter 5.). Such a scenario powerfully highlights the ineliminable nature of conflict in the political realm. However, if the emergence of AGI or superintelligence involves the dissolution of conflict from politics, it would automatically mean the obsolescence of humanity as we know it.

Concluding the section, we aimed to argue for the relevance of applying the idea of AI – both in its prevalent, weak and in its less discussed strong forms, such as AGI or superintelligence. The point of our argument was to show that there is nothing new in AI that would be unknown to the political world. The origins of the idea of AI from Hobbes's political theory to contemporary realist political theory implied that *the political* could not be entirely subjected to human oversight. Politics is precisely about balancing between the absence of order and terror. Besides there being no final solution that secures politics once and for all, there is no possibility to regulate artificial intelligence and secure all of its desired virtues and norms in advance. However, this is not a pessimistic conclusion that refuses any attempts to implement such norms; rather it is a confident argument for demonstrating that we already have a toolset – available in the political realm – for keeping AI under control.

5. Conclusion

The article argued for a (fundamentally realist) political theoretical approach to the problem of AI regulation. Our aim was to take one step back from the current debates on AI guidelines and to investigate the context in which the claims of regulation appear, and – as a result – to question the 'applied ethics' type of attempts that aim towards *a priori* laid-down rules for AI. In the article, first, we sketched the main characteristics of a realist view, then we demonstrated how this perspective highlights that AI is, primarily, a political rather than a regulatory or a technical problem. In doing so, we identified two problems: one is about the problem of *AI in politics* and the other one is what we called the *politics of AI* problem. Regarding the former one, we showed our concerns for the way AI transforms democratic politics. Concerning the latter one, we discussed the role of the state in the enforcement of AI regulation; while claiming that there is also a deeper problem of choosing, aligning, and loading values we want for AI. Finally, we addressed what current attention to AI can teach political theory. In this regard, we first demonstrated that the question of regulation is a classical and irresolvable problem of political thought, to which any attempts seem to be doomed to failure. Second, we showed that taking AI as a thought experiment may help us understand how our political world operates.

The article touched upon some further issues that should be considered not only from a (realist) political theoretical view, but from a broader scope of discussions as well. One question is about the arbitrariness of the values we seek to regulate and to be implemented in AI. While current debates are focusing on the problem of formulating AI in a way desirable for us (let us say, for humanity *per se*), there is a preceding problem of which values to choose and what to do when there are competing or even conflicting values. Although it could be justified that values such as transparency and human oversight are primary values from the perspective of political theory's general commitment to democratic values and participation, it can be argued that some directly emancipatory values such as fairness and solidarity in AI are just as important.

Finally, we revisit the criticism we applied in the article. Basically, we argued for a political realist approach due to its methodological innovations and we mentioned the usefulness of some of its substantial claims. However, there could be other approaches that criticize the so-called 'applied ethics' approach to the regulation of AI from a different perspective, and in fact, there are some attempts for a virtue ethics view (see Hagendorff 2020). We did not intend to exclude the appropriateness of such perspectives in substantive matters. Rather, we attempted to address the problem of AI from a broader perspective that takes politics seriously.

References

Armstrong, Stuart. *Chaining God. A qualitative approach to AI, trust and moral systems.* Unpublished manuscript. 2007.

Bostrom, Nick. *Superintelligence: Paths, Dangers, Strategies*. Oxford United Kingdom; New York, NY: Oxford University Press, 2014.

Burelli, Carlo. "Political Normativity and the Functional Autonomy of Politics." *European Journal of Political Theory*, 2020. https://doi.org/10.1177/1474885120918500

Chiodo, Simona. "The Greatest Epistemological Externalisation: Reflecting on the Puzzling Direction We Are Heading to through Algorithmic Automatisation." *AI & SOCIETY* 35, no. 2 (2020): 431–40. https://doi.org/10.1007/s00146-019-00905-y

Christian, Brian. *The Alignment Problem: Machine Learning and Human Values.* W. W. Norton and Company, 2020.

Damnjanović, Ivana. "*Polity* Without Politics? Artificial Intelligence Versus Democracy: Lessons From Neal Asher's Polity Universe." *Bulletin of Science, Technology & Society* 35, no. 3–4 (2015): 76–83. https://doi.org/10.1177/0270467615623877

Erdélyi, Olivia J., and Judy Goldsmith. "Regulating Artificial Intelligence: Proposal for a Global Solution." In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, 95–101. New Orleans LA USA: ACM (2018). https://doi.org/10.1145/3278721.3278731

Freeden, Michael: "Political Realism: A Reality Check." In *Politics Recovered: Realist Thought in Theory and Practice*, edited by Sleat Matt, 344–68. New York: Columbia University Press, 2018. http://www.jstor.org/stable/10.7312/slea17528.18

Galston, William A. "Realism in Political Theory." *European Journal of Political Theory* 9, no. 4 (2010): 385–411. https://doi.org/10.1177/1474885110374001

Geuss, Raymond. *Philosophy and Real Politics*. Princeton: Princeton University Press, 2008.

Hagendorff, Thilo. "The Ethics of AI Ethics: An Evaluation of Guidelines." *Minds and Machines* 30, no. 1 (2020): 99–120. https://doi.org/10.1007/s11023-020-09517-8

Hatamleh, Omar, and Tilesch George. *Betweenbrains: Taking back our AI Future.* GTPublishDrive. 2020.

Haugeland, John. *Artificial Intelligence: The Very Idea*. First MIT Press paperback edition. 1989Bradford Books. Cambridge, Mass.: MIT Press, 1985.

Héder Mihály. "A Criticism of AI Ethics Guidelines." *Információs Társadalom* 20, no. 4 (2020) https://doi.org/10.22503/inftars.XX.2020.4.5

Hobbes, Thomas. *De Corpore.* 1655.

Hobbes, Thomas. *Leviathan.* 1651.

Horton, John. "What Might It Mean for Political Theory to Be More "Realistic"?" *Philosophia* 45, no. 2 (2017): 487–501. https://doi.org/10.1007/s11406-016-9799-3

Issenberg, Sasha. *The Victory Lab: The Secret Science of Winning Campaigns*. 1st ed. New York: Crown, 2012.

Jubb, Robert. "Realism." In *Methods In Analytical Political Theory,* edited by Adrian Blau, 112–l30. Cambridge: Cambridge University Press https://doi.org/10.1017/9781316162576.008

Körösényi András. "Stuck in Escher's staircase: Leadership, Manipulation and Democracy." *Osterreichische Zeitshrift Fur Politikwissenschaft* 39, no. 3 (2010): 289–302.

Leader Maynard, Jonathan, and Alex Worsnip. "Is There a Distinctively Political Normativity?" *Ethics* 128, no. 4 (2018): 756–87. https://doi.org/10.1086/697449

Locke, John. *Two Treatises of Government.* 1689.

McKelvey, Fenwick. "The Other Cambridge Analytics: Early "Artificial Intelligence" in American Political Science." In *The Cultural Life of Machine Learning*, edited by Jonathan Roberge and Michael Castelle, 117–42. Cham: Springer International Publishing, 2021. https://doi.org/10.1007/978-3-030-56286-1_4

Morgenthau, Hans J. *Politics among Nations: The Struggle for Power and Peace*. 7th ed. Boston: McGraw-Hill Higher Education, 2005.

Pereira, Luís Moniz. "The Carousel of Ethical Machinery." *AI & SOCIETY* 36, no. 1 (2021): 185–96. https://doi.org/10.1007/s00146-020-00994-0

Rossi, Enzo. "Can Realism Move Beyond a *Methodenstreit* ?" *Political Theory* 44, no. 3 (2016): 410–20. https://doi.org/10.1177/0090591715621507

Rossi, Enzo, and Matt Sleat. "Realism in Normative Political Theory: Realism in Normative Political Theory." *Philosophy Compass* 9, no. 10 (2014): 689–701. https://doi.org/10.1111/phc3.12148

Savaget, Paulo, Tulio Chiarini, and Steve Evans. "Empowering Political Participation through Artificial Intelligence." *Science and Public Policy* 46, no. 3 (2019): 369–80. https://doi.org/10.1093/scipol/scy064

Schippers, Birgit. "Artificial Intelligence and Democratic Politics." *Political Insight* 11, no. 1 (2020): 32–35. https://doi.org/10.1177/2041905820911746

Skinner, Quentin. *From Humanism to Hobbes. Studies in Rhetoric and Politics.* Cambridge, Cambridge University Press. 2018.

Sleat, Matt. "Liberal Realism: A Liberal Response to the Realist Critique.". *The Review of Politics* 73, no. 3 (2011): 469–96. https://doi.org/10.1017/S0034670511003457

Sleat, Matt, ed. *Politics Recovered: Realist Thought in Theory and Practice*. New York: Columbia University Press, 2018.

Smuha, Nathalie A. "From a "Race to AI" to a "Race to AI Regulation": Regulatory Competition for Artificial Intelligence." *Law, Innovation and Technology* 13, no. 1 (2021): 57–84. https://doi.org/10.1080/17579961.2021.1898300

Totschnig, Wolfhart. "The Problem of Superintelligence: Political, Not Technological." *AI & SOCIETY* 34, no. 4 (2019): 907–20. https://doi.org/10.1007/s00146-017-0753-0

Williams, Bernard. *In the Beginning Was the Deed: Realism and Moralism in Political Argument*. Princeton, N.J.: Princeton Univ. Press, 2005.

Zerilli, John, Alistair Knott, James Maclaurin, and Colin Gavaghan. "Transparency in Algorithmic and Human Decision-Making: Is There a Double Standard?" *Philosophy & Technology* 32, no. 4 (2019): 661–83. https://doi.org/10.1007/s13347-018-0330-6

Diversity-sensitive social platforms and responsibility

Some ethical considerations

There is an ongoing debate on how algorithms and machine learning can and should deal with human diversity while avoiding the pitfalls of statistical stereotyping, the re-enforcement of clichés and the perpetuation of unjust discrimination. Computer scientists try to tackle these issues by developing algorithms and social-interaction protocols for mediating diversity-aware interactions between people, for instance on diversity-sensitive social platforms. At the same time, diversity-related data often comprise sensitive personal data, and their collection, storage and management increases the vulnerability of users to various misuse scenarios. Already this observation leads to the question, how do we need to conceptualize responsibility to do justice to the increased vulnerability? In this paper, I thus focus on the questions a diversity-sensitive social platform raises with regard to responsibility, and propose a tentative ethical framework of responsibility for these platforms.

Keywords: *algorithms, applied ethics, diversity, responsibility, social platforms*

Acknowledgement

The research for this paper was partially conducted as part of the European Union Horizon 2020 Project "WeNet - The Internet of us" (Grant no. 823783) and as part of the Baden–Württemberg Foundation Project "AITE – Artificial Intelligence, Trustworthiness and Explainability". A draft version was presented at the Budapest Workshop on Philosophy of Technology in December 2019. I want to thank the participants of that workshop for the helpful discussions and comments that followed. I also want to thank Jessica Hessen, Moritz Hildt and an anonymous reviewer for their helpful suggestions.

Author Information

Karoline Reinhardt, Eberhard Karls Universität Tübingen
https://uni-tuebingen.de/en/facilities/central-institutions/international-center-for-ethics-in-the-sciences-and-humanities/team/dr-karoline-reinhardt/

How to cite this article:

Reinhardt, Karoline. "Diversity-sensitive social platforms and responsibility." *Információs Társadalom* XXI, no. 2 (2021): 43–62.
https://dx.doi.org/10.22503/inftars.XXI.2021.2.4

1. Introduction

Despite the apparent diversity of humans, technology in general and digital solutions in particular still struggle to take diversity into account. Thus, diversity poses not only an ethical but also a computational challenge. There is an ongoing debate on how computerized algorithms and machine learning algorithms can and should deal with human diversity while avoiding the pitfalls of statistical stereotyping, the re-enforcement of clichés and the perpetuation of discrimination against particular groups.

Computer scientists and developers are increasingly aware of these problems, and are trying to develop technical methods, algorithms and social-interaction protocols that enable machine-mediated diversity-aware interactions between people, for instance on diversity-sensitive social platforms (cf. Reinhardt 2020a, 272). To measure diversity and to make sure that diversity is instantiated in turn requires the collection of data. Diversity-related data, however, often comprise sensitive personal data. The collection of diversity-related data, thus, touches not only on questions of informational privacy, but also increases the risk of various misuse scenarios, online (e.g. hate speech, trolling, cyberbullying, misinformation) and offline (e.g. persecution).

The often, though not always, honourable goal, to make a social platform diversity-aware, thus, comes with an increased vulnerability for the users and data subjects of the platform – a vulnerability that is increasingly exploited by some service providers and other platform users. Hence, this vulnerability alone gives us sufficient reason to think about questions of responsibility with regard to diversity-sensitive social platforms. Additionally, technical innovations, especially those with an undeniable extent of societal impact, have consequences – both intended and unintended. It is part of responsible research, as well as of responsible action in general, to be mindful of the possible consequences in advance and to set up structures that do justice to this fact. This also includes a reflection on the appropriate concepts and structures of responsibility. Concerning social platforms, this has not yet been sufficiently done, though there are of course some notable exceptions[1]. This article is, thus, intended also as a contribution to a more general debate on responsibility and social platforms.

In what follows, I outline some ethical questions that arise from the design process of diversity-aware social platforms. To do this, I first define what platforms and social platforms are and explain that the term diversity itself also raises some conceptual questions that we should bear in mind. Then, I will focus on the questions a diversity-sensitive social platform raises with regard to responsibility, and sketch a tentative ethical framework of responsibility that formulates some recommendations for relevant stakeholders.

2. Diversity-aware social platforms and their flip side

In computer science, the notion ‘platform’ refers to a unified basis on which application programs can be executed and developed. Platforms are digital structures

[1] With regard to responsibility and content moderation as well as to content generation, these include, among others, the works of Gillespie (2018a, 2018b) and Gorwa, Binns and Katzenbach (2020).

that store data and enable and mediate interactions on a large scale. There is a great variety of platforms on the market and their development and evolution are highly dynamic. Platforms include search engines, comparison and rating portals, marketplaces, trading platforms, media and content services, online games, social networks and communication services.

Herein, I refer to platforms as social platforms when their aim is to connect people, or to mediate interactions between individual persons – in contrast, for instance, to platforms that enable digital communication between industrial machines. This distinction is, of course, based on the implicit thesis that the interaction of humans – whether digitally mediated or not – raises ethical questions that are different from the ethical questions that are raised by the interaction of machines. Such interactions on social platforms include, but are not limited to, transactions, content production and consumption, or joint activities that remain within the virtual world or go beyond it. Social media sites are among the most prevalent examples of social platforms.

Increasingly, social platforms transcend the management and storage of user-created information and employ algorithmic decision-making procedures (cf. Gillespie 2018a; 2018b). This gives them considerable influence over their users – an influence that is often designed to be imperceptible. In this way, platforms have become important players in their respective fields with an often unknown and still insufficiently understood impact on people's behaviour and decisions.[2] However, although the algorithmic decision-making procedures deployed in social platforms are highly evolved, they often seem to struggle with the diversity of humans and tend to reproduce, perpetuate and amplify discrimination and stereotypes (cf. Heesen, Reinhardt and Schelenz 2021). As computer science has become more and more aware of this, some developers and designers have tried to make algorithms and interaction protocols more diversity-sensitive.

There are, of course, many different approaches to the topic of diversity. Some start from an instrumental understanding of diversity, in which diversity is understood as fostering other organizational goals; for instance, economic success (cf. Mensi-Klarbach 2012), while others take a normative approach to diversity; for instance, by emphasizing its intrinsic value (Parekh 2000). There are also representational and inclusive concepts of diversity (Young 1990), and concepts that emphasize the ambiguities of human diversity (Reinhardt 2020a). In any case, making algorithmic decision-procedures diversity-sensitive requires the collection of data that comprise – depending on the paradigm of diversity used – usually highly sensitive personal data.

A platform that collects, stores and works with data relevant to diversity, thus carries a considerable risk of increasing the vulnerability of individuals to all kinds of misuse scenarios such as harassment, stalking, political persecution and many others.[3] If a platform is, furthermore, designed to not only connect people online but also to help

[2] In recent years, we have seen a number of publications that have taken a critical stance towards platforms and the underlying algorithms and mathematical models, among them O'Neil (2016), Gillespie (2018a; 2018b), van Dijk, Poell and de Waal (2018) and Zuboff (2019).

[3] For an overview of the risks associated with social media platforms as one of the primary examples of social platforms, see Brake (2014).

them to organize and engage in joint activities in the offline world, these risks transcend the virtual lives we live and become real-world threats. That means creating a diversity-aware social platform, which *prima facie* seems to be a good thing, does in fact come with the risk of increasing the vulnerability of people to all kinds of risky and potentially horrible scenarios. This raises multiple questions, including: Who is responsible for the often unintended consequences of a diversity-aware social platform – and to what extent? How are we even to conceptualize responsibility with regard to platform-mediated interactions that by large part are powered by the aggregation of data?

3. Responsibility: Traditional concepts and evolving technologies

3.1. Responsibility in ethics and moral philosophy

There is an ongoing debate on (moral) responsibility in philosophy, which I will not attempt to summarize in its entirety here. Instead, in what follows, I start from a conception of responsibility that focuses on the responsibility for one's own actions, as formulated, for instance, by Nida-Rümelin (2011). This approach does not differ from common theories of responsibility in all aspects, but does so in one crucial respect, which I will highlight here. Since this text is, however, primarily concerned with conceptions of responsibility with regard to diversity-sensitive social platforms, I will only outline this action-centred approach here, but will not provide an independent justification of this perspective on responsibility.

Many takes on responsibility in ethics and moral philosophy share the following three elements as summarized by Noorman (2020): "A person is usually only held responsible if she had some control over the outcome of events", and the person in question was "able to freely choose to act" in that way or another. However, we "tend to excuse someone from blame if they could not have known that their actions would lead to a harmful event" (Noorman 2020). Lenk argues that responsibility could be summarized in the following way: Somebody, the bearer of responsibility, is responsible for something – an action, consequences, a state of affairs, but also to fulfil certain tasks – to an addressee, possibly before a sanctioning authority with regard to a certain (prescriptive/normative) criterion in a given context (Lenk 2016, 7; see also Höffe 2008, 326).

As intuitive as it may seem at first glance, the list of what people are in general taken to be responsible for (actions, consequences, state of affairs, tasks) is not uncontroversial and it is exactly this point where Nida-Rümelin introduces an important specification. According to Nida-Rümelin, we are only responsible for our own actions – and not for all the consequences that follow causally from our actions.

Let me elaborate this point a little: That I am ultimately responsible for my actions does not mean that their consequences play no role: Each action has a context from which certain probability distributions for possible consequences of this action result. These probability distributions are part of a complete description of the action (Nida-Rümelin 2011, 113). To put it in Nida-Rümelin's words, this means: "I bear responsibility for the possible consequences of my action, weighted according to their respective probabilities" (Nida-Rümelin 2011, 111, my translation).

And I would add taking into account those measures that I take so that negative consequences do not occur for others. For example, the safety measures I take so that, when I brake at a red light, no-one is going to rear-end me, like driving with foresight, taking a look in the rear-view mirror before braking, and not braking suddenly but slowly if possible, help to prevent bad things from happening from my action (braking at a red light). If despite all these measures (provided they are executed properly) a rear-end collision happens, it is not my responsibility. The distractions of other drivers are not my responsibility as long as I give them enough time to react properly to my actions. The collision is, then, a causal consequence of my braking at the red light (it would not have occurred if I had not braked), but I am not responsible for it. The same holds true for "coincidences, which I cannot control, that lead to the realization of one sequence of actions and not the other" (Nida-Rümelin 2011, 111).

I bear, though, a greater responsibility to take countermeasures the more likely a negative consequence of my action is for others or when a certain probability is considered unacceptable for a particular consequence of action. To give an example: Even a 5% chance that my action will result in a person's death is unacceptable, even though the number is relatively low.

So I am only responsible for those consequences that are part of the probability distribution of my action in a given context to the extent that those consequences are likely. Since I may seek to always do good to other persons, but not to harm them against their will (Nida-Rümelin 2011, 214), possible harm to others has to have priority in my deliberations on how to act.

This concept of responsibility, moreover, allows attributing responsibility even if there are no negative consequences, since it is concerned solely with the action and the probability distribution of its possible consequences. It is also irrelevant for the attribution of responsibility, according to this approach, whether the consequences actually occur. Nida-Rümelin uses the following example to illustrate this point: Someone throws paving stones from a bridge onto a busy highway. Even if, fortunately, no cars are hit and no-one is killed, the probability that this action will have fatal consequences is unacceptably high: "The favourable coincidence does not relieve the person of her responsibility. She bears responsibility for an action that involves unacceptably high risks to other people" (Nida-Rümelin 2011, 112).

On top of that, I would like to emphasize a point that is not stressed by Nida-Rümelin: Such an understanding of responsibility ultimately permits a high degree of proactivity. Since the approach does not focus on the consequences of action, it requires actors to take countermeasures against possible negative consequences before they even happen and, more importantly, to think with foresight. I will build on this action-centred understanding of responsibility in what follows, and apply it to the questions at hand and expand it where needed.

3.2. Responsibility in the age of digitalization

Traditional concepts of responsibility raise a number of fundamental philosophical questions: They work with assumptions about causality and accountability. They

also raise epistemological and psychological questions. They refer to the notion of freedom and choice and in general assume that we do have some kind of control over the outcome of our actions. All of these issues are highly controversial topics within philosophical debates and beyond. With regard to technological change and digitalization, even more problems arise, as described below.

Traditionally, philosophical ethics have focussed on human actions and interactions with respect to responsibility. Responsibility was about human actions, their intentions and consequences (Fisher 1999). However, as our world has become more and more defined by technological artefacts, it has become less and less appropriate to solely take human action into account when tackling questions of responsibility. Technological artefacts change how we view the world and ourselves (Weizenbaum 1976), but they also have an impact on our decisions and how we make them (Latour 1992). This seems in particular true with regard to computers and digital tools. They have become "active mediators" that "actively co-shape people's being in the world" (Verbeek 2006, 364). A responsibility concept that solely focusses on human-to-human interaction, thus, seems in many ways inapt in a world that is increasingly shaped and defined by technologically mediated interactions.

Another prominent problem of a traditional concept of responsibility refers to the nexus of complexity, accountability and attributing responsibility. Since "computer technologies can obscure the causal connection between a person's action and the eventual consequences" (Noorman 2020), attributing responsibility and defining who is accountable for a particular outcome of a specific technologically mediated interaction has become more difficult.

This observation is in some ways linked to a further issue, namely the complexity problem. Complex technologies can make it difficult to ascribe responsibility (Johnson and Powers 2005; cf. Noorman 2020), because humans increasingly do not understand the workings of the machines they use. Therefore, some argue: "The more complex computer technologies become and the less human beings can directly control or intervene in the behaviour of these technologies, the less we can reasonably hold human beings responsible for these technologies"(Noorman 2020; cf. Matthias 2004). Note that this point about complexity is even more true of social platforms where not only the technology is complex but also the interaction patterns on them.

A further complication when ascribing responsibility is that computerized human interactions possibly distance us (spatially and temporarily) from the outcome of our actions. With the help of a social platform, you can get in touch with people from different places, and the traces you leave on a social platform might, through their temporal persistence, have effects on others a long time after you have left them. This, in turn, leads to epistemic problems; whereby it might be quite difficult to fully comprehend what might happen with the content you produce now in the future, or what consequences it might have in other societal contexts that you might have little knowledge about.

Interactions on social platforms, furthermore, raise questions about collective responsibility, since in the case of a social platform, we not only deal with individuals that we interact with, but with a whole community of users whose actions affect

people online and offline. The notion of collective responsibility makes the group out as the bearer of responsibility – distinct from each individual member of that respective group. Cybermobbing and cyberbullying are important examples of how scale and publicity can have a particular destructive impact. Though mobbing and bullying are (sadly) not new phenomena, what is different in the online world is how persistent content is and how easily content can be replicated and traced (cf. Schmidt 2016, 290). Therefore, cyberbullying has also different psychological consequences: The knowledge that the defamations and insults used may be public and that access to them is not limited to the original perpetrators and bystanders, i.e. that they are stored and can in principle be found and shared by anyone, makes it much more difficult to deal with such attacks.

Collective responsibility is, however, a highly controversial topic in philosophy. One of the controversies is whether groups have what it takes to be attributed moral agency and thus moral responsibility. In particular, the question of whether and how we can make out something like intentionality for a collectives take centre stage. Another controversy about collective responsibility deals with how to distribute responsibility (or accountability and liability) among the members of the group in question. A third controversy revolves around the question about what effects a notion of collective responsibility would have with regard to individual responsibility. One prominent issue in this debate is whether we run the risk of replacing the notion of individual responsibility altogether.[4] In this light, some argue that collective responsibility exists, if at all, only as a cooperative responsibility, that is, as a responsibility for joint action. Everything else, such as ascribing moral agency and intentionality to groups, would be simply 'mystifications' (Nida-Rümelin 2018, 393). Cooperative responsibility assumes that at the beginning of cooperative action there is an explicit or implicit consensus on the goals to be jointly achieved (Nida-Rümelin 2011, 119). Whoever then participates in a cooperative act is jointly responsible for its execution. The causal role of one's own contribution is therefore irrelevant for the attribution of responsibility (Nida-Rümelin 2011, 125).

Coming back to the example of social platforms, as helpful as the distinction Nida-Rümelin made for acts in our offline lives, how helpful is it for highly complex, interconnected, technically mediated actions of various people who possibly have never met or not even communicated? Their actions, as in the case of cyberbullying, sometimes lead to hideous consequences, but we seem not to be always able to determine that there has been some sort of consensus on the nature of the action or its intended results. It is not the individual insult, although wrong in itself, that has disastrous consequences here, but, as mentioned above, the combination of scale, shareability and persistence. A strength of the notion of cooperative responsibility here is that that the quantitative contribution of the individual action is not decisive nor is the causal role of one's contribution, but only the shared intention to act in a certain way. What it lacks, however, with regard to the interaction patterns on social

[4] For an overview of the debate on collective responsibility, see Smiley (2017).

platforms, is that there is, in practice, not always a shared intention.[5] In light of all this, it might be worthwhile to think about a notion of responsibility for 'aggregated action': This type of responsibility shares with the concept of cooperative responsibility the observation that the individual act and the blameworthiness attributed to it may be out of proportion regarding the damage caused by the sum of all the individual acts contributing to the outcome.

In any case, if one thinks that the notion of cooperative responsibility is sound, then it does not only apply to the community of users on a platform but also to the team that is designing the platform or the application. But also in this case the question arises as to whether it is really solely the agreement to commit a joint action for which responsibility can be ascribed, or whether, at least in some cases, responsibility for aggregated action is called for.

This leads us to the professional responsibility of researchers, developers and designers for the knowledge and products they produce and the question about whether they can be held responsible for the (often unintended) consequences of the inventions they have created (cf. Verbeek 2006, 379). One problem, among others, with attributing responsibility for unforeseen and, thus, unintended consequences is the 'interpretative flexibility' of users. Since "[p]eople often use technologies in ways unforeseen by their designers. This interpretative flexibility makes it difficult for designers to anticipate all the possible outcomes of the use of their technologies" (Noorman 2020).

Connected to the problem of attributing responsibility and professional responsibility is the "problem of many hands" (cf. Jonas 1984; Nissenbaum 1994 and 1996, 28-32; van de Poel, Royakkers and Zwart 2015): Since the development of a technology like a social platform involves many actors (designers, developers, computer scientists, sociologists, ethicists), it is often difficult to ascribe responsibility for a certain outcome created by that technology. In any case, an action-oriented (and not consequence-oriented) concept of responsibility holds that even if a technology in whose development I am involved has serious consequences that are even causally related to my actions, I may not ultimately be responsible for them, as shown above. Nevertheless, I am still responsible for ensuring I do not act negligently and that means also anticipating the possible consequences of my actions within human limits – and also, I would want to argue, anticipating the possible contribution of my actions to certain outcomes, if aggregated. In particular, when I am using a tool that works by aggregating information on actions like a social platform, my responsibility also entails including other perspectives in order to foresee as many side effects as possible and to adapt the product accordingly, as well as to regularly evaluate already existing products for their possibly

[5] I am aware that a shared intention is often present even in mediated interaction. For example, if a group has the goal of harming somebody by insulting he or him online, then there is a shared intention to act. What I am getting at here is that somebody might also, for instance, 'like' and 'share' some of these insults out of different, though possibly still malicious, intentions. For instance, if the content is obviously insulting and somebody 'likes' and 'shares' it out of Schadenfreude, but with no intention of harming a real person – they might, for instance, assume that the person depicted is generic – this act is morally questionable but the person sharing and liking the content is not participating in a cooperative action with the original group that set up the insulting content.

unintended side effects and to improve them if necessary. In addition, the product should be tested on a small scale before use and all groups potentially affected by the use of the product, such as a platform, should be involved in the planning process.

Last but not least, it is worth noting that technologically mediated interactions do not take place in a vacuum, but in a societal and political framework. What about creating the legal framework that, for instance, defines mandatory certification systems that make it easier to understand which platform is trustworthy and which is not, prohibits problematic products as well as problematic ways of using products and ensures a certain level of consumer protection? In other words: Where does the political responsibility to set up a framework for the interactions begin, and where does it end?

In the face of all these problems and difficulties: Should we abandon the concept of responsibility in the age of digitalization? I do not think so. Although artefacts "may influence and shape human action, they do not determine it" (Noorman 2020). Although it might be hard to attribute responsibility, it is not impossible – in many cases, one might even argue, a digitalized society makes it all too easy to trace back who has done what and contributed to a given outcome in what way. The problem of distance through time and space also arises with regard to many other ethical issues and is not particular to responsibility under the conditions of an increasingly datafied and digitalized world. This issues is also under debate for instance with regard to global poverty relief and global distributive justice, and climate change mitigation.

The concept of responsibility is not rendered superfluous simply because it is sometimes difficult to determine who is responsible for what and in what way and to whom. If anything, the many layers to the concept show how relevant the concept is and that we cannot (and ought not) side-step it easily.

4. An ethical framework of responsibility for diversity-sensitive social platforms

In what follows, I do not propose a single unified concept of responsibility, but rather start from the observation that there are various dimensions of responsibility (cf. Höffe 2008, 326) and that we need a multidimensional concept of responsibility to capture all the relevant layers of the notion when dealing with a complex technological artefact that mediates complex human interaction patterns.

The sketch consists of seven main points: First, there is no such thing as "ethical neutrality" with regard to social platforms. Second, I stress the proactive side of responsibility (cf. Gotterbarn 2001), in particular with regard to social platforms. Third, I propose that in a computerized age we have to be not less, but rather more person-centred. This leads us, as the fourth point, to a positive conception of responsibility. Furthermore, a heightened sense of professional responsibility is warranted with regard to diversity-sensitive social platforms. Finally, the professional responsibility needs to be framed by appropriate standards and regulations.

4.1. No ethical neutrality

It is often argued that platforms or technologies in general are 'ethically neutral', that is, that they do not come with a perspective on what is right and wrong. Or, as Martin observed about the debate: "algorithms are implemented with the hope of being more neutral, thereby suggesting that the decisions are better than those performed solely by individuals. By removing individuals from decisions [...] algorithmic decisions are framed as less biased without the perceived irrationality, discrimination, or frailties of humans in the decision" (Martin 2019, 837).

By now quite a few authors have criticized the view that computing is an ethically neutral practice, or that there is such a thing as an ethically neutral technology at all (cf. Akrich 1992; Bijker 1995; Gotterbarn 2001; Friedman and Nissenbaum 1996; Latour 1992; Martin 2019; Winner 1980). Without wanting to decide the much wider debate on ethical neutrality here, I take the view that ethical neutrality cannot be assumed with regard to social platforms.[6] At the same time, I hold the view that ethical non-neutrality does not depend solely on the possible or factual consequences of the deployment of a technology.

By inviting specific forms of action and prohibiting others, algorithms have an enormous impact on the behaviour of humans (cf. Verbeek 2006, 377; Martin 2019, 836). That they do have an impact, however, is not sufficient to make that impact ethically relevant. Let me elaborate this point a little: According to Martin, "algorithms are not neutral but value-laden in that they (1) create moral consequences, (2) re-enforce or under-cut ethical principles, or (3) enable or diminish stakeholder rights and dignity" (Martin 2019, 838). This definition, however, is too broad to define ethical non-neutrality, because it is also true for natural phenomena, like lightning. Lightning might create morally relevant consequences by destroying property and killing people, it can under-cut ethical principles like, for instance, the protection of property, and one might want to argue that it also diminishes possibly a stakeholder's bodily integrity to be struck by lightning. Thus, clearly, algorithms potentially have an impact on our lives, but so does lightning.

Kramer, van Overheld and Peterson in my eyes capture the point about ethical non-neutrality more precisely than Martin, when they argue that some algorithms "cannot be designed without implicitly or explicitly taking a stand on ethical issues, some of which may be highly controversial" (Kramer, van Overheld and Peterson 2010, 251). What Kramer et al. stress is that it is humans, for instance the programmers and designers, that here take a stand on sometimes highly controversial matters – a stand that is very often not transparent to the users of that technology. As Mittelstadt et al. stress, algorithms are in fact intended to "privilege some values and interests over others" (Mittelstadt et al. 2016; cf. Martin 2019, 839), and since "[a]lgorithms are implicitly or explicitly designed within the framework of social customs" (Capurro 2019, 132), they are in fact in many ways (explicitly or implicitly) value laden. Algorithms are, thus, ethically non-neutral not because of the possibly morally relevant conse-

[6] Also compare Gillespie's illuminating discussion of platforms as constituents of public discourse, thus leaving their previously assumed role as mere intermediaries behind (Gillespie 2018a).

quences of their use, but because of the value-decisions that are already (necessarily) inscribed in their set up, regardless of whether they would ever generate negative consequences, even regardless of whether they would ever be put to use.

The non-neutrality thesis, by the way, also holds for supervised and unsupervised machine learning systems: If a machine is trained on biased data, it will 'learn' the biases from the data set and reproduce them in the decision-making process. To give an example from picture analysis and image search that I have discussed already in more detail elsewhere: In 2016, a tweet by Karbir Alli about the outcome of a search for "three black teenagers" in Google Images lead to a storm of protest (Beuth 2016), because it presented mugshots. A search for "three white kids" in contrast showed mostly pictures of happy white kids (cf. Zuiderveen Borgesius 2018, 16). Google's response to that was that these search results merely reflected the portrayal of the respective subpopulation across the web and the frequency of that portrayal (York 2016). Since the market for images of "three happy white kids" is apparently much bigger than the market for images of "three happy black kids", more stock pictures that are assumed to fit the first label are found on the web and come up as search results more frequently in Google Images (cf. York 2016).

In many ways, then, the search results reflect racism and the manifold injustices of our respective societies, such as the unequal distribution of resources and unequal market access, but also the unwillingness to acknowledge that reproducing the accompanying prejudices is not a neutral position: it is an ethically relevant decision.[7]

4.2. Not only remedial, but also proactive

Often when we talk about responsibility, we talk about it in hindsight. The question then is who is to blame for something that has gone wrong. Indeed, responsibility is closely linked to the idea of accountability and liability. But if responsibility is used synonymously with accountability, we tend to lose an important aspect; that is, the idea of foresight.[8]

Responsibility properly understood not only tells us who is to blame after something has gone wrong, it also demands us to act in a way that prevents bad things from happening.[9] It furthermore demands that we collect the information necessary to make responsible decisions and to learn from any mistakes made. It requires a professional ethics that does not regard the problem to be solved in a given situation as isolated, but takes into account that it is part of a wider context in which the

[7] For a more detailed interpretation and discussion of this example, see Reinhardt (2020a, 273f.).
[8] Based on Ladd (1989), Gotterbarn makes a distinction between positive and negative responsibility that is similar to the idea of remedial and proactive responsibility. Note, that accounts that stress the incentives a strong culture of accountability would provide for a person to act responsibly under the threat of the sanctions that he or she would otherwise face, which tends to put less emphasis on the idea that a person should act responsibly no matter whether anybody could hold that person responsible for his or her actions (Gotterbarn 2001, 226–228).
[9] Höffe (2008) distinguishes three levels of responsibility [Verantwortung]: 1) Aufgabenverantwortung [responsibility for a task] and Zuständigkeitsverantwortung [role responsibility], 2) Rechenschaftsverantwortung [accountability or answerability], 3) Haftungsverantwortung [liability].

proposed solution might have profound consequences – I will elaborate this point in the next section.

Platform developers should in any case implement algorithms that support and strengthen human agency (cf. Hartswood et al. 2016), thus making it easier for users to act responsibly in the above-mentioned person-centred way. Where possible, users should have a say in how the platform is regulated and its transparent working and easily accessible boards of appeal should be in place. Therefore, platforms should establish participatory structures for their users. Also, there have to be monitoring procedures in place that raise awareness of the signs of unintended algorithmic behaviour; for instance, discriminatory decisions. Generally, societies, governments, journalists but also platform owners and operators should invest in increasing data literacy and support users in becoming competent with regard to their online decisions. Establishing standing techno-ethical committees that have an eye on platform operations would also help to detect problems early on (ibid.). Political decision-makers can increase the plurality of platforms by, for example, supporting platform start-ups, thus, increasing the options for users to choose from. [10]The legal framework for social platforms has to be evaluated and updated regularly. Political decision-makers have to examine whether adjustments of civil and criminal law are necessary in certain areas and they have to ensure that the training of legal practitioners is adjusted accordingly. Monitoring and consultation procedures should be established to support the political decision-making process.

4.3. Not less, but more person-centred

The loss of control, autonomy and self-determination in a digitalized world is a common theme in fictional digital dystopias. These often depict a world in which humans lose or have given up everything that makes them capable of acting in an accountable and responsible way (cf. Reinhardt 2020b, 113). The current debate on responsibility in a digitalized world tends to lead away from persons as the ultimate bearer of responsibility:

Algorithms are surrounded by an "air of rationality or infallibility" (Zuiderveen Borgesius 2018, 8). They are commonly trusted because of the belief in their technical 'superiority' in terms of neutrality, objectivity, reliability and accuracy. This puts human decision-makers in a problematic situation: often they have an information deficit and little knowledge of how the system arrives at its results.[11] Prevalent time pressure in many work environments further encourages a pragmatic and approving attitude towards the recommendations of the algorithmic systems deployed (Heesen, Reinhardt and Schelenz 2021, 140).

[10] *Plattform Lernende Systeme* formulated this idea for all systems that use machine learning (Plattform Lernende Systeme 2019).

[11] This fact is often associated with the keywords 'black box', lack of interpretability and explainability or lack of transparency (literature). However, the topic goes much further and also includes fundamental questions about the handling of technical expert systems, their use in decision-making processes and the time and resources we are willing to spend on important decisions, as well as the framework conditions of the work settings that involve algorithmic systems.

In the philosophical debate on responsibility and digitalization, there are now several ways to deal with these and other related observations. These, however, often re-enforce the impression of the supposed powerlessness of the persons involved: One strand in the debate is advancing responsibility from a new angle, namely asking the question whether computers and other computing entities ought to be regarded as moral agents that bear responsibility for their actions. Some authors in fact arrive at the conclusion that we should widen our picture of moral responsibility and expand our understanding of moral agency to computers and other machines (Misselhorn 2019). This is one way in which the current debate leads away from personal responsibility. Another strand is focused on enforcing responsibility, thus bringing about a "strong culture of accountability" (Nissenbaum 1996, 26), in which rule abiding behaviour is much more likely, simply because people would face sanctions. Note that this approach is based on an "instrumental" understanding of responsibility: Ultimately, it is about bringing about a desirable state of affairs, thus, also leading away from persons as bearers of responsibility as the focal point.

I think that we should instead stress the idea of personal responsibility – in particular, in a digitalized world. Responsibility should not only be seen as a means to an end, but rather as a feature of humaneness. The capacity to take on responsibility is an important feature of the human condition. We bring ourselves into this world as moral agents not by shying away from responsibility but by taking it on.[12]

4.4. A positive notion of responsibility

Stressing the concept of personal responsibility also stresses our nature as moral agents and represents an important counterpoint to a debate that focusses on the incomprehensibility and opacity of technological artefacts and the dependence of human decisions on them: Even if my actions are mediated by the interaction protocols on a social platform, I am still responsible for my actions and, as a consumer or user, I also bear a responsibility for my choices. So, for example, there is no context in which participation in cyberbullying, hate speech or any other harmful – and sometimes even illegal – activity is ever okay, simply because it is conducted online. From an ethical perspective, this is a trivial fact. Our current practices, however, seem to speak a different language.[13] A positive notion does not reduce responsibility to accountability and liability for wrongs.

[12] Cf. Höffe's notion of sittliche Verantwortung [moral/ethical responsibility]: "Accepting responsibility is moral [sittlich], as long as one does not take it on because of expected rewards and punishments, but because one recognizes oneself as responsible for fellow human beings, the world and oneself, and acts according to this responsibility as a person" (Höffe 2008, 327; my translation).

[13] One example of legal frameworks going in the direction of stressing personal responsibility is the principle of human final decision of the GDPR: "The data subject shall have the right not to be subject to a decision based solely on automated processing, including profiling, which produces legal effects concerning him or her or similarly significantly affects him or her" (European General Data Protection Regulation §22(1)). This means that algorithmic decisions with the said effects must be professionally assessed by a person before they become effective. With regard to applied ethics, stressing personal responsibility also means that we are not to delegate decisions of moral relevance to machines. Or in other words: "An ethics of AI or algorithms can only be an ethics of those who deal with these techniques" (Lenzen 2020, 64; my translation), and not of ethical decisions by machines.

Taking on a positive concept of responsibility could with regard to social platforms, also entail that individuals or organizations set up collaborative online platforms themselves and develop techniques for fair machine learning via experimental data processing (cf. Veale and Binns 2017; Hagendorff 2019, 55). Watchdog institutions could be established and supported. Exchange forums could help to identify and counteract malpractices. Citizen journalism, blogging or professional data journalism could help with raising awareness about the possible effects of our conduct online and to the possible misconduct of and via platforms (cf. Heesen, Reinhardt and Schelenz 2021, 141).

4.5. A heightened sense of professional responsibility

There may be areas where computerized algorithmic decisions and AI applications have little or no effect on human interactions. The opposite is true for social platforms. They are aimed at people. Therefore, developers need to be aware that their actions will ultimately affect real people in the real world and they need to take this into account in their design and programming decisions.

What is more with regard to a diversity-sensitive social platform, as explained at the beginning, is that the collection of diversity-relevant data is associated with an increased degree of vulnerability for the users. In order to do justice to this fact, developers must exercise particular caution. These aspects call for a heightened awareness of the professional responsibility of developers.

It is important to note that the professional responsibility of developers cannot be side-stepped with reference to the choices of individuals: Though it is correct that people are (in general) responsible for their own choices and actions, we cannot impose our own duties on others by reference to their free choice. To give an example, securing the privacy of data and information is imperative with regard to the safety of the users of a social platform, in particular a diversity-sensitive one. This obligation should not to be transferred to the individual user via (possibly incomprehensible) general terms and conditions.

Overall, however, it is important that computer scientists and developers are not left alone with the high level of responsibility they bear, but are supported by guidelines and other instruments for implementing professional responsibility. These include legal standards and public control procedures, development standards and certifications, corporate codes, voluntary commitments to orient development work, appropriate protection of whistleblowers (Heesen, Reinhardt and Schelenz 2021, 138) and professional ethical training that raises awareness for ethical issues.

4.6. Framework responsibility

As has already been mentioned, social platforms do not exist in a vacuum but are integrated into social and legal structures. Political decision-makers as well as civil society are called upon to further develop this framework to cover the issues that

may arise from social platforms. Legal standards and binding certifications can set framework conditions for all providers on the platform market. They can provide criteria for trustworthy digital products and provide incentives to develop products accordingly. If necessary, it is advisable to supplement these standards with mandatory audits (Heesen, Reinhardt and Schelenz 2021, 139). Gillespie has enlisted a number of amendments to current law that might be worthwhile to implement with regard to platforms. To mention but a few: Legal standards could entail transparency obligations not only with regard to data being processed, but also on how interactions on the platforms are moderated (Gillespie 2018a, 213f.). There should be minimum standards for content moderation in place. "Platforms could each be required", he adds, "to have a public ombudsman" and an expert advisory panel (ibid., 214). Major platforms could be required to invest in digital literacy programmes "to better address online harassment, hate speech and misinformation" (ibid.).

4.7. Responsibility for diversity on all levels

If a social platform wants to promote diversity and not only collect diversity-sensitive data for exploitative reasons, like for instance micro-targeting, then it should take on responsibility for diversity on all levels: While merely increasing diversity in technical development teams is no guarantee that the whole spectrum of the needs and requirements of different social groups will be taken into account, the composition of a development team can nevertheless help in reflecting the needs and requirements of different user groups (Heesen, Reinhardt and Schelenz 2021, 138). The teams developing and running the platform should consist, therefore, of developers with different perspectives and backgrounds. The platform community should, furthermore, be perceived, acknowledged, appreciated and welcomed as a diverse and heterogenous group. For the training of AI systems that are deployed, it is, moreover, important to ensure representation of the diverse societal groups in the training data (ibid.) and it should be closely monitored whether new grounds of discrimination arise from the specific nature of machine learning (Data Ethics Commission 2019). There should also be a platform constitution or charter in place that acknowledges and welcomes diversity as a core feature of the platform and that bans discriminatory behaviour.

Research and development in the field of algorithmic decision systems and, accordingly, in the field of artificial intelligence and machine learning, is increasingly concerned with diversity and non-discrimination. There are initiatives promoted by various professional associations, such as the Code of Ethics of the Association for Computing Machinery (ACM), or covered in conferences and research contexts, such as "Fairness, Accountability and Transparency in Machine Learning" (FAccTML) or "Discrimination-Aware Data Mining" (DADM). Such forms of diversity-sensitive research take on challenges such as data distortion and stereotyping (cf. Heesen, Reinhardt and Schelenz 2021, 137f.), thus showing an effort to act responsibly with regard to diversity. Some approaches explicitly focus on the concerns of marginalized groups, such as the Design Justice Network (cf. Constanza-Chock 2020).

5. Conclusion: A multidimensional notion of responsibility

Based on the observation that computer science is increasingly interested in issues of diversity and non-discrimination, while at the same time the representation of diversity in, for example, social platforms is linked to some ethical questions, I investigated what an ethical framework of responsibility for a diversity-sensitive social platform might look like. I started by outlining some of the dangers and risks that go hand in hand with making a social platform diversity-sensitive and argued that the increased vulnerability of the users of such a platform is a sufficient reason to take a closer look at the responsibility at stake in this particular context. Then, I went on to show how traditional concepts of responsibility face difficulties when dealing with algorithm-mediated computerized interactions between humans. I proposed that these difficulties make it all the more urgent to think through what responsibility might entail with regard to social platforms. My starting point was an action-centred account of responsibility as opposed to a consequences-centred account.

I then sketched a framework of responsibility for social platforms and proposed a multidimensional understanding of responsibility with regard to diversity-sensitive social platforms. The sketched framework stresses that platforms, in particular social platforms, are not ethically neutral. The responsibility concept we should employ should be person-centred, positive and deal with the remedial as well as the proactive side of responsibility. A social platform in addition calls for a heightened sense of professional responsibility on the side of the people developing, designing, operating and managing the platform. This professional responsibility, however, has to be framed by societal and legal formal and informal rules that give guidance to the individual decisions of the professionals involved. Beyond that, a diversity-sensitive platform has to establish procedures and work environments that are non-discriminatory and foster diversity on all levels – not only within the platform community. Finally, from the above, I arrived at a multidimensional concept of responsibility for diversity-sensitive social platforms that includes but is not limited to the following aspects:

- Professional responsibility: If I help creating technological items and products, I bear the responsibility to ensure, as far as humanly possible, that they will do no harm to people, the environment and democratic structures.

- Tool-mediated responsibility: If I create tools that could foreseeably lead to disastrous consequences, I am responsible for ensuring that the appropriate counter-measures are taken.

- Knowledge-mediated responsibility: I am also responsible for the foreseeable effects of the publication or non-publication of my research.

- Responsibility for my actions: Even if my actions are mediated by interaction protocols on a social platform, I am still responsible for them.

- Consumer/User responsibility: I am responsible for which platform I choose and for considering what consequences and side-effects my actions on a particular platform may have. I am responsible for deciding which tasks I delegate to the platform and for informing myself on how they are carried out — and for adjusting my decision accordingly.

- Framework responsibility: If I am in a position to work towards setting up mandatory and non-mandatory standards and guidelines, curricula for computer scientists, etc. that ensure the non-harmful use of technologies, I bear the responsibility to do so.

If we do not focus on one aspect of responsibility alone, we will get to see the different actors and types of responsibility that are at play with regard to a diversity-sensitive social platform. Through such a positive, proactive and multidimensional understanding of responsibility, we can avoid the impression that responsibility has only to do with the question of who can be called to account if something goes wrong. Instead, the empowering aspect of taking on responsibility should be stressed and a sense of foresight fostered while a dense net of overlapping dimensions of responsibility is woven to ensure that responsibility is taken seriously on all levels and all sides with regard to diversity-sensitive social platforms. The considerations here can, however, only be taken as a preliminary starting point for a much more thorough going debate on responsibility, social platforms and diversity-sensitivity.

In many cases, social platforms function as amplifiers where the single act does no, or little, harm, but the aggregation can do significant harm. Though everybody is responsible for their own wrongdoing, who is responsible for the amplified surplus effect of the aggregation? With regard to social platforms, but not only for social platforms, it might be worthwhile to work out a thorough concept of responsibility for aggregated actions different from the collective responsibility or cooperate responsibility. In future research, it would be, furthermore, useful to see how this framework could be applied to other technologies and what other dimensions of responsibility emerge from such an application, possibly leading to amendments and modifications of the framework.

References

Akrich, Madleine. “The description of technological objects.” In *Shaping technology/building society: Studies in sociotechnical change*, edited by Wiebe E. Bijker and John Law, 205–224. Cambridge, MA: MIT Press, 1992.

Beuth, Patrick. “Nein, die Suchmaschine ist nicht rassistisch.” Published 9 June 2016. Accessed March 30 2021. https://www.zeit.de/digital/internet/2016-06/google-three-black-teenagers-suchmaschine-rassismus.

Bijker, Wiebe. *Of bicycles, bakelite, and bulbs: Towards a theory of sociological change*. Boston MA: MIT Press, 1995.

Brake, David R. *Sharing our Lives Online. Risks and Exposure in Social Media*. New York: Palgrave Macmillan, 2014.

Capurro, Rafael. "Enculturating algorithms." *Nanoethics* 13 (2019): 131–137.

Costanza-Chock, Sasha. *Design justice. Community-led practices to build the worlds we need*, Cambridge, MA: The MIT Press, 2020.

Fisher, John Martin. "Recent work on moral responsibility." *Ethics* 110, no. 1, (1999): 93–139.

Friedman, Batya and Helen Nissenbaum. "Bias in computer systems." *ACM Transactions on Information Systems*, 14, no. 3, (1996): 330–347.

Gillespie, Tarleton. "Platforms are not Intermediaries." *Georgetown Law Technology Review*, 198, no. 2, (2018a): 198–216.

Gillespie, Tarleton. *Custodians of the Internet. Platforms, Content Moderation and the Hidden Rules that Shape Social Media.* New Haven: Yale University Press, 2018b.

Gotterbarn, Donald. "Informatics and professional responsibility." *Science and Engineering Ethics* 7 no. 2, (2001): 221–230.

Gorwa, Robert, Reuben Binns and Christian Katzenbach. "Algorithmic Content Moderation: Technical and Political Challenges in Automation of Platform Governance." *Big Data & Society*, (2020): 1–15.

Hagendorff, Thilo. "Maschinelles Lernen und Diskriminierung: Probleme und Lösungsansätze." *Österreichische Zeitschrift für Soziologie* 44, Supplement 1, (2019): 53–66.

Hartswood, Mark et al. "A Social Charter for Smart Platforms". Published 2016. Accessed 30 March 2021. https://eprints.soton.ac.uk/410307/1/SmartSocietySocialCharterforSmartPlatforms_final.pdf.

Heesen, Jessica, Karoline Reinhardt, and Laura Schelenz. "Diskriminierung durch Algorithmen vermeiden: Analysen und Instrumente für eine demokratische digitale Gesellschaft." In *Diskriminierung und Antidiskriminierung. Beiträge aus Wissenschaft und Praxis,* edited by Gero Bauer, Maria Kechaja, Sebastian Engelmann, and Lean Haug, 129–147. Bielefeld: transcript, 2021.

Höffe, Otfried. "Verantwortung." In *Lexikon der Ethik,* edited by Otfried Höffe, 326–327. München: C.H. Beck, 2008.

Johnson, Deborah, Thomas Power. "Computer systems and responsibility: A normative look at technological complexity." *Ethics and Information Technology* 7 (2005): 99–107.

Jonas, Hans. *Das Prinzip Verantwortung. Versuch einer Ethik für die technologische Zivilisation*, Frankfurt am Main: Suhrkamp, 1984.

Kramer, Felicitas, Kees van Overheld, and Martin Peterson. Is there an ethics of algorithms? *Ethics and Information Technology* 13, 2010: 251–260.

Ladd, John. "Computers and Moral Responsibility: A framework for an ethical analysis." In *The Information Web: Ethical and Social Implications of Computer Networking,* edited by Carol Gould, 207–228. Boulder CO: Westview Press, 1989.

Latour, Bruno. "Where are the Missing Masses? The Sociology of a Few Mundane Artefacts." In *Shaping technology/building society: Studies in sociotechnical change*, edited by Wiebe E. Bijker and John Law, 225–258. Cambridge, MA: MIT Press, 1992.

Lenk, Hans. "Verantwortlichkeit und Verantwortungstypen." In *Handbuch Verantwortung,* edited by Ludger Heidbrink, Claus Langbehn, Janina Sombetzki, 1–29. Wiesbaden: Springer VS, 2016.

Martin, Kirsten. “Ethical Implications and Accountability of Algorithms.” *Journal of Business Ethics* 160, (2019): 835–850.

Matthias, Andreas. “The responsibility gap: Ascribing responsibility for the actions of learning automata.” Ethics and Information Technology, 6 (2004): 175–183.

Mensi-Klarbach, Heike. “Der Business Case für *Diversität und Diversitätsmanagement*”. In Diversität und Diversitätsmanagement, edited by Regine Bendl, Edeltraud Hanappi-Egger, Roswitha Hofmann, 299–326. Wien: facultas, 2012.

Mittelstadt, Brent Daniel, Patrick Allo, Mariarosaria Taddeo, Sandra Wachter, and Luciano Floridi. “The ethics of algorithms: Mapping the debate.” *Big Data & Society* 3, no. 2, (2016): 1–21.

Misselhorn, Catrin. *Grundfragen der Maschinenethik*, Stuttgart: Reclam, 2019.

Nida-Rümelin, Julian. *Verantwortung*. Stuttgart: Reclam, 2011.

Nida-Rümelin, Julian. *Humanistische Reflexionen*. Berlin: Suhrkamp 2018.

Nissenbaum, Helen. “Computing and Accountability.” *Communications Association for Computing Machinery* 37, no.1, (1994): 72–80.

Nissenbaum, Helen. “Accountability in a Computerized Society.” *Science and Engineering Ethics* 2, no. 1, (1996): 25–42.

Noorman, Merel. “Computing and Moral Responsibility.” In: *The Stanford Encyclopedia of Philosophy* (Spring 2020 Edition), edited by Edward N. Zalta. Accessed 30 March 2021. https://plato.stanford.edu/archives/spr2020/entries/computing-responsibility/.

O’Neil, Cathy. *Weapons of Math Destruction. How Big Data Increases Inequality and Threatens Democracy*. London: Penguin, 2016.

Parekh, Bhikhu. *Rethinking Multiculturalism: Cultural Diversity and Political Theory*. London: Macmillan, 2000.

Plattform Lernende Systeme. “Innovation nutzen, Werte schaffen. Neue Geschäftsmodelle mit Künstlicher Intelligenz Bericht der Arbeitsgruppe Geschäftsmodellinnovationen.” Accessed March 30 2021: https://www.plattform-lernende-systeme.de/ files/Downloads/Publikationen/AG4_ Bericht_231019.pdf.

Reinhardt, Karoline. “Between Identity and Ambiguity. Some Conceptual Considerations on Diversity.” *Symposion* 7, no. 2, (2020a): 261–283.

Reinhardt, Karoline. “Digitaler Humanismus. Jenseits von Utopie und Dystopie.” *Berliner Debatte Initial* 31, no. 1, (2020b): 111–123.

Schmidt, Jan-Hinrik. “Ethik des Internets.” *Handbuch Medien- und Informationsethik*, edited by Jessica Heesen, 284-292. Stuttgart: J.B. Metzler, 2016.

Smiley, Marion. “Collective Responsibility.” *The Stanford Encyclopedia of Philosophy* edited by Edward N. Zalta, (Summer 2017 Edition). Accessed March 30 2021. https://plato.stanford.edu/archives/sum2017/entries/collective-responsibility/.

van de Poel, Ibo, Lambèr Royakkers, Sjoerd D. Zwart. *Moral Responsibility and the Problem of Many Hands*, London: Routledge, 2015.

van Dijk, José, Thomas Poell and Martijn de Waal. *The Platform Society. Public Values in a Connective World*. New York: Oxford University Press, 2018.

Veale, Michael and Reuben Binns. “Fairer machine learning in the real world. Mitigating discrimination without collecting sensitive data.” *Big Data & Society* 4, no. 2, (2017): 1–17.

Verbeek, Peter-Paul “Materializing Morality: Design Ethics and Technological Mediation.” *Science, Technology, and Human Values* 31, no. 3, (2006): 361–380.
Weizenbaum, Joseph. *Computer Power and Human Reason. From Judgement to Calculation.* San Francisco: W.H. Freeman and Company, 1976.
Winner, Langdon. “Do Artifacts Have Politics?” *Daedalus* 109, no. 1, (1980): 121–136.
York, Chris. “Three black teenagers: Is Google Racist? It’s not them, it’s us.” Published 8 June 2016. Accessed March 30 2021. https://www.huffingtonpost.co.uk/entry/three-black-teenagers-google-racism_uk_575811f5e4b014b4f2530bb5, published June 8 2016.
Young, Iris. *Justice and the Politics of Difference*. Princeton: Princeton University Press, 1990.
Zuboff, Shoshana. *The Age of Surveillance Capitalism. The Fight for a Human Future ar the New Frontier of Power*. London: Profile Books, 2019.
Zuiderveen Borgesius, Frederik. “Discrimination, artificial intelligence, and algorithmic decision-making.” published by Directorate General of Democracy, Council of Europe, Strasbourg 2018.

Ethical and legal implications of using AI-powered recommendation systems in streaming services

Recommendation engines are commonly used in the entertainment industry to keep users glued in front of their screens. These engines are becoming increasingly sophisticated as machine learning tools are being built into ever-more complex AI-driven systems that enable providers to effectively map user preferences. The utilization of AI-powered tools, however, has serious ethical and legal implications. Some of the emerging issues are already being addressed by ethical codes, developed by international organizations and supranational bodies. The present study aimed to address the key challenges posed by AI-powered content recommendation engines. Consequently, this paper introduces the relevant rules present in the existing ethical guidelines and elaborates on how they are to be applied within the streaming industry. The paper strives to adopt a critical standpoint towards the provisions of the ethical guidelines in place, arguing that adopting a one-size-fits all approach is not effective due to the specificities of the content distribution industry.

Keywords: *recommendation system, streaming service, audiovisual media, fundamental rights*

Author Information

Kinga Sorbán, University of Public Service, Institute of the Information Society
https://orcid.org/0000-0002-9288-7897

How to cite this article:

Sorbán, Kinga. "Ethical and legal implications of using AI-powered recommendation systems in streaming services."
Információs Társadalom XXI, no. 2 (2021): 63–82.
https://dx.doi.org/10.22503/inftars.XXI.2021.2.5

1. Introduction

In 2007, a new kind of service was introduced in the United States that allowed members to watch movies and television shows instantly in an online environment (URL 1). The service was called Netflix and initially launched with 1000 titles and with the aim to become a competitor of traditional DVD-rental services. As we now know, the new business model revolutionized the entertainment industry, allowing users to access a massive catalogue of audiovisual works from the comfort of their couches. Demand has grown massively year on year, with Netflix re 203.66 million subscribers in 2021 (URL 2). The service has not only proven to be a wildly successful business model, but streaming content online has become a cultural phenomenon that has even given rise to its own slang terminology, consisting of terms such as 'Netflix and chill' and 'binge-watching' (URL 3). Binge-watching is a term that refers to the phenomenon of watching multiple episodes of a television programme in rapid succession (URL 4). This is a common practice; according to a survey conducted in 2013, 61% of Netflix users binge-watch TV series regularly (URL 5). Several research studies have indicated that binge-watching can be a harmful phenomenon at the level of the individual, stating that such a viewing pattern may lead to addiction symptoms (Riddle et al. 2018) similar to "other behavioural addictions, such as loss of control and pleasure anticipation" (Forte et al. 2021, 1) but also depression and polarization. However, the interest of the streaming service providers is to fuel binge-watching to enhance users' screen time and to develop stronger user engagement, all with the purpose of realizing more revenue. Among various other tools, stronger user engagement is achieved by recommending more content to watch; preferably content that spikes the individual user's interest, leading to their further consumption. It is not a coincidence that when we watch content on a platform (not only on Netflix but on Hulu, Amazon Prime and even YouTube) we keep bumping into other interesting content. Sometimes we may feel that service providers are reading our minds and know exactly what we want (or what we think we want). However providers are not using some mind-reading magic, they are using something perhaps even better: recommendation engines. Recommendation engines are commonly used in the entertainment industry to keep users glued in front of their screens. These engines are becoming increasingly sophisticated as machine learning tools are being built into ever-more complex AI-driven systems that enable providers to effectively map user preferences. The utilization of AI-powered tools, however, has serious ethical and legal implications, and not just exclusively limited to the field of content distribution. Some of the emerging issues are already being addressed by ethical codes, developed by international organization and supranational bodies. The present study aimed to address the key challenges posed by AI-powered content recommendation engines. Consequently, this paper introduces the relevant rules present in the existing ethical guidelines and elaborates on how they are to be applied within the streaming industry. The paper strives to adopt a critical standpoint towards the provisions of the ethical guidelines in place, arguing that adopting a one-size-fits all approach is not effective due to the specificities of the content distribution industry.

2. "We have to go back" – A brief history of recommendation systems

Recommendation engines are not new inventions, actually they have been around for almost two decades. Knowing how these recommendation systems work can enable us to map the issues that may arise due to their widespread application. It is important to note that this study only gives an outline sketch of the functioning of these systems, it does not strive to give a specific or comprehensive analysis. Descriptions thus may be restricted to general outlines of the concepts discussed, but this level of interpretation should provide the reader with the necessary background to assess the extent of the issues related to AI-powered recommendation systems. At a basic level, recommendation systems consist of machine learning algorithms, which are a subset of AI. "Machine learning is a method of data analysis that automates analytical model building. It is a branch of artificial intelligence based on the idea that systems can learn from data, identify patterns and make decisions with minimal human intervention" (URL 6). The machine learning algorithms used to build recommendation systems can be categorized based on the method they use for filtering. Three categories can be identified from the perspective of the filtering method: content-based, collaborative and knowledge-based filtering. The first recommendation systems implemented content-based filtering, while the more advanced ones applied collaborative-filtering methods. Further, the early recommendation engines relied solely on user ratings when making suggestions.

2.1. Content-based filtering

Content-based filtering is a filtering method based, on the one hand, on assigning features/descriptors to every content in the database and, on the other hand, on profiling a user's behaviour using data extracted from the explicit user contributions (rating previously accessed titles or keyword searches). In this model, the user feeds the recommendation system with relevant information, which is used to generate a user profile. The recommendation system assigns items to the list of search results or recommended titles if there is a match between the descriptive attributes of certain media content (movies or series) and the characteristics of the media content used to build the user profile. Aggarwal describes the operation of these systems through providing the example of a user called John who gave a high rating to the movie *Terminator* (Aggarwal 2016, 14). As the descriptors of *Terminator* match the majority of the genre keywords for *Alien* and *Predator*, these movies will be recommended to John (Aggarwal 2016, 14). The following figure illustrates how the system works:

The advantage of content-based recommendation systems is that they do not need a particularly large dataset, as the recommendations are specific to a certain user. These systems are, however, limited, as they only recommend items with similar properties. Sticking to the example of John, if he prefers the genres science fiction, horror and action, he will never be recommended *The Crown*, a historical drama series. This is considered to be a disadvantage as it "tends to reduce the diversity of the recommended items" (Aggarwal 2016, 15).

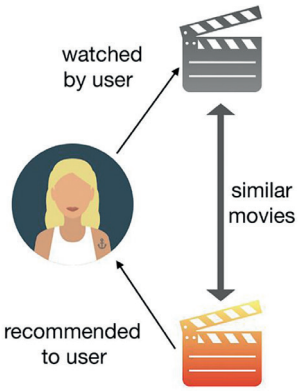

Figure 1. Example of a content-based recommendation system (Source: URL7)

2.2. Collaborative filtering

Collaborative filtering utilizes other users' profiles when making recommendations. The main idea behind these systems is that when users have displayed similar interests and rating patterns in the past, it is likely they are going to have similar preferences in the future. In the 1990s, there were attempts to predict user preferences in order to tailor search result lists and recommendations with collaborative-filtering methods. One of the first recommendation engines was GroupLens, which was used as collaborative-filtering system for Usenet news. A pilot trial was performed which started by inviting users from selected newsgroups to rate pieces of news on a scale of 1–5 (Konstan et al. 2000). Ratings then were used to generate predictions embedded into the recommendations. Later, service providers enhanced their systems, adding more relevant factors to the assessment matrix. For instance, besides active and explicit user contributions (ratings), providers started to incorporate active but implicit contributions to their recommendation engines (such as an assessment of the user's clicking history or watching time).

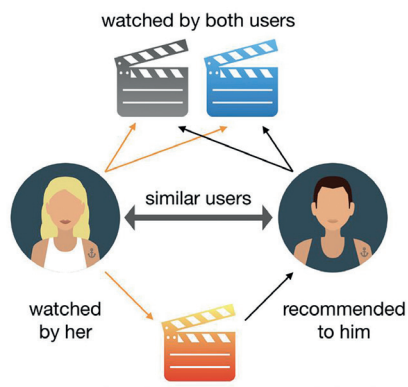

Figure 2. Example of a collaborative recommendation system (Source: URL 7)

Naturally, this is also a simplified description of collaborative filtering, but for the purpose of this study, it is enough. Modern recommendation engines do not use a clean version of the described filtering methods; indeed, the majority of the existing recommendation engines combine elements of different filtering techniques. These recommendation engines are often called hybrid recommendation systems.

3. "The truth is out there" – How AI-powered recommendation engines work in a nutshell

At this point one may ask why are AI-powered recommendation systems the focus of interest all of a sudden, if the base technology, i.e. machine learning algorithms, has been around for decades? Perhaps because AI has only recently reached the level of development that makes their functioning comparable to human thinking and allows them to perform tasks requiring (close-to-) human intelligence. There are many definitions of AI, but there is one common element in all of them: AI should be able to mimic intelligent human behaviour. For the purposes of this study, we use the working definition set out by the European Commission's Communication on AI (URL 8). According to the proposal:

> *"Artificial intelligence (AI) refers to systems that display intelligent behaviour by analysing their environment and taking actions – with some degree of autonomy – to achieve specific goals. AI-based systems can be purely software-based, acting in the virtual world (e.g. voice assistants, image analysis software, search engines, speech and face recognition systems) or AI can be embedded in hardware devices (e.g. advanced robots, autonomous cars, drones or Internet of Things applications)."*

However, AI is not a uniform technology, in fact there are several categories of AI, depending on their level of independence and scope of functioning. Colloquially, AI is often identified with what Kurzweil called "strong artificial intelligence (SAI)" (Kurzweil 2005). Here, SAIs are considered intelligent agents that are able to perform any intellectual task a human, in other words, it refers to conscious machines with full human cognitive abilities (URL 9). However, it has been stated that when singularity is reached, machine intelligence will exceed human intelligence; thus humans will become unable to understand and control technological development (Kurzweil 2005). Although technology can develop at a frightening speed, SAI yet remains within the domains of science fiction. The majority of AIs currently in use correspond to the notion of narrow artificial intelligence (NAI), as also introduced by Kurzweil (URL 10). NAI systems are only capable of performing specific tasks – albeit with high efficacy – but they lack the cognitive complexity of human thinking.

Due to the exponential development in the areas of algorithm programming, computational power and the availability of massive, easily accessible and transferable data pools, AIs have undergone unprecedented development in the past couple of years. Advances in building deep neural networks have led to the invention of

deep learning, which is a subcategory of machine learning. The connection between AI, machine learning and deep learning are shown in the figure below:

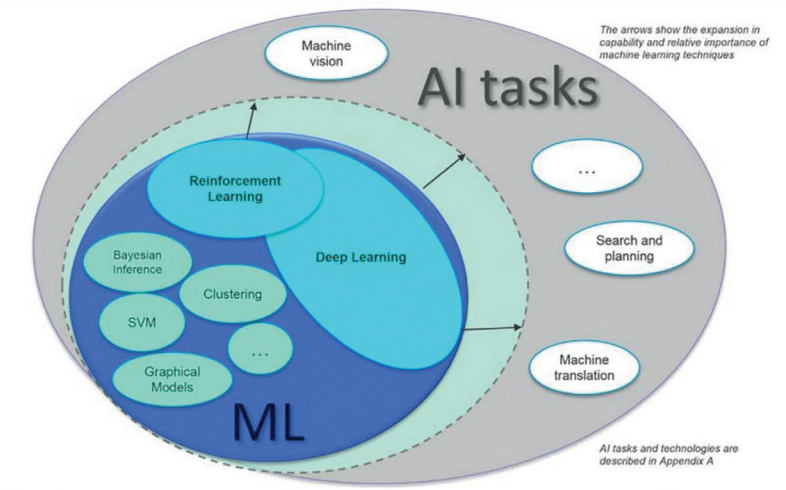

Figure 3. The connections between AI, machine learning and deep learning. Source: (Cambridge Consultants 2019)

Deep learning is a machine learning technique that uses artificial neural networks, which mimic the structure of the human brain (Cambridge Consultants 2019, 20). Deep learning techniques are now applied within the most advanced recommendation systems, like Netflix's.

4. "Winter is coming" – The risks recommendation systems pose to certain fundamental rights

A key aim of this study was to assess the ethical issues related to recommendation systems through the lens of human rights. A human-rights-based approach is relevant, because fundamental rights are core values that are recognized globally and are set out by various international legal instruments. Furthermore, legal instruments adopted within the European Union, for instance, are all rooted in these fundamental rights, and all the institutions of the EU and its Member States are bound to abide by Charter of Fundamental Rights of the EU (2012). The human-rights-based approach is favourable for one more reason: the EU's Ethics Guidelines for Trustworthy AI (AI HLEG 2019) and the proposed AI Code (URL 11) share the view that a human-rights-driven approach is the key to building trustworthy AI. There are some approaches that take into account the human rights dimensions of recommendation systems (but do not focus solely on the fundamental rights aspects of AI-powered recommendation systems). Milano et al. suggest a taxonomy in which recommendation systems are categorized along two dimensions (Milano et al. 2020). The first dimension catalogues the risks identified in connection to recommendation systems

based on whether they negatively affect the utility of some stakeholders or constitute a rights violation. The second dimension categorizes risks based on the severity of the impact: some may cause immediate harm, while some only cause an exposure to the future risk of harm. This study focuses solely on the rights dimension of the recommendation system-related risks. It has to be noted though that some overlaps may be observed between some elements of these categories. For example, inaccurate recommendations are normally considered a utility issue, yet if they persistently appear on a larger scale, there is an inherent risk of rights violation (such as unfair treatment). In 2018, controversy revolved around Netflix personalizing movie posters shown to users. The company was accused of personalizing the movie poster selection based on ethnicity, as black viewers were presented posters featuring black cast members (URL 12). What seemed to be at first just a flaw in the algorithmic design turned out to be a risk to the principles of equality and non-discrimination.

Fundamental rights are impacted by AI-powered recommendation systems extensively; the following rights are particularly affected:

- Dignity – human autonomy
- Integrity
- Privacy
- Freedom of information
- Equality, non-discrimination
- Diversity.

Projecting the ethical issues identified through a literature review onto the catalogue of fundamental rights laid down in international legal instruments (especially in the Charter of Fundamental Rights of the European Union), the following risk map was drawn up. However it must be borne in mind that the identified risks intertwine and some of the identified risks affect different rights (especially the lack of transparency).

Affected right	Risk
Dignity – human autonomy	Lack of transparency – black box
Integrity	May lead to addiction Inappropriate content
Privacy	User profiling and data leakage Data publishing Algorithm design User interface design Experimentation on user groups
Freedom of information	Filter bubble Lack of transparency – black box
Equality, non-discrimination	Activity bias Algorithmic bias Cognitive bias of the user
Diversity	Lack of transparency – black box with lack of diversity in the recommendations

Table 1. Risk map of AI-powered recommendation systems

4.1. Dignity – human autonomy

Human autonomy when the user browses the system to choose media content to consume is a mirage to some degree, because the options recommended by the algorithm are filtered media content, deemed to be relevant for the individual user by the recommendation system. Such filtering limits the list of available options, thereby in reality curbing the users' freedom of choice.

The functioning of recommendation systems remains hidden from the user, with such tools operating in the background, unnoticed as part of the user experience. Algorithmic decision-making tools – not limited to recommendation systems – in fact work as a 'black box' (Pasquale 2015) system, where users cannot tell why and how the system generated one specific output. One reason behind this is the fact that we are talking about sophisticated program codes and complex mathematics often guarded as trade secrets of the provider. One may argue that the release of the source codes does not really carry relevant information to the general public. The European Union Agency for Fundamental Rights' (FRA) report highlights that one aspect of preserving human dignity is to inform people about the use of AI, enabling them to provider informed consent (FRA 2020, 60). Transparency in the case of AI thus equals explicability: Knowledge of the principles of functioning and the factors that were taken into consideration when compiling the list of recommended content that would be enough to contribute to user awareness and to allow them to reach informed consent.

4.2. Integrity

According to the Charter of the Fundamental Rights of the EU, the right to integrity of the person means respect for one's physical and mental integrity. Research shows that the use of recommender systems paired with psychological factors, such as a lack of self-control or lack of self-esteem complemented with certain motives (the motive of information seeking) can lead to excessive usage (Hasan et al. 2018). Excessive internet usage and content consumption are known can have negative impacts on individuals' psychosocial well-being (Young 2004.), and can have negative consequences on individuals, such as emotional problems, relational problems, sleep-difficulties and performance problems (Andreassen 2015). Recommender systems that are designed to manipulate users – sometimes with subliminal techniques – thus may adversely affect the users' mental integrity.

Inappropriate content (Milano et al. 2020) may also have negative impacts on a person's integrity, although there is no common understanding what the term 'inappropriate content' means. Inappropriate could mean content that is erroneously suggested, contradicting a user's preferences and predictions reached by analysing the factors assessed by the recommendation system. Inappropriate can also mean that recommendations are not culturally appropriate for the individual users or certain user groups (Souali et al. 2011). This was the case when the Christian community of Brazil petitioned to remove the movie titled *The First Temptation of Christ* from Net-

flix's catalogue because it portrayed Jesus as a homosexual figure (URL 13). Inappropriate can be also interpreted in terms of certain vulnerable audience groups, such as minors. The popular American teen drama titled *13 Reasons Why* was severely criticized for romanticizing suicide, yet it appeared on the lists of recommendations of teenage users. The show is very popular among adolescents, although it could pose a serious risk for mentally unstable people, or people with mental health issues (URL 14). There are rules in the European audiovisual media regulation which aim to restrict the free flow of content that is detrimental to minors. For instance, content that can cause serious harm to the physical, moral or mental development of children should be restricted to adult audiences. This is an obligatory provision of the AVMS Directive since 2010 and applies to on-demand streaming providers as well, pursuant to which they must tune their recommendation systems to take into consideration the user's age and the parental control settings.

4.3. Privacy

Privacy is one of the key challenges identified and the protection of personal data is the most cited fundamental right in the AI-related discourse. The right to privacy has paramount importance in the case of recommendation systems, which profile users to fine-tune content recommendations. The factors that are taken into consideration by recommendation systems – introduced in Section 1. – are designed with regard to the availability of user data. There are five problematic areas that were identified in relation to recommendation systems by Paraschakis, three of which (1. 2. and 5.) are privacy related (Paraschakis 2017):

1. user profiling and data leakage
2. data publishing
3. algorithm design
4. user interface design
5. experimentation on user groups.

Paraschakis draws attention to the fact that behavioural profiling is often done without acquiring informed consent, as privacy notices hidden behind hyperlinks that follow "I consent" checkboxes often remain unread by the user (Paraschakis 2017). Unsolicited data collection is also common, because user profiles are generally enhanced with data obtained from external sources, such as cookies, social networks or information brokers, despite the fact that the integration of external sources can lead to vulnerability and could lead to data breaches. Friedman et al. consider the actions of external adversaries who attempt to de-anonymize data one of the biggest privacy-related risks to recommendation systems (Friedman et al. 2015). Paraschakis adds that companies often release large datasets from their services that contain private data. Although personally identifiable information (such as names and email addresses) are anonymized, there are quasi-identifiers (birth date, gender, location) that can be combined to identify users. Companies

also often test new versions of their recommendation systems on randomly selected user groups. The most popular method for testing new algorithms is A/B testing, in which two variants of the same webpage is shown to different user groups (a control and treatment group). A famous example of such an experiment is research conducted through Facebook as part of an emotion experiment in 2014 (Kramer et al. 2014). In the experiment, members of different user groups were shown differently curated news feeds. Users who were shown more positive posts reported feeling happier, while the people who had seen negative images more frequently felt unhappy and showed signs of depression. Besides the issue of the unethicalness of the experiment, the research also drew attention to the lack of informed consent of the users, who weren't informed prior to the experiment that they were part of such a research, or of the handling of their personal data for the purposes of the research.

4.4. Freedom of information, freedom of expression

Freedom of expression involves freedom of information and means the right to receive and impart information and ideas without interference. Yet, the main objective of recommendation systems is to interfere with the flow of information by selecting relevant information (or information considered relevant) for each individual user, with the aim to enhance the user experience and promote engagement. By consuming overly-personalized media content, users can easily become isolated from media content that is outside their comfort zone and ideas different from their own ideology, resulting in them getting stuck in cultural and ideological bubbles, named filter bubbles (Pariser 2011.) The user may not necessarily notice getting into such a filter bubble due to the elaborate design of the filtering system and the lack of transparency.

It is important to examine the operation of recommender systems from the perspective of the content creators (artists) as well. Streaming services are generally good platforms for emerging creators and narrow-niche genres, because streaming service providers facilitate the worldwide (or regional) distribution of audiovisual works of any genres, and as audiences can find media content that is often not available through traditional distribution. Due to the long-tail effect (Anderson 2006), offering niche content is profitable for streaming service providers. However, these providers are also in a gatekeeper position, in that they have a direct impact on the media content they make available to the public (Koltay 2019, 82). This means that streaming service providers can arbitrarily control what gets popular and what gets lost among the myriad of content, reducing the visibility of certain creators and widening the gap between well-known global studios and smaller studios making art films or niche movies. We do know that some creators sign deals with streaming providers to produce, to distribute and even to feature their works, resulting in a further imbalance between the larger and richer and well-known outlets and the small studios. In this system, service providers have little to no regard to the credibility of the information conveyed, blurring the boundaries

between fictional works and documentaries, which carries the risk that, due to the fact that they can reach millions of people, they can amplify the spread of false or misleading information. Netflix received harsh criticism for signing a deal with Gwyneth Paltrow's Goop, which is known to feature a pseudoscientific lifestyle documentary (URL 15), but was also criticized for producing a similar documentary series titled *Down to Earth* with Zac Efron (URL 16), and for the documentary *Seaspriacy*, which was accused of containing misleading claims about commercial fishing (URL 17).

4.5. Equality, non-discrimination

The Charter of Fundamental Rights of the EU (2012) considers all people equal and sets out that any discrimination based on protected characteristics – such as sex, race, colour, ethnic or social origin, genetic features, language, religion shall be prohibited– (Article 21). Furthermore, it sets out that the EU shall respect cultural, religious and linguistic diversity (Article 22). However, recommendation systems are biased by design, as they draw up patterns, which they then use to generalize users. The FRA's report also draws attention to the fact that the "very purpose of machine learning algorithms is to categorize, classify and separate" (FRA 2020, 68). Baeza-Yates differentiates between three types of biases that characterize recommendation systems and distort the list of recommended media content (Baeza-Yates 2020):

1. Activity bias, which refers to the distorting effects of the attributes that are automatically assigned to users upon browsing and searching, such as gender, age, location, language of the service.

2. Algorithmic bias, which refers to the distortion that can be traced back to the programming of the algorithm. Recommendation systems work with sets of variables, weigh each factor and rank each property differently, where the principles of weighing and ranking are coded into the system by programmers having their own biases. One form of algorithmic bias is observation bias (Farnandi et al. 2018), which refers to the feedback loops generated to specific groups of users. The term "feedback loop" has been used in software development for some time now and it refers to a situation where the outputs of a system are loaded back to be used as inputs. Using the outputs generated by the system as teaching data amplifies bias and leads to the development of filter bubbles (Mansoury et al. 2020; Jiang et al. 2019). Observation bias is also caused by population imbalance (Farnandi et al. 2018), whereby existing social patterns are reflected in the system's decisions. Observation bias can mostly occur during the application of collaborative-filtering tools, which rely on interpersonal relationships (other, similar people's preferences) to filter information (Bozdag 2013). Bozdag also mentions popularity bias – which means popular content often gets highlights and thus gets even more popular – hindering the diversity of recommendations.

3. The cognitive biases including confirmation bias and other behavioural biases of the user, which also affect the functioning of the recommender system, as the user is able to tune and teach the system through his or her choices. Unconscious decisions are taken into consideration by the recommender system, leading to the formulation previously discussed filter bubbles. According to Pariser, confirmation bias is often enhanced by personalization algorithms, because consuming information or media content that conforms to one's taste and ideology causes pleasure (Pariser 2011), while diverse opinions and content can lead to cognitive dissonance.

4.6. Diversity

The entertainment industry has often been described as an industry fuelling blockbuster culture. (Anderson 2006.) The term 'blockbuster' has been used for movies since the 1970s (one of the first blockbusters was Steven Spielberg's *Jaws*) and refers to fast-paced and exciting movies, that tend to generate interest beyond the cinema (Shone 2004) and are capable of reaching an extremely wide audience (see the Marvel Cinematic Universe). Although there is a vast amount of available audiovisual works, there are only a limited number of blockbusters. AI-powered recommendation systems may have the effect that they "reinforce the popularity of already popular products" (Fleder et al. 2009, 679), as they are more likely to appear on the top of the list of recommendations and in the list of many difference audience categories. However, it is also argued that these systems can enable members of the audience to find niche content (Anderson 2006).

It is worth highlighting that the promotion of European audiovisual culture with regulatory tools is not unbeknownst in EU law, and in fact it is an obligation already present in the Audiovisual Media Services Directive (AVMSD). The AVMSD explicitly obliges on-demand service providers (such as Netflix and other streaming service providers) to foster the European audiovisual culture and movie industry by reserving at least a 30% share of European works in their catalogues (Article 13). Additionally, the providers also have to ensure the prominence of those works.

Mehrota et al. (2018) point out that modern recommendation systems serve two-sided markets, so algorithms must be optimized in a way to take the interest of the supply side (artists) into consideration as well.

5. "To boldly go…" – Ethical codes as tools to tackle the challenges of AI

The risks of AI have been recognized by several international organizations. In recent years, several AI ethical codes have been drafted to mitigate these risks. Notable examples are the following:

– The Ethics Guidelines for Trustworthy AI of the European Commission's High-level Expert Group (AI HLEG 2019) on Artificial Intelligence;

- The OECD's Recommendation of the Council on Artificial Intelligence (OECD 2019);
- The Beijing AI Principles drafted by the Beijing Academy of Artificial Intelligence (BAAI 2019); and
- Guidelines adopted by the IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems titled "Ethically Aligned Design: A Vision for Prioritizing Human Well-being with Autonomous and Intelligent Systems" (IEEE 2019).

This paper discusses two of these guidelines in detail: HLEG's Ethics Guidelines and the OECD's Recommendation as these instruments are the most relevant from the perspective of the EU's legal framework. Additionally, it has to be noted that all the ethical codes are surprisingly similar to each other as they grasp AI from the same perspective, whereby they place values such as trustworthiness and fairness at the centre. This section not only gives a detailed description of these guidelines, but specifically highlights which provisions are relevant in terms or recommendation systems and why. The second part of this section then provides constructive criticism of the ethical codes discussed.

5.1. How do ethical codes drive the future of AI?

The OECD's Recommendation of the Council on Artificial Intelligence sets out 5 value-based principles for a responsible stewardship of trustworthy AI. The principles for AI are that it should operate in line with:

- The pursuit of inclusive growth, sustainable development and the well-being of humankind.
- Human-centred values and fairness. This principle includes due consideration for the rule of law, human rights and democratic values. In order to abide by this principle, specific mechanisms and safeguards should be implemented into AI systems, such as a capacity for human determination.
- Transparency and explicability.
- Robustness, security and safety.
- Accountability.

These five principles are targeted towards those driving the development of AI, such as those who design and operate systems. To complement these five principles, complementary recommendations were added for policy-makers to take into consideration.

The HLEG's guidelines characterize trustworthy AI as meaning lawful, ethical and robust. The EU's AI ethics principles are rooted in the respect for fundamental rights, and thus the ethical principles set out by the guidelines are based on tangible rights set out by existing international legal instruments. The guidelines list 4 ethical principles, terming them ethical imperatives:

1. respect for human autonomy
2. prevention of harm
3. fairness
4. explicability.

Respect for human autonomy means that humans must be able to maintain self-determination over themselves and are entitled to be protected from manipulation, coercion, deception and conditioning: AI systems thus should refrain from applying techniques that manipulate human beings and should be designed to rather augment, complement and empower human skills. The principle of prevention of harm means that AI systems should be designed to protect human dignity, as well as mental and physical integrity, which is basically the ethical implementation of the first law of robotics, made famous by Isaac Asimov in his work of science fiction *I, Robot* (Asimov 1950). The principle of fairness means – among others – that AI systems should ensure that individuals and groups are free from unfair bias, discrimination and stigmatization. Explicability means that the processes should be transparent and an explanation should be provided why the system reached a particular decision, in order to build and maintain the user's trust.

The issues that were in identified in relation to how recommendation systems affect certain fundamental rights are also risks to the principles set out by the ethics codes. The Ethics Guidelines for Trustworthy AI sets up a non-exhaustive list of requirements to achieve trustworthy AI, which are all relevant to mitigate the risks of AI-powered recommendation systems to fundamental rights as long as the industry players are willing to align their behaviour to them. The list of requirements consists of requirements such as human agency and oversight, technical robustness and safety, privacy and data governance, transparency, diversity. However as argued in the next subsection ethical principles and requirements do not create domain-specific obligations and are not enforceable thus they are not sufficient tools to alleviate concerns.

5.2. Critical remarks concerning the ethical codes

A criticism that is often voiced towards codes of ethics in general can be projected to AI specific codes as well; namely that ethics codes tend to be the results of long negotiations between industry members and the industry members and state actors, where the end text is almost always a result of many compromises. This tends to lead to a set of diluted norms, stripped down to the most important core values and principles recognized by all parties. Setting common values are important elements of marking out a normative framework, but they are not very effective if the norms that should fill the frames are missing.

Given their nature, ethical principles tend to be worded vaguely, making the norms set out overbroad. Here, the values and principles set out by the AI ethical codes are important elements of a normative framework, but not exactly AI specific. Héder notes that the provisions set out in AI ethics guidelines can be applied to any novel technology; for example, if one were to change the term "AI" to the term

"water boiler", the guidelines would still be interpretable (Héder 2020). Indeed, the majority of the discussed principles have been set out already in different scientific areas, such as bioethics. There is also the issue of general documents not being able to react to domain-specific issues in an appropriate manner. NAIs are used in several sectors nowadays, ranging from application in autonomous vehicles to facial recognition. Although they have common elements that can be addressed by horizontal codes of ethics, they also pose special challenges that are not dealt with by general ethics codes. For example, freedom of information is a right that is significantly impacted by AI-powered recommendation systems, meanwhile it is almost completely irrelevant in the case of autonomous weapons systems, in which case the right to life has more relevance. Current ethical codes are thus unable to tackle those issues that stem from the specific functions that AIs applied in specific areas have, thus suggesting regulatory blind spots exist.

As these codes are often too vague, they are unable to influence the signatories' behaviour when it comes to real-life application. As these are soft-law instruments and do not contain tangible obligations, abiding by the norms and the manner in which compliance is realized is largely dependent on the willingness of the members of the industry. The lack of mechanisms for creating compliance is a particular weak point of ethics codes according to many scholars (Hagendorff 2020; Larsson 2020). Without setting out mandatory rules to oblige parties to a certain conduct or to refrain from certain behaviours, there is no way to impose sanctions on those who fail to act in accordance with the principles. The infringement of ethical codes may also result in disadvantage (the disapproval of society, loss of clients, etc.), but these differ from the sanctioning framework set out by legal norms. Larsson's main reason for concern is that it is unclear what the relationship is between ethics codes and pieces of legislation and notes that ethical guidelines are essentially being drafted by industry players with the incentive to avoid stronger state-regulation (Larsson 2020).

6. Concluding remarks – How can recommendation systems promote culture and diversity?

To effectively tackle the domain-specific issues of AI-powered applications and to liquidate the regulatory blind spots identified in Section 5, a more detailed set of norms should be drafted for each individual application domain. Horizontal AI guidelines are too general to tackle the issues that require sectoral tailor-made solutions. In the scientific literature, there have been several suggestions made to address the risks of AI-powered recommendation systems such as those identified in Section 4., which should provide a good starting point for further regulatory initiatives.

Increasing the transparency would be one cure to the majority of the problems discussed, because making recommendation systems more transparent would positively contribute to human autonomy, diversity and privacy. The developers of AI should be mandated to explain how recommendation systems work, what factors they are considering to generate the outputs and how they handle personal data to

generate recommendations. Héder warns that if transparency is not defined at an appropriate level, AI development may be hindered (Héder 2020). Furthermore, he argues that transparency has to be combined with some measure of intelligibility to avoid "pretend transparency". Educating users to move safely and comfortably in the digital world is not a novel thought. The concept of digital literacy was developed as long ago as the 1990s and refers to "the ability to understand and use information in multiple formats from a wide variety of sources when it is presented via computers" (Gilster 1997). Media literacy – which is a narrower concept – refers to those skills that allow users to use the media; for example, to access information and to critically assess media content. Along these concepts, there is a need to introduce the concept of algorithmic literacy (URL 18). Algorithmic literacy should involve the skills to critically assess the recommendations made by algorithms, to exploit the recommendation systems in a manner that serves the best interest of the individual and to develop an awareness of algorithmic biases and how to avoid their influence.

Ensuring a diversity of content is a factor that should be incorporated into algorithms (Castells et al. 2015). The recitals of the AVMSD (recital 35) propose the need to ensure prominence by labelling metadata, facilitating access and setting up dedicated sections in catalogues. The AVMSD does not explicitly mention recommendation systems as a tool to enhance visibility, but does mention that fine-tuning the algorithms to recommend European content can also be considered a viable option to increase the reach of European audiovisual works. Optimizing recommendation systems to promote audiovisual culture and the interests of the supply side would definitely mean a shift from subordinating company policies to audience engagement, and in the long term these measures could contribute to a more diverse media landscape. As members of the audience would be able to find European and national works along with the works of emerging artists, the fine-tuning of recommendation systems could contribute to the promotion of cultural exploration.

Besides enhancing the algorithmic designs, the previously mentioned risks can be mitigated by giving the users more freedom to customize their experience. In the area of privacy, explicit privacy controls should be incorporated into the systems, including allowing the users to decide which data is to be shared and with whom (Paraschakis 2017). Customizable settings should be introduced in filtering as well, because if one can choose to filter adult content (which is what parental control tools are about), one should also be able to select their preferences in terms of other factors (such as genres, actors, directors). More options to customize the service would also contribute to enhancing the autonomy of the user.

References

Aggarwal, C.C. *Recommender Systems: The Textbook*. Springer, 2016.

Anderson, C. *The Long Tail*. Hyperion: New York, 2006.

Andreassen, C.S. "Online Social Network Site Addiction: A Comprehensive Review." *Current Addiction Reports*, 2 (2) (2015): 175–184.

Asimov, I. *I, Robot*. Gnome Press, 1950.

Baeza-Yates, R. "Bias in Search and Recommender Systems." *RecSys '20: Fourteenth ACM Conference on Recommender Systems*, September 2020. https://doi.org/10.1145/3383313.3418435

Beijing Academy of Artificial Intelligence. *Beijing AI Principles*. 2019.

Bozdag, E. "Bias in algorithmic filtering and personalization." Ethics and Information Technology 15, (2013): 209–227. https://doi.org/10.1007/s10676-013-9321-6

Cambridge Consultants: Use of AI in Online Content Moderation 2019 Report Produced on Behalf of Ofcom. 2019.

Castells P., Hurley N.J., Vargas S. "Novelty and Diversity in Recommender Systems." In: Ricci F., Rokach L., Shapira B. (eds) *Recommender Systems Handbook*. Springer, Boston, MA., 2015. https://doi.org/10.1007/978-1-4899-7637-6_26

Fleder, D., & Hosanagar, K. "Blockbuster Culture's Next Rise or Fall: The Impact of Recommender Systems on Sales Diversity." *Management Science*, 55 (5) (2009): 697–712. http://dx.doi.org/10.1287/mnsc.1080.0974 "

Forte, G., Favieri, F., Tedeschi, D., Casagrande, M. "Binge-Watching: Development and Validation of the Binge-Watching Addiction Questionnaire." Behavioral Sciences. 11, no. 2: 27. (2021) https://doi.org/10.3390/bs11020027

Friedman, A., Knijnenburg. B., Vanhecke, K., Martens, L., Berkovsky, S. "Privacy aspects of recommender systems." In: *Recommender Systems Handbook*, 649-688 New York: Springer Science+Business Media, 2015.

Gilster, P. *Digital Literacy*. John Wiley & Sons, 1997.

Hagendorff, T. "The Ethics of AI Ethics: An Evaluation of Guidelines." *Minds & Machines* 30, (2020): 99–120. https://doi.org/10.1007/s11023-020-09517-8

Hasan, M.R., Jha, A.K., and Liu, Y. "Excessive use of online video streaming services: Impact of recommender system use, psychological factors and motives." *Computers in Human Behaviour*, 80 (2018):220–228.

Héder, M. "A Criticism of AI Ethics Guidelines." *Információs Társadalom* 20, no 4. (2020) https://doi.org/10.22503/inftars.XX.2020.4.5

High-Level Expert Group on Artificial Intelligence. Ethics Guidelines for Trustworthy AI. (2019)

The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems. Ethically Aligned Design: A Vision for Prioritizing Human Well-Being with Autonomous and Intelligent Systems (2019).

Ray Jiang, Silvia Chiappa, Tor Lattimore, András György and Pushmeet Kohli. “Degenerate Feedback Loops in Recommender Systems.” In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, Honolulu, USA. 2019.

Koltay, A. *New Media and Freedom of Expression: Rethinking the Constitutional Foundations of the Public Sphere*. Hart Publishing, 2019.

Konstan, J.A. et al. “GroupLens: Applying collaborative filtering to Usenet news.” In *February 2000 Communications of the ACM* 40(3) (2000): 78. https://doi.org/10.1145/245108.245126

Kramer, A., Guillory, JE., Hancock, JT. “Experimental evidence of massive-scale emotional contagion through social networks.” *Proceedings of the National Academy of Sciences*, 111 (24) (2014):8788–8790. https://doi.org/10.1073/pnas.1320040111

Kurzweil, Ray (2005), *The Singularity is Near*. Viking Press, 260.

Larsson, S. “On the Governance of Artificial Intelligence through Ethics Guidelines.” *Asian Journal of Law and Society*, 7(3) (2020):437–451. https://doi.org/10.1017/als.2020.19

Masoud Mansoury, M., Abdollahpouri, H., Pechenizkiy, M., Mobasher B. and Burke R. “Feedback Loop and Bias Amplification in Recommender Systems.” In *CIKM ‘20: Proceedings of the 29th ACM International Conference on Information & Knowledge Management*, 2145–2148. 2020. https://doi.org/10.1145/3340531.3412152

Mehrota, R., McInerney, J., Bouchard, H., Lalmas, M. and Diaz, F. “Towards a Fair Marketplace: Counterfactual Evaluation of the trade-off between Relevance, Fairness & Satisfaction in Recommendation Systems.” In *CIKM ‘18: Proceedings of the 27th ACM International Conference on Information and Knowledge Management*, 2243–2251. 2018. https://doi.org/10.1145/3269206.3272027

Milano S., Taddeo, M. and Floridi L. “Recommender Systems and their Ethical Challenges.” *AI & Society*, 30 (2020):957–967.

OECD. *OECD’s Recommendation of the Council on Artificial Intelligence.* OECD/LEGAL/0449. 2019.

Paraschakis, D. “Towards an ethical recommendation framework.” In *Conference Proceedings 11th IEEE International Conference on Research Challenges in Information Science*, 211–220, 2017. https://doi.org/10.1109/RCIS.2017.7956539

Pariser, E. *The filter bubble: What the internet is hiding from you*. London: Penguin Press, 2011.

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

Riddle, K., Peebles, A., Davis, C., Xu, F., & Schroeder, E. “The addictive potential of television binge watching: Comparing intentional and unintentional binges.” *Psychology of Popular Media Culture*, 7(4) (2018):589–604.

Shone, T. *Blockbuster*. London: Simon & Schuster UK, 2004. 27–40.

Souali, K., El Afia, A. and Faizi, R. “An automatic ethical-based recommender system for e-commerce.” In *2011 International Conference on Multimedia Computing and Systems*, 1–4, 2011. https://doi.org/10.1109/ICMCS.2011.5945631

Young, K. "Internet Addiction: A New Clinical Phenomenon and Its Consequences." *American Behavioural Scientist*, 48 (2004):402–415. https://doi.org/10.1177/0002764204270278

URL 1: McFadden, C. "The Fascinating History of Netflix." Interesting Engineering, Accessed 17. May 2021. https://interestingengineering.com/the-fascinating-history-of-netflix

URL 2: Dean, B. "Netflix Subscriber and Growth Statistics: How Many People Watch Netflix in 2021?" Backlinko, Accessed 17 May 2021. https://backlinko.com/netflix-users

URL 3: McCluskey, B. "9 Netflix slang terms besides 'Netflix and Chill' every binge watcher needs to know." Business Insider, Accessed 17. May 2021. https://www.businessinsider.com/netflix-slang-and-their-meanings-2017-3

URL 4: Merriam-Webster Dictionary. Accessed 17. May 2021. https://www.merriam-webster.com/dictionary/binge-watch

URL 5: West, K. "Unsurprising: Netflix Survey Indicates People Like To Binge-Watch TV." Cinema Blend, Accessed 17. May 2021. https://www.cinemablend.com/television/Unsurprising-Netflix-Survey-Indicates-People-Like-Binge-Watch-TV-61045.html

URL 6: SAS. "Machine learning" Accessed 17 May 2021. https://www.sas.com/en_us/insights/analytics/machine-learning.html#:~:text=Machine%20learning%20is%20a%20method,decisions%20with%20minimal%20human%20intervention.

URL 7: Grimaldi, E. "How to build a content-based movie recommender system with Natural Language Processing." Towards Data Science, Accessed 17. May 2021. https://towardsdatascience.com/how-to-build-from-scratch-a-content-based-movie-recommender-with-natural-language-processing-25ad400eb243

URL 8: Communication from the Commission to the European Parliament, the European Council, the Council, the European Economic and Social Committee and the Committee of the Regions on Artificial Intelligence for Europe, Brussels 25.4.2018 COM(2018) 237 final. Accessed 17 May 2021. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=COM%3A2018%3A237%3AFIN

URL 9: Marr, B. "What Is The Difference Between Weak (Narrow) And Strong (General) Artificial Intelligence (AI)?" LinkedIn, Accessed 17 May 2021. https://www.linkedin.com/pulse/what-difference-between-weak-narrow-strong-general-artificial-marr/?trackingId=PeGWhIxST8m69dhMG4pACA%3D%3D

URL 10: Kurzweil, R. "Long Live AI." Forbes, Accessed 17. May 2021. https://www.forbes.com/forbes/2005/0815/030.html?sh=426123fb7e8f

URL 11: 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, Accessed 26. August 2021. https://eur-lex.europa.eu/legal-content/EN/TXT/?qid=1623335154975&uri=CELEX%3A52021PC0206

URL 12: Iqbal, N. "Film fans see red over Netflix 'targeted' posters for black viewers." The Guardian, Accessed 17. May 2021. https://www.theguardian.com/media/2018/oct/20/netflix-film-black-viewers-personalised-marketing-target

URL 13: Reuters. “Netflix show portraying Jesus as gay sparks anger in Brazil.” The Guardian, Accessed 17 May 2021
https://www.theguardian.com/world/2019/dec/17/netflix-show-portraying-jesus-as-gay-sparks-anger-in-brazil
URL 14: MollyKate, C. “Why ‘13 Reasons Why’ Can Be Triggering for People Coping With Mental Illness.” Teen Vogue, Accessed 17 May 2021.
https://www.teenvogue.com/story/13-reasons-why-can-be-triggering-coping-with-mental-illness
URL 15: Mahdawi, A. “Goop has a Netflix deal – this is a dangerous win for pseudoscience.” The Guardian, Accessed 17. May 2021.
https://www.theguardian.com/commentisfree/2019/feb/08/goops-deal-with-netflix-is-a-dangerous-win-for-pseudoscience
URL 16: Dodgson, L. ‘All the problematic pseudoscience shared by Zac Efron’s health guru and guests in his new Netflix show ‘Down to Earth’.” Insider, Accessed 17 May 2021.
https://www.insider.com/pseudoscience-in-zac-efron-new-netflix-show-down-to-earth-2020-7
URL 17: McVeigh, K. “Seaspiracy: Netflix documentary accused of misrepresentation by participants.” The Guardian, Accessed 17. May 2021.
https://www.theguardian.com/environment/2021/mar/31/seaspiracy-netflix-documentary-accused-of-misrepresentation-by-participants
URL 18: Pew Research Center. “Theme 7: The need grows for algorithmic literacy, transparency and oversight.” Accessed 20. May 2021.
https://www.pewresearch.org/internet/2017/02/08/theme-7-the-need-grows-for-algorithmic-literacy-transparency-and-oversight/

The emperor is naked: Moral diplomacies and the ethics of AI

With AI permeating our lives, there is widespread concern regarding the proper framework needed to morally assess and regulate it. This has given rise to many attempts to devise ethical guidelines that infuse guidance for both AI development and deployment. Our main concern is that, instead of a genuine ethical interest for AI, we are witnessing moral diplomacies resulting in moral bureaucracies battling for moral supremacy and political domination. After providing a short overview of what we term 'ethics washing' in the AI industry, we analyze the 2021 UNESCO Intergovernmental Meeting of Experts (Category II) tasked with drafting the Recommendation on the Ethics of Artificial Intelligence and show why the term 'moral diplomacy' is better suited to explain what is happening in the field of the ethics of AI. Our paper ends with some general considerations regarding the future of the ethics of AI.

Keywords: *moral diplomacies, moral bureaucracy, AI ethics, AI guidelines, ethics washing*

Acknowledgement

This work was supported by a grant of the Romanian Ministry of Education and Research, CNCS -UEFISCDI, project number PN-III-P1-1.1-TE-2019-1765, within PNCDIIII, awarded for the research project *Collective moral responsibility: from organizations to artificial systems. Re-assessing the Aristotelian framework*, implemented within CCEA & ICUB, University of Bucharest (2021-2022).

Author Information

Constantin Vică, Faculty of Philosophy, University of Bucharest
https://orcid.org/0000-0001-8975-8827
Cristina Voinea, Bucharest University of Economic Studies
https://orcid.org/0000-0003-4654-0697
Radu Uszkai, Bucharest University of Economic Studies
https://orcid.org/0000-0001-5250-8015

How to cite this article:

Vică, Constantin, Cristina Voinea, Radu Uszkai. "The emperor is naked: Moral diplomacies and the ethics of AI."
Információs Társadalom XXI, no. 2 (2021): 83–96.
https://dx.doi.org/10.22503/inftars.XXI.2021.2.6

1. Introduction

Artificial intelligence (AI) is a shining star within the technology world. All other technological innovations and artefacts pale in comparison with what AI, in all its shapes and sizes, promises to offer. However, as the saying goes, all that glitters is not gold. AI technologies have pushed the significance of dual-use to the extreme: whether we think about autonomous weapons, facial recognition technologies or already mundane decision-making software, all of these applications can be used for both good and bad purposes. For example, decision-making algorithms can improve efficiency, but they can also reinforce racial prejudices and biases as they may discriminate based on race or gender (Buolamwini and Gebru 2018; Noble 2018). Other AI systems, such as scoring systems, identify and exploit weaknesses that individuals may not be aware of themselves (Citron and Pasquale 2014). And while discrimination, manipulation or exploitation have plagued societies since the dawn of civilization, unlike with human decision-making, AI systems can operate at scale, instantly and automatically, with the potential to affect people in the flash of a second, "at orders of magnitude and at speeds not previously possible" (Yeung, Howes and Pogrebna 2019). It is precisely these actual and potential harms that AI could create that have driven the massive interest in the ethics of AI.

In this paper, we explore the implications and consequences of the particular interests of both private companies and states alike for the development of ethical guidelines for AI systems. In the first section, we look at some critiques of private companies' focus on the development of ethical codes of conduct and guidelines for ethical AI. We show that most researchers tend to focus on the problem of 'ethics washing', which is the superficial and even hypocritical use of ethics for the avoidance of state regulation. The criticism of companies' attempts to self-regulate is based on the belief that they will always strive to advance their own interests, thus their efforts of devising ethical or responsible AI systems will not do away with the wider problems generated by the societal deployment of these technologies. However, a hidden presupposition behind these types of analyses is that if private companies should not be left alone to their own devices, then states should take the lead in the efforts to advance ethics in the field of AI. In the following section, we show that, in practice, states don't fare too well in this domain either. We take as a case study UNESCO's attempt to create yet more guidelines for ethical AI, in order to show that both transnational organizations and states alike use ethics as a locus of power. We advance the term 'moral diplomacy' to describe the strategy of using the language of morality, by transnational organizations, states and the industry alike, to protect and advance forms of technology that can advance certain economic and political interests. In the concluding remarks, we claim that the fight over 'AI ethics' is actually a political fight and that the ethical guidelines and regulations for AI advanced by 'moral diplomacies' are just a way of signalling allegiance to certain ethical values and principles, without actually moving towards their accomplishment.

2. From ethics washing to the bureaucratization of ethics

While the ethical implications of AI have been addressed since the 1960s, the emergence of machine learning and neural networks has brought ethical debates in to the mainstream (Morley et al. 2019). 'AI ethics' is now sort of a buzzword in the field, as it is employed to name and describe a whole array of moral, legal, societal and political concerns associated with the development and implementation of AI technologies. One of the most frequently employed tools that is believed could help resolve the ethical issues generated by AI are documents containing ethical principles, frameworks, checklists and guidelines to aid the development and implementation of AI technologies. These documents are considered a universal panacea for the potential harms generated during technological development and implementation, a fact shown by the diversity and multiplicity of organizations that have rushed to issue such documents, from industry, to governments, transnational organizations, academia and NGOs. An exhaustive list of all these documents and organizations would probably take the whole space of this paper, so here we settle with mentioning just a few of them: IEEE's (2019) *Ethically Aligned Design "Vision"*, *Artificial Intelligence at Google* (2018) manifesto, OpenAI, Partnership on AI or The Foundation for Responsible Robotics. Jobin et al. (2019) identified no less than 84 documents of a non-legal nature (research and position papers excluded) expounding ethical principles and guidelines for AI. Most of these documents are issued by private companies, followed by governmental agencies, while academic institutions, supposedly the only impartial and objective organizations, are the last issuers of such recommendations and guidelines (AI Ethics Lab 2020; AlgorithmWatch 2020).

Although the attention paid to ethics in AI development and deployment is a heartening development, the focus on ethical guidelines is not without its critics. One of the first problems identified by the critics is that most of these documents are principle-based, embracing a deontological approach (Mittelstadt 2019). Principles are highly abstract standards for good, but they tend to be vague as their application is most of the time context sensitive. As a consequence, principle-based AI guidelines have been criticized for not being sufficiently action-guiding (Hagendorff 2020b; Héder 2020). This means that it is not clear to AI practitioners how to put these principles into practice, as principles, by themselves, do not play a role in informing and training the moral reasoning needed for ethical behaviour in a practical context (Greene, Hoffmann and Stark 2019). This is further proven by the fact that, despite the richness in ethical guidelines, 79% of tech workers report that they would like more practical, down-to-earth instructions on how to deal with and address ethical problems in technology development (Miller and Coldicott 2019).

The ineffectiveness of professional codes of conduct or of any sort of guidelines for the development of responsible AI is further complicated by the fact that AI systems can be used in a wide range of domains, from medicine to warfare and many others. Further, AI developers do not have a common background, as they come at AI from various domains and will be specialized in different disciplines, which also means that they might have different moral obligations to attend to (Filipović,

Koska and Paganini 2018). However, ethical guidelines tend to reduce AI developers to a single expertise, which cannot but obscure the complexity of reality (Mittelstadt 2019). Moreover, any sort of deviation from these principles would be hard to notice and also difficult to punish, as these documents lack enforcement mechanisms (Hagendorff 2020a; LaCroix and Mohseni 2020).

Another contentious issue connected to these ethical guidelines is the lack of diversity of their creators. In his analysis of 22 AI ethics guidelines, Hagendorff (2020b) shows that the ratio of female to male authors is 31.3%, and makes an interesting observation that those reports authored primarily by men tended to focus on particular issues, such as privacy or transparency, ignoring the fact that when AI systems are deployed, they become embedded in complex sociotechnical systems. This shows that male-dominated reports tend to oversimplify the problems these technologies give rise to when they complement or even substitute human decision-making, ignoring important issues such as welfare, fairness or even ecological concerns (Hagendorff 2020b). What is more, Jobin et al. (2019), in their analysis of the corpus of the principles and guidelines on ethical AI, noticed an underrepresentation of developing regions, such as Africa, Central and South America and Central Asia, which of course denotes an existing global power imbalance that it seems is even perpetuated in AI ethics debates. This raises questions of global fairness, but it also denotes a sort of technological determinism implicit in most documents. It is almost as if humans can only react to these technologies as if they are a force that we cannot shape (Greene, Hoffmann and Stark 2019). Further, most documents have as their locus design processes, mostly ignoring business or political decisions, revenue models or the incentive mechanisms that after all shape design processes (Yeung, Howes and Pogrebna 2019).

If the lack of specificity and diversity were the sole issues with these ethical guidelines for the development and deployment of AI, then there would be no significant reasons to worry. After all, these are problems that could, in principle, be solved by more careful deliberation and consideration of the purposes and application of ethical guidelines/codes of conduct. Another, more important worry, though, is that these high-level principles and documents are used as a façade by the industry, and essentially as a ploy to delay or plainly avoid policy-maker's reasons to pursue regulation. To put it more simply, the underlying idea in almost all of these documents is that states' role in regulating AI technologies can be sidelined, while the role of the private sector should be overly-emphasized (Wagner 2018). In 2019, the term 'ethics washing' was first used by the philosopher Thomas Metzinger to describe the instrumentalization of ethics by industry, in his critique of the European Commission Ethics Guidelines for Trustworthy AI (Metzinger 2019). Responsible for the creation of this document was the 52-member High-Level Expert Group on Artificial Intelligence (HLEG AI), which was heavily dominated by industry, with only four ethicists part of the team. Metzinger complains that the guidelines issued by HLEG AI are "lukewarm, short-sighted and deliberately vague" precisely because ethics is instrumentalized in order to "distract the public and to prevent or at least delay effective regulation and policy-making" (Metzinger 2019).

The inspiration for this term ethics washing comes from the already popular term 'greenwashing'. The suffix '-washing' is used to denote a gap between the behaviour of a business or government and how that behaviour is framed or communicated to the public (Peukert and Kloker 2020). While greenwashing refers to the discrepancy between the claims companies make about the environmental impact of their products/services and their actual environmental impact (Voinea and Uszkai 2020), ethics washing denotes the proclaimed adherence to ethical standards by AI companies in order to escape regulation and to reassure customers and other stakeholders of their ethical commitment (Bietti 2020; Wagner 2016; Peukert and Kloker 2020; Rességuier and Rodrigues 2020). Besides in the creation of AI working groups meant to issue guidelines for ethical AI, ethics washing is also manifested in ethics partnerships for AI, such as in the employment by industry of in-house philosophers and ethicists with little or no influence on design processes or business operations (Bietti 2020), and also in the funding by Big Tech of academic work on responsible or ethical AI, which is really meant to obscure problems regarding business practices or the political implications of AI systems (Abdalla and Abdalla 2021; Ebell et al. 2021).

The use and abuse of ethics within the technology world seems to be a strategy employed by various stakeholders in order to create the impression, for both the public and governments, that internal self-regulation by science and industry is more than enough for dealing with the risks raised by AI systems (Bietti 2020). In a paradoxical turn of events, ethics is now used to protect and foster the status quo, while eliminating the possibility of moral progress in the technological world. Many important and stringent ethical implications of AI technologies, such as the social and political impacts of algorithmic decision-making, the environmental implications of data processing for AI, and the rise of fake news/propaganda/deep-fakes, as well as the private funding of public research institutions in the field of AI remain virtually unaddressed within these documents (Hagendorff 2020b).

In what follows, we claim that ethics washing is not the most appropriate way to describe the instrumentalization of ethics in the technology world, because it tends to frame the avoidance of regulation of technology companies by public authorities as something that is bad in itself. But the question of whether governments are better placed to regulate complex, constantly evolving and changing technologies, such as ML-based AI, remains unaddressed. We advance the term 'moral diplomacy' to describe the strategy of using the language of morality, by both transnational organizations and the industry alike, to protect and advance forms of technology that can advance certain economic and political interests. Just as the moral diplomacy conceived by US President Woodrow Wilson was an instrument of fighting back against governments that opposed or were hostile to American interests, so the moral diplomacies in today's AI landscape are a way of advancing political and economic interests and of nipping in the bud discussions addressing important questions about power arrangements. In the following section, we show what moral diplomacies may consist of by analyzing the UNESCO Intergovernmental Meeting of Experts (Category II) tasked with drafting the Recommendation on the Ethics of Artificial Intelligence.

3. The birth of moral diplomacy and AI governance

Despite the views of Immanuel Kant (1998), in day-to-day life, ethics is seldom "pure", that is based solely on supreme or ultimate moral principles. When it comes to applying ethics in practice, like in the development of ethical guidelines for AI systems, ethics is surrounded by many other 'vocabularies' and intellectual disciplines: law and legal thinking (especially human rights), economic and institutional approaches, but also political stakes or social opportunities, etc. The main peril for ethics in the highly dynamic landscape of AI is for it to become just a pretext or a decorative, floral *adagio* in attempts to protect and entrench the *status quo*. When ethics becomes mere etiquette, it fails to provide deliberative mechanisms, sound judgment and true answers. Other risks are not to be neglected either: ethics could become an instrument of struggle or persuasion, a motive for negotiation (that involves *trading* and not *pondering* values or principles) or even a way of 'washing' the image of companies. To put it simply, ethics is in danger of becoming a (cultural or even technological) instrument of domination.

What happens *with* AI ethics and the attempts to codify it (in the form of recommendations or White Papers) is the continuation of a trend that started in early modernity. In search of impartiality, ethics is de-personalized, becoming an art of legalization, architectonics or building systems (Iftode 2021). Moreover, ethicists are beginning to lose sight of a fundamental problem in ethics, that is, moral motivation (Iftode 2021). Moral motivation cannot be merely extrinsic, it cannot lie only in the power of a law, a precept, or of any recommendation, no matter how convincing it is. It should be clear for everyone that governing AI systems, for their lifetime cycle, through ethical codes and guidelines, or recommendations is not a solution, but is increasingly becoming part of the problem. AI is not like the commons – be it pastures, rivers or Wikipedia – for which there are models of good collective governance (Ostrom and Hess 2007). AI systems are not common resources (although maybe data should be), and perhaps that is why the open-source development model has not caught on in the AI research world.

We call 'moral diplomacies' the widespread arrangements of negotiating and gaining *consensus* on the moral guidelines for AI development. Until now, there have been at least three notable productions of moral diplomacies: the OECD Recommendation of principles (adopted on May 22, 2019), the EU guidelines, and the in-the-making UNESCO Recommendation. The outputs of moral diplomacies are documents, in a word-based format, which necessarily implies the emergence of moral bureaucracies capable of interpreting and making decisions on their basis. This is a mechanism similar to academic or medical ethical committees, or Institutional Review Boards (IRBs), the institutions putting ethical codes into practice (Molina and Borgatti 2019). In short, ethical AI governance, transcribed in codes or recommendations, is a product of moral diplomacies, further creating moral bureaucracies.

In what follows, we focus on UNESCO's approach to AI ethics, mainly because it is one of the most transparent and open to inquiry[1] cases of moral diplomacy, allowing

[1] This goes hand in hand with the subjective reason for choosing UNESCO: one of the authors was an expert participating in the discussions.

a detailed analysis. Not only was the draft Recommendation made public (UNESCO 2021a), but also the Intergovernmental Meeting of Experts (Category II) tasked with creating the Draft Recommendation on the Ethics of Artificial Intelligence was livestreamed and kept online afterwards (UNESCO 2021b). It is also important to stress that this forthcoming Recommendation is non-binding, i.e. it has no legal effects and creates no obligations (compared to a Convention, which should be instilled in national legislations), and it will not come into effect before being accepted by member states in another high-level meeting, namely the UNESCO General Conference. Before the Intergovernmental Meeting, there was an arduous process of drafting the Recommendation, prepared by the Ad Hoc Expert Group (AHEG) based on wide multistakeholder consultations. The Recommendation included a preamble and 141 articles structured around the aims, objectives, values, principles and areas of policy action (UNESCO 2021). It was accompanied by a preliminary study and a final report. Also, before the meeting, the member states were invited to send their comments and amendments, which in turn produced a huge document of almost 1000 pages. So, the amount of work and the outputs was highly impressive. From this point on, the deliberation began in earnest. Keep in mind here that our short analysis is limited to the first session of the Intergovernmental Meeting of Experts (26–30 April 2021).

This debate is representative for the making of public AI ethics. If we take the ideal model of discourse ethics (Habermas 1990; Bohman and Rehg 2017) as a frame of reference, we can see that not all of its "pragmatic presuppositions" have been fulfilled. First of all, (1) we need to use the linguistic expressions in the same way in order to ensure we have the same meanings in play; then, (2) none of the relevant arguments can be ruled out. Third, we must take into consideration (3) only the strength of the arguments, and not their rhetorical power of persuasion. For things to work, (4) all the participants must be motivated to find the best argument. Last but not least, (5) no-one should be excluded. The result of the deliberative process should be the intellectual empowerment of the participants, and its foundation lies precisely in the equal respect accorded to everybody involved. Undoubtedly, equal respect was given to all the participants who were able to intervene and propose amendments to the articles of the Recommendation. Condition (5) was met, in that no-one present was excluded. It should be noted, however, that not all states were represented, with some having only observers there (such as the USA, which had withdrawn from UNESCO) or were altogether absent.[2] The first condition was impossible to fulfil in practice; for example, participants had different meanings for some of the more contentious terms, such as 'gender equality' or 'universal', meanings that did not necessarily converge. Further, some arguments were ruled out – which contradicts condition (2) – only because of different experts' rhetorical power of persuasion – thus going against condition (3). Further, the experts were not limited to ethicists and AI researchers, many of them were human rights lawyers and activists, or diplomats, appointed by their states to advance specific values and principles resonating with their own foreign

[2] A novelty of this meeting was its online format: it took place on Zoom.

and domestic policies. What was really problematic, though, was the unfulfillment of condition (4): here, the aim was not to find the best arguments, but to block or support various positions without appealing to moral grounding, rather to political expediency.

The specific procedure rules for UNESCO clearly stipulate that discussions on the Recommendation draft may advance by consensus. And here resides the first difference between the philosophical and the diplomatic employment of ethics. Ethical debate cannot have either as a mere goal, or as a method, consensus at any price. Consensus could be the goal of institutions in gaining uniform practices, but philosophical grounding is bound to the strength of the argument and truth finding. Undoubtedly, in a pragmatic sense, reaching agreement is important, but not at the price of distorting its foundation. During this first round of debates, one key question was to find the *sources of normativity* for the Recommendation. Here, the divide was apparent from the beginning (especially during the April 28 meeting): some state representatives insisted on human rights law as having priority and should be the only universal, normative source of the document (UNESCO 2021b). Others insisted on ethics and its particular contribution as an extension into areas that human rights law cannot cover; as an answer to the focus on ethics, some others decried this as 'ethics washing'. One of the participants (observer) said: "The language of ethics has the merit of shedding light on the blind spots in current (positive) international law. At present, the Recommendation has nothing to add to the current legal framework." Also, one ethicist bluntly expressed his opinion in the chat box by stating: "I understand that ethics is out of the scope of this discussion" and then adding a touching quote from Vladimir Jankélévitch, "Evil is the disjunction of virtues, it is to have a virtue without the others". This reaction says a lot about the actual divorce between legalistic thinking and ethical deliberation and, even more, about the unrealistic expectation that an ethical code or recommendation will make AI systems virtuous. Indeed, the delegates spent almost a whole day of the session rejecting an amendment on the role of international law which, in the end, was made less prominent within the Recommendation. We do not wish to comment on whether this is good or bad, right or wrong, but we want in fact to stress that the focal points of the debate did not seem to have in view the ethics of AI, but rather the concealment of the political interests that would like to instrumentalize AI. Article 11 of the draft, "While all the values and principles outlined below are desirable per se, in any practical context there are inevitable trade-offs among them, requiring complex choices to be made about contextual prioritization, without compromising other principles or values in the process, especially human rights and fundamental freedoms" (UNESCO 2021, 7), was the biggest bone of contention. The prolonged debate around it, taking up nearly an entire day, is paradigmatic of the whole context of drafting ethical principles and norms for AI systems. Because the locution 'trade-off' has different meanings within different domains of discourse, the debates regarding what it refers to more precisely, and what ethical values and principles should be prioritized in case of a conflict, almost blocked any advance in reaching agreement.

The inherent conflict between states has moved into the realm of AI ethics, as another attempt to move from the power of arms to the power of speech (which is a fundamental way of preserving peace through culturalization). AI ethics is a territory unclear to many, ideal and conceptual, but with immense material implications. In this game, the clash between lawyers and ethicists is obvious. And even more obvious is the struggle between the 'old world' of human rights, and the 'new world' prone to use AI in governing populations. It parallels the symbolic struggle between universalists (the Western World) and generalists (Iran, China or Venezuela) over the nature of human rights. All this is part of an 'ethical arms race' between organizations (international or industry alike) for exerting influence upon the future of AI development. In all this context, ethics in its practical or applied exercise becomes the loser, the abandoned puppet. As long as these kinds of documents are non-binding, the effort seems directed towards something different from the red-lines or the way AI should be governed for the common good. The incompatibility between ethics, a pluralistic and revisable system (or even fully particularistic sometimes), with its trade-offs, limitations and balancing, and international law, with its positive, rigid foundation, is hard to overcome. This adds to the main issue, namely that when consensual methods are applied in ethics, the risk is that the achieved compromise will totally reduce the normative power of ethical guidelines.

Almost everything that is ethically 'revolutionary' in this kind of document has been or could be eliminated by 'diplomatic games'. For philosophers and ethicists, it is frustrating to see that several conditions of discourse (or argumentation) ethics are not fulfilled and, even worse, that deliberation becomes bartering. Ethics was seen as part of the art of politics by Aristotle. For international organizations, ethics has rather assumed the role of a shield against recognizing the political nature of the creating institutions. For example, in the UNESCO document, there are no remarks about power and the power relationships built around AI. But power asymmetries are real and actual. Moreover, top-down approaches, based on human rights normativity, are necessary, but not sufficient alone.

4. Concluding remarks: Regulating AI, a catch-22 situation?

The presupposition behind criticisms of companies' capacities to self-regulate through ethical guidelines is that states, especially democratic ones, are better suited to take the lead and to impose clear red-lines concerning the development and deployment of AI systems. In the above discourse, we showed that states don't fare too well either in this domain. The ethics of AI, as it is approached today by industry or transnational organizations and states, is yet another proxy for advancing various types of interests – be they financial in the case of private companies, or political in the case of states. This is another example of the fact that technologies are not mere neutral functional tools, but are also ways of doing politics by other means, as Winner argues (Winner 1986). Technologies are political because they are shaped by human choices and institutional structures,

and in their turn, they shape the way things are done in a society. The politics of AI systems lies in the fact that such systems can change the distribution of power at a societal level, empowering some, while making others even more vulnerable than before (Voinea 2016). Currently, the 'fight' over AI ethics is actually a fight over the specific forms of power and authority that these technologies should incorporate.

While not as absurd and paradoxical as Yossarian's conundrums from Joseph Heller's famous 1961 novel *Catch-22*, the ethics of AI seems to be in a catch-22 situation. On the one hand, the recent push for the industry's self-regulation has proven to be unsuccessful. Companies have shown only an instrumental interest in the tools that normative and applied ethics can bring to the table for regulating AI. Naturally, one might think that the solution to the drawbacks of this strategy might be to bring states and international organizations in to fill in the gaps, like most economists tend to think that we should do when we face a market failure.

As our analysis of the UNESCO Intergovernmental Meeting of Experts tasked with drafting the Recommendation on the Ethics of AI shows, this heuristic approach is not useful in our case. Our main claim is that, at its core, the issue lies with the fact that the ethical guidelines and regulations for AI are advanced by what we term 'moral diplomacies', which are employed by both private (i.e. the dominant companies from the industry) and public organizations (i.e. states and other international organizations) for elevating their status by the use of 'moral talk' or, as Tosi and Warmke (2020) put it, for grandstanding purposes.

Whether it's for avoiding more robust regulation and attracting better employees, like it would be in the case of a company like Google (Voinea and Uszkai 2020), or for politicians to signal to the electorate that they care about Responsible AI (post-industrial democracies) or to make their opposition to Western democracies and their WEIRD morality (Haidt 2012) internationally known, it has become clear that we cannot solve a political problem with ethical ramifications (the regulation of AI) just by simply drafting codes of ethics and establishing moral bureaucracies. Even if we were to leave aside the classical criticism of bureaucracies and bureaucrats as being simply budget maximizers (Niskanen 1971; 1994), an opaque ethical infrastructure that does not contribute to the development of moral and intellectual virtues for the individuals who actually work with AI (Constantinescu et al. 2021) would be nothing more than a waste of both public and private resources, and with potentially deleterious consequences.

This quasi-pessimistic outlook on the future of AI ethics can be supplemented by an even further troubling implication for ethicists who want to have an impact outside just academia. Some ethicists might wish to shape the outlook of the industry on AI by seeking employment at Google or other major players in the industry. For others, the option of ensuring ethical checks and balances is part of public AI moral bureaucracies. Our claim is that any ethicists who might strive to advance an unbiased agenda for ethical AI and at least aim to marginally improve the current *status quo* of AI ethics will probably face what Walzer famously labelled as "the problem of dirty hands" (Walzer 1973). For example, an ethicist working for Google might have to accept some privacy intrusions for profit-maximizing pur-

poses in order to push for a more robust concern of the company for eliminating unfair biases in the ways in which the company processes data, for instance. Similarly, working as an AI moral diplomat for a Western democracy might mean that a person would need to sacrifice some of their principles either due to the electoral interests of their employer (i.e. the Government and/or political party in power) or because intercultural negotiations might entail an unsettling balancing of human rights in order to push an agenda that could be acceptable for countries with a different moral *weltanschauung*, i.e. world view. The only question that remains, then, is what is the acceptable threshold after which compromises with both industry and states or international organizations alike becomes morally unacceptable.

References

Abdalla, Mohamed, and Moustafa Abdalla. "The Grey Hoodie Project: Big Tobacco, Big Tech, and the Threat on Academic Integrity." *ArXiv:2009.13676 [Cs]* (April 2021). https://doi.org/10.1145/3461702.3462563.

AI Ethics Lab. "Tool: The Box." *Toolbox: Dynamics of AI Principles*, June 2020, https://aiethicslab.com/the-box/.

AlgorithmWatch."AI Ethics Guidelines Global Inventory by AlgorithmWatch." Retrieved May 8 2021. https://algorithmwatch.org/en/ai-ethics-guidelines-global-inventory/.

Bietti, Elettra. "From Ethics Washing to Ethics Bashing: A View on Tech Ethics from within Moral Philosophy." In *FAT* '20: Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 210–19. New York: Association for Computing Machinery, 2020. https://doi.org/10.1145/3351095.3372860.

Bohman, James, and William Rehg. "Jürgen Habermas." In *The Stanford Encyclopedia of Philosophy*, edited by Edward N. Zalta, Metaphysics Research Lab, Stanford University, Fall 2017. https://plato.stanford.edu/archives/fall2017/entries/habermas/.

Buolamwini, Joy, and Timnit Gebru. "Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification." In *Proceedings of Machine Learning Research. Conference on Fairness, Accountability and Transparency*, 81 (2018): 77–91.

Miller, Catherine, and Rachel Coldicott. "People, Power and Technology: The Tech Workers' View." Retrieved June 8 2021. https://doteveryone.org.uk/report/workersview/.

Citron, Danielle Keats, and Frank Pasquale. "The Scored Society: Due Process for Automated Predictions". *Washington Law Review* 89, no. 1 (January 2014): 1–33.

Constantinescu, Mihaela, Cristina Voinea, Radu Uszkai, and Constantin Vică. "Understanding responsibility in Responsible AI. Dianoetic virtues and the hard problem of context." Unpublished manuscript, April 2021. https://www.researchgate.net/publication/352519451_Understanding_responsibility_in_Responsible_AI_Dianoetic_virtues_and_the_hard_problem_of_context

Ebell, Christoph, Ricardo Baeza-Yates, Richard Benjamins, Hengjin Cai, Mark Coeckelbergh, Tania Duarte, Merve Hickok, Aurelie Jacquet, Angela Kim, Joris Krijger, John MacIntyre, Piyush Madhamshettiwar, Lauren Maffeo, Jeanna Matthews, Larry Medsker, Peter Smith, and Savannah Thais. "Towards Intellectual Freedom in an AI Ethics Global Community." *AI and Ethics* 1, no.2 (May 2021): 131–38. https://doi.org/10.1007/s43681-021-00052-5.

Filipović, Alexander, Christopher Koska, and Claudia Paganini. "Developing a Professional Ethics for Algorithmists." *Working Paper. Bertelsmann Stiftung* 2018. Retrieved May 8 2021. https://www.bertelsmann-stiftung.de/en/publications/publication/did/developing-a-professional-ethics-for-algorithmists.

Google. "Artificial intelligence at Google: Our principles." 2018. Retrieved May 8, 2021. https://ai.google/principles/.

Greene, Daniel, Anna Lauren Hoffmann, and Luke Stark. "Better, Nicer, Clearer, Fairer: A Critical Assessment of the Movement for Ethical Artificial Intelligence and Machine Learning." In *Proceedings of the 52nd Hawaii International Conference on System Sciences*, edited by Tung X. Bui, 2122–2131. Honolulu: HICSS, 2019.

Habermas, Jürgen. *Moral Consciousness and Communicative Action*. Cambridge (MA): MIT Press, 1990.

Hagendorff, Thilo. "AI Virtues–The Missing Link in Putting AI Ethics into Practice." *ArXiv Preprint ArXiv:2011.12750*, 2020a.

Hagendorff, Thilo. "The Ethics of AI Ethics: An Evaluation of Guidelines." *Minds and Machines* 30, no.1 (March 2020b): 99–120.

Haidt, Jonathan. *The Righteous Mind. Why Good People are Divided by Politics and Religion*, London: Penguin, 2012.

Héder Mihály. "A criticism of AI ethics guidelines." *Információs Társadalom* XX, no. 4 (2020): 57–73.

IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems. "Ethically Aligned De-sign: A Vision for Prioritizing Human Well-being with Autonomous and Intelligent Systems, First Edition." 2019. Retrieved May 8, 2021. https://standards.ieee.org/content/ieee-stand-ards/en/industry-connections/ecautonomous-systems.html

Iftode, Cristian. *Viața Bună. O Introducere În Etică*. București: Trei, 2021.

Jobin, Anna, Marcello Ienca, and Effy Vayena. "The Global Landscape of AI Ethics Guidelines." *Nature Machine Intelligence* 1, no. 9 (2019): 389–99.

Kant, Immanuel. *Groundwork of the Metaphysics of Morals*. Cambridge: Cambridge University Press, 1998.

LaCroix, Travis, and Aydin Mohseni. "The Tragedy of the AI Commons." *ArXiv Preprint ArXiv:2006.05203* (2020).

Metzinger, Thomas. "Ethics washing made in Europe." *Der Tagesspiegel*. August 4, 2019. https://www.tagesspiegel.de/politik/eu-guidelines-ethics-washing-made-in-europe/24195496.html.

Mittelstadt, Brent. "Principles Alone Cannot Guarantee Ethical AI." *SSRN Scholarly Paper* ID 3391293 (2019). https://doi.org/10.2139/ssrn.3391293.

Molina, José Luis, and Stephen P. Borgatti. "Moral Bureaucracies and Social Network Research." *Social Networks*, (November 2019). https://doi.org/10.1016/j.socnet.2019.11.001.

Morley, Jessica, Luciano Floridi, Libby Kinsey, and Anat Elhalal. "From What to How. An Overview of AI Ethics Tools, Methods and Research to Translate Principles into Practices." *ArXiv Preprint ArXiv:1905.06876* (2019).

Niskanen, William. A. *Bureaucracy and Public Economics*, Aldershot: Edward Elgar, 1994.

Niskanen, William A. Bureaucracy and Representative Government, Chicago: Aldine Atherton, 1971.

Noble, Safiya Umoja. *Algorithms of Oppression: How Search Engines Reinforce Racism*. New York: NYU Press, 2018.

OECD. "Recommendation of the Council on Artificial Intelligence." *OECD Legal Instruments*, 2019. https://legalinstruments.oecd.org/en/instruments/oecd-legal-0449.

Ostrom, Elinor, and Charlotte Hess. "A Framework for Analyzing the Knowledge Commons." In *Understanding Knowledge as a Commons: From Theory to Practice*, edited by Charlotte Hess and Elinor Ostrom, 41–81. Cambridge (MA): MIT Press, 2007.

Peukert, Christian, and Simon Kloker. "Trustworthy AI: How Ethics Washing Undermines Consumer Trust." In *Proceedings of the 15th International Conference on Wirtschaftsinformatik, Potsdam*, 2020. https://Doi. Org/10.30844/Wi_2020_j11-Peukert.

Rességuier, Anaïs, and Rowena Rodrigues. "AI Ethics Should Not Remain Toothless! A Call to Bring Back the Teeth of Ethics." *Big Data & Society* 7, no. 2 (July-December 2020): 1-5. https://doi.org/10.1177/2053951720942541.

Tosi, Justin, and Brandon Warmke. *Grandstanding. The use and abuse of moral talk*. New York: Oxford University Press, 2020.

UNESCO. "Draft Text of the Recommendation on the Ethics of Artificial Intelligence." *SHS/IGM-AIETHICS/2021/APR/4*. Paris: UNESCO, 2021a. https://unesdoc.unesco.org/ark:/48223/pf0000376713.

UNESCO. "Intergovernmental Meeting of Experts (Category II) related to a Draft Recommendation on the Ethics of Artificial Intelligence." 2021b. http://webcast.unesco.org/events/2021-04-REC-Ethics-of-AI/.

Voinea, Cristina, and Radu Uszkai. "Do Companies Engage in Moral Grandstanding?" In *Proceedings of the International Management Conference*, edited by Ion Popa, Cosmin Dobrin, Carmen Nadia Ciocoiu, 1033–1039. Bucharest: ASE University Press, 2020.

Voinea, Cristina. "Governance without Governors: Politics through Algorithms and Big Data." *Revista de Filosofie*, LXIII, no. 6 (2016): 583–595.

Walzer, Michael. "Political Action: The Problem of Dirty Hands." *Philosophy & Public Affairs* 2, no. 2 (Winter, 1973): 160–180.

Wagner, Ben. "Algorithmic Regulation and the Global Default: Shifting Norms in Internet Technology." *Etikk i Praksis - Nordic Journal of Applied Ethics* 10, no. 1 (2016): 5–13.

Wagner, Ben. "Ethics as an Escape from Regulation: From Ethics-Washing to Ethics-Shopping." In *Being Profiling. Cogitas Ergo Sum: 10 Years of Profiling the European Citizen*, edited by Emre Bayamlıoğlu, Irina Baraliuc, Liisa Janssens, and Mireille Hildebrandt, 1–7. Amsterdam: Amsterdam University Press, 2018.

Winner, Langdon. *The Whale and the Reactor: A Search for Limits in an Age of High Technology*. Chicago: University of Chicago Press, 1986.

Yeung, Karen, Andrew Howes, and Ganna Pogrebna. "AI Governance by Human Rights-Centred Design, Deliberation and Oversight: An End to Ethics Washing." *SSRN Scholarly Paper* ID 3435011 (2019). https://doi.org/10.2139/ssrn.3435011.

PETER KONHÄUSNER,
MARIA MARGARITA CABRERA FRIAS
AND DAN-CRISTIAN DABIJA

Monetary Incentivization of Crowds by Platforms

The platform industry is currently on the rise, and with so many platforms, acquiring users and getting them to engage can be challenging. To address this, many platforms are relying on crowdfunding, network effects and incentives, including monetary incentives. But what techniques are platforms using to monetarily incentivize their crowd? Although the study of platform dynamics has been on the rise, including research on crowdsourcing, network effects and incentivization, there is no present research being done on the methods being implemented by platforms to use monetary incentives on their crowd. This paper uses an inductive empirical method based on grounded theory, with data gathered from 15 different platforms that are known to be using a monetary incentivization method, to analyze and categorize the different strategies used by platforms and their marketing objectives. This paper presents useful information to assist managers to make the right decisions regarding monetary incentives and for fostering the potential of their crowd.

Keywords: *platforms, monetary incentives, crowdsourcing, network effects, incentivization*

Author Information

Peter Konhäusner, Babeș-Bolyai, Faculty of Economics and Business Administration, Department of Marketing, Teodor Mihali 58-60, RO-400591 Cluj-Napoca, Romania
https://orcid.org/0000-0001-6717-1304

Maria Margarita Cabrera Frias, bbw Hochschule, Leibnizstrasse 11-13, 10625 Berlin, Germany
https://orcid.org/0000-0002-2446-5430

Dan-Cristian Dabija, Babeș-Bolyai, Faculty of Economics and Business Administration, Department of Marketing, Teodor Mihali 58-60, RO-400591 Cluj-Napoca, Romania
https://orcid.org/0000-0002-8265-175X

How to cite this article:

Konhäusner, Peter, Maria Margarita Cabrera Frias and Dan-Cristian Dabija. "Monetary Incentivization of Crowds by Platforms."
Információs Társadalom XXI, no. 2 (2021): 97–118.
https://dx.doi.org/10.22503/inftars.XXI.2021.2.7

Introduction

In the second quarter of 2020, seven of the top ten global companies by market capitalization, including Apple, Microsoft and Amazon, had sharing platforms (PricewaterhouseCoopers 2020). With so many platforms in the market (from social media platforms to industry-disrupting infomediaries, like Airbnb), and new ones being launched every day, acquiring users, and getting them to engage or participate can become a challenge. To address this, many platforms rely on the power of the crowd, network effects and incentives, including monetary incentives (Katmada et al. 2011).

The study of platform dynamics has been on the rise over the last decades, with recent research done on crowdsourcing (Sayedi and Baghaie 2017; Moysidou and Hausberg 2019), network effects (Evans and Schmalensee 2017; Parker et al. 2017) and incentivization (Katmada et al. 2011; Toker-Yildiz et al. 2017). Nevertheless, as desk research shows and to the best of our knowledge, there is no present research being done on the methods being used by platforms to offer monetary incentives to their crowd, which opens up a research gap that the present study aims to fill. Therefore, the research question addressed in this paper focuses on how and why platforms use monetary incentivization to engage their crowd.

By examining this topic, the present paper will add theoretical input to the discussion about the use of incentivization techniques by platforms (Ashander et al. 2019; Bratu 2019a; Bratu 2019b; Furnham 2019; Mircica and Sion 2019). In terms of practical input, understanding the different methods used by platforms will assist managers to make the right decisions to foster the potential of their crowd.

While the theoretical aspects are covered by desk research, this paper aims to expose the empirically possible ways to monetarily incentivize crowds on platforms using an inductive empirical method based on grounded theory (Glaser and Strauss 1967). Datasets from 15 different platforms that are known to use monetary incentives, including Reddit, Groupon and TikTok, were collected and analyzed to categorize the different strategies as the outcome of the paper. After the data were compiled and structured, clusters were created by grouping together platforms showing similar characteristics, resulting in a theoretical approach of categorization regarding the typical characteristics of platform strategies using monetary incentivization for engaging their crowd. From a theoretical perspective, we provide useful insights regarding how the theories of crowdfunding and platform economics are interdependent yet influence each other. From a managerial perspective, we present useful information for platform owners regarding how to use incentives with their audience to increase user acquisition and engagement, by providing an overview of the different methods used by various platforms to reach their objectives, and that can be key components of an overall marketing strategy.

The paper is structured as follows: Section one presents an overview of the research on platform dynamics, network effects and incentivization, while in section two, crowdfunding theory is reviewed as a possibility to use and build a community. Section three deals with the research methodology, while section four reports the

empirical findings and describes the different incentive strategies used by the selected platforms. Section five presents discussions of the analysis and categorization, and then the paper ends with the main conclusions of the study.

This paper presents useful information for platforms looking for ways to use incentives with their audience to increase user acquisition and engagement, with the paper providing an overview of the different methods used by various platforms to reach their objectives, and that represent key components of their overall marketing strategy.

1. Platform Dynamics, Network Effects and Incentivization

Platforms are defined by Zhu and Furr (2016) as "intermediaries that connect two or more distinct groups of users and enable their direct interaction". They allow individuals and companies to pursue their own transactions by using the infrastructure and services of a core organization (Hegel et al. 2008), and they create an ecosystem that promotes shared values in place of traditionally transactional relationships (Brown 2016). Platforms can often benefit from the sharing economy, where "consumers and organizations have opportunities to collectively innovate [and] create value" (Lim 2020; Stare and Jaklič 2020). Unlike products, which usually generate single revenue streams, platforms have the potential for multiple revenue streams, which is why many organizations have entered the platform industry, with such platforms either being created outright as platforms or starting out as products and then making the leap to platforms to serve their niche (Zhu and Furr 2016).

Interactions on a digital platform work like any economic or social exchange in the real world, meaning there is an exchange between the producer and consumer of information, goods or services and some sort of value, e.g. a currency (Parker et al. 2017; Culkin 2019; Oláh et al. 2020; Stare and Jaklic 2020). These interactions have clear relationships that enable business success, while ensuring a common goal or purpose is provided, and a strong sense of trust within the network is present to support the exchange (Brown 2016). The consumer can also switch sides and become a producer – in which way, he would be called a "prosumer" (Dabija et al. 2019; Meilhan 2019).

An information exchange allows the parties to decide whether and how to engage in a transaction. This means that platforms facilitate the exchange of information so that further transactions can occur. On some platforms, the exchange of information can be the desired outcome; for example, forums (e.g. Craigslist, Reddit), while on other platforms, after exchanging information, the parties can decide to also exchange goods or services (e.g. eBay, YouTube, Uber). Depending on the category, the entire exchange process can happen within the platform, while in other cases, the exchange could be organized within the platform but continues outside of it. Finally, there can be an exchange of a value unit, which could be traditional currency (money) or other types of desired value, such as in-app coins or attention in the form of likes, views and influence (Parker et al. 2016).

1.1. Network effects

One of the main features of platforms is the network effect, which can be used to benefit their user and usage growth. According to Parker et al. (2016), the network effect can be defined as the positive or negative change of a platform's value in relation to the number of users of the platform. A platform that successfully generates network effects will not only increase its value for the user, but also the overall platform value; "in other words, when a user joins the platform, the value of the platform to all other users increases" (Posthumus 2017). Social networks are the most popular communication tools to attract potential customers now (Nadanyiova et al. 2020; Pop et al. 2021). Some social networks use controversial online marketing techniques to grow their market (Héder 2019).

Although the use of network effects within platforms seems more prominent today, scholars began observing such effects as long ago as the 1970s, while a significant growth in network effect theories took place in the 1990s (Evans and Schmalensee 2017). According to Parker et al. (2016), "positive network effects refer to the ability of a large, well-managed platform community to produce significant value for each user of the platform". Belleflamme and Peitz (2016) discuss how network effects can be generated in quite different ways depending on who creates them and who is affected by them. A situation where two user groups affect each other is called a two-sided network effect (Evans and Schmalensee 2017). In this case, each user group is responsible for the attraction of the other, creating a consistent cycle of engagement and retention of users. One type of two-sided network effect is cross-side network effects, which occur when an increase of one user group affects the value of the platform for the opposite user group (Eisenmann et al. 2006). Two prominent examples of platforms that rely on cross-side network effects are Uber and eBay, where more possible passengers attract more drivers, more buyers attract more sellers, and vice-versa (Eisenmann et al. 2006).

In contrast to cross-side network effects, same-side effects occur when an impact of users on one side of a network affects the value of the platform for that same side (Eisenmann et al. 2006). For example, social media platforms often create positive same-side effects through the amount of people active in their network (Petrovic 2010; Dabija et al. 2017; Sârbu et al. 2018; Atwell et al. 2019). Also, platforms like Uber or eBay use same-side network effects to attract users because peer groups using the platforms result in the encouragement of more people to join the platform. Another example includes the facilitation of job searches through the usage of network effects (Lemke 2019).

Careful curation of the value shared on the platform is needed to maintain the platform ecosystem (Grudin et al. 2019). The value creation on platforms is linked to network effects, attracting more-demanding consumers to the platforms when value is created, which again attracts more providers offering value (Hagiu and Yoffie 2016; Ślusarczyk et al. 2020). However, it is important to understand how certain attributes contribute to customer satisfaction (Suzuki et al. 2019). These positive network effects create competition for the best price-value-combination, because charging or requiring payment can discourage the entry, participation, value creation and consumption, depending on where the charging occurs. It is therefore important to analyze the value that is being created, so that the right point of mon-

etization can be found without harming a platform's ecosystem. Platforms can be monetized by either generating revenue from the supply or the demand side of the platform, by the successful transactions between the two sides or through providing access to the whole community for external third parties, like advertisers (Posthumus 2017). This means that users could be charged for access to the value created on the platform, producers could pay for the access to a community, both can be charged for the access to interactions, or both can pay for curation mechanisms that enhance the interactions (Parker et al. 2016). A successful user experience strategy must thus analyze the interactions of the users and the providers to identify the sources of excess value that the platform generates to select where the monetization can take place without inhibiting the growth by network effects (Durlauf 2019).

1.2. Incentivization

In many cases, a platform can begin free or with a discounted pricing to generate the first network effects. Afterwards, it can move towards the "freemium" strategy of charging for extras. It can also have free or discounted prices for one side while the other side pays. These incentives usually occur when one side highly values the presence of the other side on the network (Parker et al. 2016). Monetary incentives can be a way to promote desired behaviours, such as survey responses (Hansen 1980) or encouraging word-of-mouth referrals (Wirtz and Chew 2002). This effect is strong mainly when combined with social incentives and influence, as social interactions are found to be particularly significant (Toker-Yildiz et al. 2017).

Different incentives are used in crowdsourcing platforms as these are very highly dependent on user participation. The incentives in crowdsourcing platforms are usually individual (self-learning, enjoyment and altruism), social or monetary, among others (Katmada et al. 2011; Pedregosa et al. 2020). Financial rewards trigger extrinsic motives to get compensation and can be a good option when social or individual rewards are missing (Blaškova et al. 2017). Crowdsourcing platforms often use monetary rewards combined with reputation systems or other incentives, such as self-marketing. Katmada et al. (2011) exemplify this with iStockPhotos, an online stock image platform, where users can submit their photos and receive commission. Financial rewards can increase participation but should be used with caution, as there is scepticism towards them, and they can decrease intrinsic motivation or push people to try to cheat the system. Using small monetary incentives as an initial motivating factor and then utilizing other rewards, such as prizes, to achieve sustained engagement can result in more sustainable results (Katmada et al. 2011).

2. Crowdfunding

Crowdfunding is a subcategory of crowdsourcing that was introduced by Howe (2006) and that can be used by organizations to gather monetary funds. Furthermore, the organizations seeking crowdfunding as well as the platforms connecting

the supporters and campaign organizers depend on the crowd. Generally, crowdsourcing follows the principles that external stakeholders are supporting the company by adopting core processes. This is the reason why crowdsourcing is seen as a combination of the concepts of the "crowd" and "outsourcing" (Opstal 2013, 86). In this regard, looking at crowdfunding, stakeholders are giving money to a project or a company. Crowdfunding can be split up into four different categories: donation-, lending-, reward- and equity-based crowdfunding. The basic principle involves a crowd of people (many single "crowdfunders") giving money for a project or venture in response to a campaign run by the project organizer (most of the time the company or individual behind the project). While the basic principle is the same for all four crowdfunding categories, the types of interaction and transaction vary from crowdfunding category to category (Meyskens and Bird 2015; Gierczak et al. 2016); e.g. donation-based crowdfunding aims at giving money to people and projects in need, while the aim for lending-based crowdfunders is to earn interest. Reward-based crowdfunding benefits people by providing goods or services in return for their investment, while equity-based crowdfunding aims at a long-term relationship and rewards, including a return of profits as well as co-determination (Agrawal et al. 2014; Pedregosa et al. 2020; Konhausner et al. 2021).

Crowdfunding campaigns can have various levels of reach, from activating the global community to seeking support from a local group of people (Mollick 2014). By taking the potential conversion rate from lead to crowdfunder as well as the level of engagement of the targeted community into consideration, the campaign organizers can estimate the outcome of a proposed campaign. As diverse as the crowdfunding levels of reach are, the definitions of the crowd and, therefore, the implications of this are as well. The crowd can be a group of as few as two people, but also a community of billions of users, such as Facebook users (Sternberg and Todd 1995).

The goals of crowdsourcing can vary from idea generation to support for product development, while the general goal of crowdfunding is primarily seen as a financial benefit for the project. Besides that, an additional benefit for project organizers can be identified in the marketing effect of the campaigns, namely shaping the public image of the campaign object by communicating proactively about the campaign as well as the goal of the campaign (Friedman 2013; Konhäusner et al. 2021). The negative connotation of crowdfunding, which companies using crowdfunding often encounter, namely that one of the reasons why they are using crowdfunding could be that no bank, financial institution or investor is willing to give them money, can be countered with the positive, long-term effects delivered on the marketing side (Sayedi and Baghaie 2017; Pedregosa et al. 2020).

The crowd, nevertheless, must be able to see the clear benefit of the campaign to engage the crowd and to win their commitment (Belleflamme et al. 2014). This commitment can be differentiated between either the short-term or the long-term commitment of the crowd that a project organizer is aiming for. These differ to some degree, whereby short-term commitment can be accomplished by fulfilling the basic expectations on the return of the campaign, such as interest, goods or services, while long-term commitment embraces more factors, like communication and the integration of the crowdfunder. If the expectations of the crowdfunders are not met,

a project organizer might face a trust issue (Zheng et al. 2016; Hollowell et al. 2019). The best case would be to transform customers into long-term investors of the company (Ordanini et al. 2011).

For a project organizer, one of the first questions will be where to find potential supporters for his or her endeavour. Online platforms can disrupt different industries, where they offer a way to disintermediate processes and to reach new audiences (Hagel et al. 2008). For the project organizer, the trust a community has in the selected platform can be of utmost importance as it can influence the success of the campaign (Moysidou and Hausberg 2019). The success is also directly influenced by the platform dynamics.

A typical crowdfunding campaign is, as pointed out, defined by supporters giving money via a platform to a project organizer (Lukkarinen et al. 2016). There is also the possibility of long-term relationships stemming out of crowdfunding campaigns, usually, if there is a long-term return involved. The question remains though: What would happen if a project organizer were to give back money to the supporters. This can normally be managed by the platforms where project organizers and supporters meet. The platforms could also, potentially, give a monetary incentivization to the users to engage on the platform.

3. Research Methodology

After highlighting the characteristics of crowdfunding and describing the features of platform economics the linkage between platforms and crowds can be unidirectionally explained: Crowds need platforms to interact and to participate in projects. On the other side, the need of platforms for crowds in terms of loyal users is imminent. Without users, the platforms will not attract new projects and will fail to acquire new users. In this regard, marketing for the platforms is one of the main components of a sustainable business approach. As platform research mainly focuses on network effects and the adaption of the traditional marketing mix (Sridhar et al. 2011), research on the option of a platform providing monetary incentivization to the crowd is lacking.

Thus, the research question arises: Which methods of monetary incentivization are used by platforms to attract, hold and engage users, i.e. the crowd? As this perspective is a new and innovative approach, a method is needed that can examine this phenomenon and extrapolate it to a bigger scope.

This paper uses the inductive empirical method based on grounded theory (Glaser and Strauss 1967; Tie et al. 2019) to uncover empirically possible ways to monetarily incentivize crowds on platforms. This method involves constructing and discussing theories based on the collection of data through various sources, such as interviews and observations (Faggiolani 2011). As for this paper, the aim was to gather data from different platforms (current platforms as well as on platforms that have gone out of business or changed their marketing strategy) that are or were using monetary means to incentivize their users to use and stick to the platform. This should then, in turn, lead to a categorization of the different strategies applied,

which can form key parts of an overall marketing strategy. Using grounded theory, the phenomenon of users' individual decisions can be made more understandable (Aldiabat and Navenec 2011).

The data used in this empirical research was gathered by desk research and processed according to grounded theory (see the underlined steps below, which are in accordance with Bernard and Ryan 2010). A total of 15 platforms from various industries were identified based on their industry relevance as well as the growth in their specific niches. Table 1 shows the selected platforms.

#	Name	#	Name
1	Reddit	10	Zoom
2	PayPal	9	GrubHub
3	TikTok	10	Xbox
4	FatKat Club	11	Groupon
5	Good Shepard Entertainment	12	Duolingo
6	Airbnb	13	Dropbox
7	Uber	14	Trello
8	eBay	15	WeAre8
9	99designs		

Table 1. Overview of the selected platforms

In the first step, the data of the platforms and businesses were descriptively analyzed and tagged with keywords (codes). The incident, as the typical observational unit of grounded theory (Glaser and Strauss 1967), is the use of monetary incentivization techniques to form/retain a crowd. The hypothesis for this grounded theory approach is that different clusters with objects having similar characteristics can be derived from the data gathered in the first step. The data were clustered into different groups according to the analysis of the shared characteristics (concepts). These clusters were named as categories in the last step and their common attributes were described. The research resulted in a theoretical approach with categories embracing the typical characteristics, which need further verification, indicating if they are generally applicable for platforms using monetary incentivization for engaging their crowd. Further research steps have been gained and noted throughout the whole process.

4. Research Findings

The research is based on desk research that brought up 15 different cases, which are summarized in Table 2. Besides the case number in the table view of the descriptive analysis of the platforms and businesses, the name of the platform and the marketing method of interaction with the crowd are highlighted. Also, the objective

that can be attributed to the specific method regarding the platform is pointed out. Furthermore, the incentivization is briefly qualitatively described by the authors and a rough starting point of time when the platform started using the method is noted down (step 1 – codes). References to the individual cases are directly included in Table 2.

The data was gathered by doing research on the relevant platform as well as on the marketing techniques used by the platform in the past and currently. The table has been filled with data including the incentivization method used by the platforms (in the form of a keyword), the maturity of the method (describing, if it is a short-term (up to two years), medium-term (up to five years), or long-term marketing (more than five years) approach), and an explanation of the incentivization approach in a descriptive form. The maturity was analyzed taking the observed incentivization objective and method used into consideration. Besides the date, where the specific method was first introduced or used on the platform as well as the primary objective of the method has been noted down in the table (e.g. user acquisition). According to Bernard and Ryan (2010), this step is the coding phase, which is the foundation of the grounded theory approach.

The forum and communication platform Reddit engaged their community by offering shares in their project, which can be seen as a long-term loyalty programme to strengthen the community and to carve out its user orientation even more. The announcement of promoting user involvement in this way echoed throughout the internet in 2014 (D'Orazio 2014).

As an experiment to grow faster and to acquire new users quicker, the payment service provider PayPal initially offered new users a free 5-dollar voucher for sign up. This is an example of a cost per action (CPA)method. The downside of this method for PayPal was that users were free to decide where they would spend the money, which led to a lot of new registrations, a massive amount of money spent in terms of vouchers, but a minimal engagement of users as many seemed to create new user accounts just to get the voucher, then leaving the account untouched afterwards (Parker et al. 2016; O'Connell 2020).

The social media platform TikTok uses a variety of different methods to acquire, retain and engage users on their platform. Many of the methods, which include a referral programme, a rewards programme and a creators' fund, were introduced in 2020 and TikTok believes they should lead to user and engagement growth (TikTok 2020). The two gaming-focused equity-based crowdfunding platforms FatKat Club (planned launch in 2021) and Good Shepherd Entertainment are trying to target crowds that are supporting games. In contrast to general equity-based crowdfunding platforms, gaming-focused offerings tend to have a higher community engagement factor due to the nature of the subject matter. Both platforms offer rewards as well as revenue share and curation mechanisms, but FatKat is also open to non-accredited investors, which would mean targeting a whole new audience (Pereira 2020; Good Shepherd Entertainment 2020).

#	Name	Method	Incentivization explained
1	Reddit	Equity	Community got offered shared of the project
2	PayPal	CPA	PayPal gave a 5 US$ voucher to every new registrar
3	TikTok	CPA	-Referral program: where users get an incentive to invite friends -Rewards program: users get incentives to engage with content -Creators fund: users get incentives to become content Creators and monetary Rewards for the quality of the content incentive varies according to Countries and some are temporary
4	FatKat Club	CPA	Rewards, Revenue Share, Curation, Non-Accredited Investors (opening up a whole new market)
5	Good Shepard Entertainment	CPA	Rewards, Revenue Share, Curation
6	Airbnb	CPA	Airbnb Plus: incentive program to help hosts cover the cost of fixes they'll need to complete to become verified for Airbnb Plus T-Points: Users get points for joining Airbnb and booking stays. These points can later be redeemed for rewards with partners. (Japan)
7	Uber	CPA	Voucher for the first ride and for referring new users
8	eBay	CPA	Commission is paid for every sale referred; also additional information for sellers for better marketing is provided
9	99designs		coupons for the first order placed on the platform?
10	Zoom		
9	GrubHub	CPA	Refer a Friend: 10-15 voucher for the first order with referal code
10	Xbox	indirect CPA	Microsoft sold the console (first Xbox) under the production cost (resulting in a negative marginal return) to grow the installed base
11	Groupon	CPA	Grupon Bucks: Rewards that users can gain by participating in activities and can be used within the platform Refer-A-Friend: Users can get Grupon Bucks if they refer a new user and they make a purchase
12	Duolingo	indirect CPA	free months of premium membership for referred users
13	Dropbox	indirect CPA	free storage for referred users
14	Trello	indirect CPA	free premium months for referred users
15	WeAre8	CPA	community is paid for watching social content & interacting

Table 2a Overview from our platform research

#	Name	Objective / Concept	Introduction Date
1	Reddit	User Involvement	2014
2	PayPal	User Acquisition	1999/2000
3	TikTok	User Acquisition, user engagement and content creation	Creators fund: Since July 2020 Referral and rewards program are temporary and might come and go as needed by the platform
4	FatKat Club	User Retention through equity-based crowdfunding in games	2021 (official start of website / program)
5	Good Shepard Entertainment	User Retention through equity-based crowdfunding in games	2011 (as a company, later as a platform)
6	Airbnb	User Acquisition and engagement	n.d.
7	Uber	User Acquisition	2009
8	eBay	Sales Promotion	2014 (relaunched 2019)
9	99designs		n.d.
10	Zoom		n.d.
9	GrubHub	User Acquisition	n.d.
10	Xbox	User Acquisition	2001; 2014
11	Groupon	User acquisition and engagement	n.d.
12	Duolingo	User Acquisition	n.d.
13	Dropbox	User Acquisition	n.d.
14	Trello	User Acquisition	2015
15	WeAre8	User Engagement	n.d.

Table 2b Overview from our platform research

#	Name	Maturity	Reference
1	Reddit	long-term	D'Orazio 2014
2	PayPal	short-term	Parker et al. 2016; O'Connell 2020
3	TikTok	short-term	https://www.tiktok.com/legal/terms-of-use?lang=en https://www.tiktok.com/legal/virtual-items?lang=en https://www.tiktok.com/legal/creator-program-terms-of-service?lang=en https://www.tiktok.com/legal/tiktok-referral-program-terms?lang=en
4	FatKat Club	medium-term	Pereira 2020
5	Good Shepard Entertainment	medium-term	https://www.goodshepherd.games/
6	Airbnb	short-term	https://www.airbnb.com/help/article/2368/how-does-the-airbnb-plus-advance-incentive-program-work https://www.airbnb.com/t-point/
7	Uber	short-term	http://heartofcodes.com/marketing-strategy-of-uber/ https://www.annexcloud.com/blog/ubers-marketing-strategy-in-7-steps/
8	eBay	short-term	https://pages.ebay.com/seller-center/listing-and-marketing/ebay-partner-network.html
9	99designs		
10	Zoom		
9	GrubHub	short-term	https://www.grubhub.com/legal/referral-terms
10	Xbox	mid-term	Tassi 2014; Parker et al. 2016
11	Groupon	short-term	https://www.groupon.com/legal/specialprograms
12	Duolingo	short-term	https://www.duolingo.com/plus
13	Dropbox	long-term	https://help.dropbox.com/de-de/accounts-billing/space-storage/earn-space-referring-friends
14	Trello	long-term	https://trello.com/recommend https://www.makeuseof.com/tag/get-free-trello-gold-can/
15	WeAre8	mid-term	https://www.weare8.com/

Table 2c Overview from our platform research

The hospitality disruptor Airbnb is focusing on user acquisition and engagement in the short term with an incentivization programme as well as a gamification approach (Chou 2019). One of their incentives is support for hosts to become verified for Airbnb. Another of their programmes takes advantage of gamification, by allowing users to get T-Points whenever they book accommodation. This approach is currently in the test phase in Japan (Airbnb 2020).

To acquire new users for its ride platform, Uber offers vouchers for the first ride as well as for referring new users. Moreover, there are temporary in-app discount offers for users coming back after a while away to encourage them to use the service again or to incentivize users to make longer trips during non-peak times (Miller 2016). Another major platform, auction platform eBay, has outsourced most of its campaign-specific marketing to the sellers themselves by providing them with the tools needed to promote their own articles that they put on sale on the platform. Thereby the users themselves drive new users and keep up the user engagement. Moreover, commission is paid to users who are referring new users who make a new sale (eBay 2020).

The food delivery platform GrubHub operates on a short-term cost per action (CPA) basis, offering 10 to 15 dollars in the form of a voucher for referring a new customer upon their first order. This method is quite common among food delivery platforms, although the voucher value of GrubHub is high and hence, the long-term orientation of the user commitment in that case could be questionable (GrubHub 2020). The gaming console Xbox sold its hardware under production cost causing a negative profit margin just to grow its installed base of future users, who will then buy the games and subscribe to services. This medium-term strategy can be described as an indirect CPA (Tassi 2014; Parker et al. 2016).

Referral methods as well as gamification features can be found on the voucher platform Groupon, where users can earn points for participating in activities and can mediate new users. These techniques can also trigger competitive thinking and lead to more platform usage (Groupon 2020). Using an indirect CPA, language learning app Duolingo offers free premium membership months for referring new users. This is an indirect monetary investment, exchanging possible revenue for onboarding new users (Duolingo 2020).

Similar methods can be seen being used at cloud-service provider Dropbox, which is offering free storage for referred users, while task-management tool Trello is giving away premium months for referrals. Calculating the value platforms spend for new users by incentivizing existing users with intangible services reveals the value referrals have for platforms (Patkar 2015; Dropbox 2020). Social media platform WeAre8 pays users money for using the platform and for staying loyal. Added gamification elements are also triggering higher user engagement (Brown 2014; WeAre8 2020).

5. Categorization and Discussions

Mutual characteristics of the datasets gathered were identified and clusters were formed (step 2 – concepts). These definitions of concepts lead to the formation of categories (step 3 – categories), which are presented here in a joint table with the

characteristics (see Table 3). Some of the datasets can be inserted into more than one category which validates the fuzzy transitions between methods.

#	Category Name of Cluster	Cases (#)	Specified Attributes	Probability of Success	Comments	Total covered by cases observed
1	Acquisition-Driven	2, 3, 7, 11	short-term user acquisition	low entry barriers positively affect the acquisition (see also Parker et al. 2016)	short-term marketing activities to drive users to the platform; network effects are needed to bind them longer	26.67%
2	Integration-Driven	1, 15	long-term direct user dialogue	unclear, but probably high (see equity-crowdfunding)	Platform case #15 unclear because of individual usage pattern	13.33%
3	Retention-Driven	3, 4, 5, 8, 10, 15	mid-term user binding, goal is to keep the user in the ecosystem	depends on competition and other offers	For all platforms keeping the user in the ecosystem is the goal, but in thee cases the focus of the platform interaction with the users is to bind them mid- to long-term	40.00%
4	Activation-Driven	3, 6, 9, 12, 13, 14	long-term engagement (also because of the design of the platform) and referral, gamification elements	gamification elements and reminders have a positive effect on the success of the long-term activation	Some Activation-Driven Platforms are also using marketing instruments to stimulate the community	40.00%

Table 3 Data categorization (own creation)

The two most prevalent categories (both accounting for 40%) of monetary incentivization of platforms for their respective crowds aim at retention as well as activation. Retention-driven approaches are aimed at a medium-term binding of users to keep the users in the ecosystem. The primary objective is to not let the user migrate to another platform rather than raising his or her engagement. This latter objective can be a positive side effect but is not the purpose of the activity. Keeping in mind that users are more difficult to win back than to keep in the system at nearly all costs (Bruhn 2016), platforms try to lock users into the system hoping for them to become more active and deliver more value.

Activation-driven approaches, on the other hand, use elements like gamification to raise the engagement of users in the long-term. Another objective is to trigger referrals caused by the belief of the users that the platform is adding value. Additionally, the design of the platform aims to deliver an enhanced user experience. Marketing-driven approaches (approximately 26% of the sample) aim at short-term user acquisition for the platform, ignoring the medium- to long-term value a user can bring to the community as well as to the company. Besides the effective use of word-of-mouth marketing's low entry barriers benefit quick user-acquisition.

Approximately 13% of the datasets reviewed involved integration-driven approaches aimed at an intensive, long-term user dialogue. This can be triggered by integration of the user into the company as a shareholder or a close stakeholder. One of the reasons why this approach is not widely applied is that the long-term

consequences are unknown and cannot be assessed yet. Also, the benefits of such a close integration can be too blurry for companies to agree on this close partnership with the crowd.

Research in the platform field has also focused on the different categories mentioned. The results of this paper are complementary to the findings of Bezzubtseva and Ignatov (2013) in terms of the typology of users a collaboration platform is aiming for. Also, research has reported how mobility platforms are aiming to attract more users through different means (Malzahn et al. 2020). Geng et al. (2019) point out that big data analysis can be used for the improving the user acquisition of industrial data platforms. Gutierrez-Leefmans and Holland (2019) highlight how platforms can be seen as business models for small- and medium-sized enterprises and focused their research on user retention by implementing an activity system. Granfeldt and Nyqvist (2019) concentrate on retention mechanisms for users on multisided platforms but did not take equity-based approaches into consideration. In terms of user activation, Lee and Kim (2019) propose a toy-focused approach for an open-source platform using a 3D printer for the community. As another aspect covered, the effects of message interactivity and platform self-disclosure on user activation were discussed by Adam and Klumpe (2019).

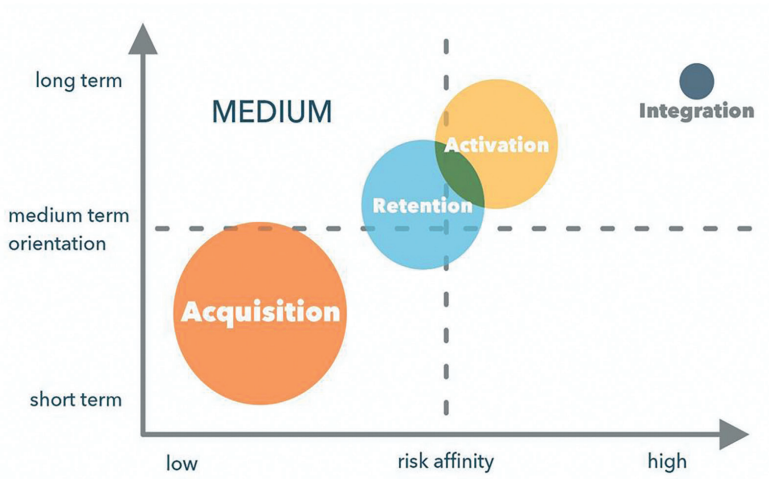

Figure 1 Monetary user-incentivization categories for platforms (own concept)

From a scientific standpoint, this research adds to the research about platform dynamics from a monetary marketing perspective. It highlights the possibility that platforms can use direct monetary means to incentivize users to employ a platform and to engage on the platform. The approach aiming for user integration (by giving equity of the platform to different users) is a relatively new, innovative way to bind users, who are already emotionally loyal, to the platform economically for the long term. Out of the generated categories (see Table 3), a theoretical approach (step 4 of the grounded theory) can be derived, which is depicted in Figure 1. The four categories are plotted on a matrix consisting of the dimensions “risk affinity”

(of the platform owner) and timely orientation of the methods used for monetary user incentivization. The risk affinity is derived from the depth of the integration of the user into the ecosystem of the platform. As a user is only slightly involved in the platform if newly acquired, but heavily involved if he/she owns equity in the platform, these two categories form the outer layers. The respective size of the depicted spheres shows the observed quantity of the respective category.

Risk affinity in the model is observed from the perspective of the platform owner. The more the platform ecosystem is open towards the community and the more impact the crowd can have, the higher the risk for the platform it may lose control or may become even more dependent on the community. The short-term acquisition of users without long-term engagement leads, therefore, to a minimum risk for the platform, while giving a voice to the community, while integrating them deeply into the business model raises the risk.

6. Conclusions

Rather than using crowdfunding to finance platforms, the managers of platform businesses are now asked to proactively act to retain high-potential users who can help secure the long-term success of the venture. To win the "battle" for users on the platform market, platform owners are called to "arms", attempting and utilizing new methods in marketing, such as monetary incentivization, to foster all aspects of their interaction with the user, from acquisition over activation and retention to intra-company integration.

This information can be useful for managers when assessing at what point of the marketing strategy they currently are and how offering monetary incentives can support them to reach the company goals. It is interesting to see how more companies are opting for using monetary incentives to reach their medium-term goals of retention and activation, with a moderate risk, followed by acquisition and addressing short-term goals. On the other side, long-term objectives can be difficult to measure, making them less attractive for brands.

Among the limitations of the present study, we can highlight the fact that the research focused only on successful international platforms. A further look at smaller platforms acting on certain national or regional markets could reveal new information. Furthermore, the dimensions for the time pattern as well as risk affinity in the theoretical approach could be discussed critically. Additionally, as shown in the literature review section, many businesses are acting like platforms themselves already, but were not considered in this research as they are not categorized as pure platform businesses.

The research shows that many platforms tend to focus on short- to medium-term methods to acquire new users rather than engaging with existing users for the long term and acting on their experience and loyalty to the platform, which could – in return – lead to a higher value creation and a higher return than from unpredictable new users. Further research will show if there is a trend towards user integration on platforms in the future. It would be interesting to see how the distribution of incen-

tivization methods among platforms has changed and will further change over time. Another approach for further research would be to add cultural dimensions to the model, such as the origin of the platform or the respective management. Furthermore, users could be interviewed about their motivation for engaging in a platform, which could be passive (as in reading only) or active (as in influencing the business actively). The most interesting questions after all could be, who benefits most out of the integration-driven approach in the long run – the platform or the user – and what are the main positive outcomes of this innovative method?

References

Adam, Martin, and Johannes Klumpe. "Onboarding with a chat – The Effects of Message Interactivity and Platform self-disclosure on user Disclosure Propensity." In *Proceedings of the 27th European Conference on Information Systems (ECIS)*, Stockholm & Uppsala, Sweden, June 8–14, 2019. Accessed December 19, 2020. https://aisel.aisnet.org/ecis2019_rp/68.

Agrawal, Ajay, Christian Catalini, and Avi Goldfarb. "Some Simple Economics of Crowdfunding." *Innovation Policy and the Economy* 14 (2014): 63–97. https://doi.org/10.1086/674021.

Airbnb. "How does the Airbnb Plus Advance Incentive Program Work? – Airbnb Help Center." Airbnb, online, 2020. Accessed December 19, 2020. https://www.airbnb.com/help/article/2368/how-does-the-airbnb-plus-advance-incentive-program-work.

Aldiabat, Khaldoun M., and Carole-Lynne Le Navenec. "Philosophical Roots of Classical Grounded Theory: Its Foundations in Symbolic Interactionism." *The Qualitative Report* 16, no. 4 (2011): 1063–1080. https://nsuworks.nova.edu/tqr/vol16/iss4/9.

Ashander, Laura, Jana Kliestikova, Pavol Durana, and Jaromir Vrbska. "The Decision-Making Logic of Big Data Algorithmic Analytics." *Contemporary Readings in Law and Social Justice* 11, no. 1 (2019): 57–62. doi:10.22381/CRLSJ11120199

Atwell, Gary J., Eva Kicova, Ladislav Vagner, and Renata Miklencicova. "Parental Engagement with Social Media Platforms: Digital Mothering, Children's Online Privacy, and the Sense of Disempowerment in the Technology-Integrated Society." *Journal of Research in Gender Studies* 9, no. 2 (2019): 44–49. https://doi.org/10.22381/JRGS9220193.

Belleflamme, Paul, and Martin Peitz. "Platforms and network effects." *Working Paper Series University of Mannheim*, 2016. Accessed December 19, 2020. https://madoc.bib.uni-mannheim.de/41306.

Belleflamme, Paul, Thomas Lambert, and Armin Schwienbacher. "Crowdfunding: Tapping the right crowd." *Journal of Business Venturing* 29, no. 5 (2014): 585–609. https://doi.org/10.1016/j.jbusvent.2013.07.003.

Bernard, H. Russell, and Gery W. Ryan. "Analyzing Qualitative Data: Systematic Approaches." California, CA: Sage Publication , 2010.

Blaškova, Martina, Ruta Adamoniene, and Ruta Petrauskiene. "Appliance of Public Senior Executives Competences for Municipality Activity Efficiency Development." *Engineering Economics* 28, no. 5 (2017): 575–584. https://doi.org/10.5755/j01.ee.28.5.17743.

Bratu, Sofia. "Can Social Media Influencers Shape Corporate Brand Reputation? Online Followers' Trust, Value Creation, and Purchase Intentions." *Review of Contemporary Philosophy* 18 (2019a): 157–163. doi:10.22381/RCP18201910

Bratu, Sofia. “Algorithmically Constructed Identities: Networked Digital Technologies, Dynamic Behavioral Big Data Collection, and Automated Decision-Making.” *Contemporary Readings in Law and Social Justice* 11, no. 2 (2019b): 49–55. doi:10.22381/CRLSJ11220197

Brown, Charlie. “3 Questions to Ask Before Adopting a Platform Business Model.” *Harvard Business Review*. Accessed December 19, 2020. https://hbr.org/2016/04/3-questions-to-ask-before-adopting-a-platform-business-model.

Brown, Eileen. “New social network Tsu shares ad revenue with content creators.” Accessed December 19, 2020. https://www.zdnet.com/article/new-social-network-tsu-shares-ad-revenue-with-content-creators.

Bruhn, Manfred. “Kundenorientierung: Bausteine für ein exzellentes Customer-Relationship-Management (CRM) .” dtv Verlagsgesellschaft, 2016.

Bezzubtseva, Anastasia, and Dmitry I. Ignatov. “A Typology of Collaboration Platform Users.” *arXiv preprint* (2013) arXiv:1312.0162.

Culkin, Brigitte. “Is Platform Capitalism Sustainable? Digital Business Models, On-Demand Labor, and Economic Growth.” *Journal of Self-Governance and Management Economics* 7 no. 1 (2019): 31–36. https://doi.org/10.22381/JSME7120195.

Chou, Yu-kai. “Actionable Gamification.” Nederland, ‘s-Hertogenbosch: Van Haren Publishing, 2019.

Dabija, Dan-Cristian, Raluca Băbuţ, Vasile Dinu, and Mădălina Lugojan. “Cross-Generational Analysis of Information Searching based on Social Media in Romania.” *Transformations in Business & Economics* 16, no. 2 (2017): 248–270. http://www.transformations.knf.vu.lt/41/article/cros.

Dabija, Dan-Cristian, Brânduşa Bejan, and Vasile Dinu. “How Sustainability Oriented is Generation Z in Retail? A Literature Review.” Transformations in Business & Economics 18, no. 2 (2019): 140–155, http://www.transformations.knf.vu.lt/47.

D’Orazio, Dante. “Reddit announces it will give $5 million to its users in the form of ‘Notes’.“ online, 2014. Accessed December 19, 2020 https://www.theverge.com/2014/12/20/7427491/reddit-notes-announced-give-5-million-dollars-to-users.

Dropbox. “Freunde zu Dropbox einladen und zusätzlichen Speicherplatz verdienen - Dropbox-Hilfezentrum.” Accessed December 19, 2020. https://help.dropbox.com/de-de/accounts-billing/space-storage/earn-space-referring-friends.

Duolingo. “14-tägige Gratis-Testzeit - Duolingo Plus. Probier’s aus!.” Accessed December 19, 2020. https://www.duolingo.com/plus.

Durlauf, Maria. “The Commodification of Digital Labor in the Gig Economy: Online Outsourcing, Insecure Employment, and Platform-based Rating and Ranking Systems.” *Psychosociological Issues in Human Resource Management* 7, no. 1 (2019): 54-59. https://doi.org/10.22381/PIHRM7120196.

eBay. “The eBay Partner Network.” Accessed December 19 2020. https://pages.ebay.com/seller-center/listing-and-marketing/ebay-partner-network.html.

Eisenmann, Thomas R., Geoffrey G. Parker, and Marshall W. Van Alstyne. “Strategies for Two-Sided Markets.” *Harvard Business Review*, Accessed December 19, 2020. https://hbr.org/2006/10/strategies-for-two-sided-markets.

Evans, David S., and Richard Schmalensee. “Network Effects: March to the Evidence, Not to the Slogans.” *Antitrust Chronicle* (2017): 1–9. https://doi.org/10.2139/ssrn.3027691.

Faggiolani, Chiara. "Perceived Identity: applying Grounded Theory in Libraries." *JLIS* 2, no. 1 (2011): 4592-2-4592-34. http://dx.doi.org/10.4403/jlis.it-4592.

Friedman, Seth. "You Can Use Crowdfunding Platforms as Marketing Research Tools." Accessed December 19, 2020. https://www.youtube.com/watch?v=UWjVDhXejfw.

Furnham, Philipp. "Automation and Autonomy of Big Data-driven Algorithmic Decision-Making." *Contemporary Readings in Law and Social Justice* 11, no. 1 (2019): 51–56. doi:10.22381/CRLSJ11120198

Geng, Daoqu, Chengyun Zhang, Chengjing Xia, Xua Xia, Qilin Liu, and Xinshuai Fu. "Big data-based improved data acquisition and storage system for designing industrial data platform." *IEEE Access 7* (2019): 44574-44582. https://ieeexplore.ieee.org/iel7/6287639/8600701/08681030.pdf.

Glaser, Barney, and Anselm Strauss. "The discovery of grounded theory: Strategies for qualitative research." Aldine, Chicago: Routledge, 1967.

Good Shepherd Entertainment. "Good Shepherd Entertainment." Accessed December 19, 2020. https://www.goodshepherd.games.

Granfeldt, Axel, and Max Nyqvist. "Fostering Network Effects: How to achieve user retention on multisided platforms." Accessed December 19, 2020. https://www.diva-portal.org/smash/get/diva2:1322861/FULLTEXT01.pdf.

Groupon. "Special Programs." Accessed December 19, 2020. https://www.groupon.com/legal/specialprograms.

GrubHub. "Referral Terms – GrubHub." Accessed December 19, 2020. https://www.grubhub.com/legal/referral-terms.

Grudin, Johanna, Marek Durica, and Lucia Svabova. "Labor Market Flexibility in Platform Capitalism: Online Freelancing, Fluid Workplaces, and the Precarious Nature of Employment." *Psychosociological Issues in Human Resource Management* 7, no. 1 (2019): 72–77. https://doi.org/10.22381/PIHRM7120199.

Gutierrez-Leefmans, Manuela, and Christopher Patrick Holland. "SME platforms as business models: A user-centric activity-system approach." *Cuadernos de Administración (Universidad del Valle)* 35, no. 64 (2019): 52–77. http://dx.doi.org/10.25100/cdea.v35i64.7248.

Hagel, John, John Seely Brown, and Lang Davison. "Shaping strategy in a world of constant disruption." 10, 80–89. Boston (2008). Mass.: Harvard Business School Publ. Corp. https://hbr.org/2008/10/shaping-strategy-in-a-world-of-constant-disruption.

Hagel, John, John Seely Brown, and Lang Davison. "Shaping strategy in a world of constant disruption." *Harvard Business Review, October 1, 2008.* https://www.researchgate.net/publication/228395755_Shaping_strategy_in_a_world_of_constant_disruption

Hagiu, Andrei, and David B. Yoffie. "Network Effects." In The Palgrave Encyclopedia of Strategic Management. London: Palgrave Macmillan UK, 2016. https://doi.org/10.1057/978-1-349-94848-2_552-1.

Hansen, Robert A. "A Self-Perception Interpretation of the Effect of Monetary and Nonmonetary Incentives on Mail Survey Respondent Behavior". *Journal of Marketing Research* 17, no. 1 (1980): 77–83. https://doi.org/10.2307/3151120.

Héder Mihaly. "A black market for upvotes and likes." *Információs Társadalom* 19, no. 4 (2019): 18–39. https://dx.doi.org/10.22503/inftars.XIX.2019.4.2.

Hollowell, Jane Catherine, Zuzana Rowland, Tomas Kliestik, Jana Kliestikova, and Victor V. Dengov. "Customer Loyalty in the Sharing Economy Platforms: How Digital Personal Reputation and Feedback Systems Facilitate Interaction and Trust between Strangers." *Journal of Self-Governance and Management Economics* 7, no. 1 (2019): 13–18. https://doi.org/10.22381/JSME7120192.

Katmada, Aikaterini, Anna Satsiou, and Ioannis Kompatsiaris. "Incentive Mechanisms for Crowdsourcing Platforms." Accessed December 19, 2020 https://doi.org/10.1007/978-3-319-45982-0_1.

Konhäusner, Peter, Bing Shang, and Dan-Cristian Dabija. "Application of the 4Es in Online Crowdfunding Platforms: A Comparative Perspective of Germany and China." *Journal of Risk and Financial Management* 14, no. 2 (2021): 49. https://doi.org/10.3390/jrfm14020049.

Lee, Chang-Beom, and Seung-In Kim. "A Study on the Activation Plan of Web-based Open Source Platform using 3D Printer-Focused on Platform Toy." *Journal of Digital Convergence* 17, no. 6 (2019): 341–347 https://doi.org/10.14400/JDC.2019.17.6.341.

Lemke, Roderick. "Digital Services Mediated by Online Labor Platforms: Contingent Work Arrangements, Job Precariousness, and Marginal Social Identities." *Psychosociological Issues in Human Resource Management* 7, no. 1 (2019): 66–71. https://doi.org/10.22381/PIHRM7120198.

Lim, Weng Marc. "The sharing economy: A marketing perspective." *Australasian Marketing Journal (AMJ)* 28 (2020): 4–13. https://doi.org/10.1016/j.ausmj.2020.06.007.

Lukkarinen, Anna, Jeffrey E. Teich, Hannele Wallenius, and Jyrki Wallenius. "Success drivers of online equity crowdfunding campaigns." *Decision Support Systems* 87 (2016): 26–38. https://doi.org/10.1016/j.dss.2016.04.006.

Malzahn, Birte, Peter Konhäusner, and Ngoc Huyen Dao. "Chancen und Hinderungsgründe einer urbanen Mobilitätsplattform aus Anwendersicht. Anwendungen und Konzepte der Wirtschaftsinformation." *Anwendungen und Konzepte der Wirtschaftsinformatik* 11(2020): 71–78. https://ojs-hslu.ch/ojs3211/index.php/akwi/article/view/11/11.

Meilhan, Deborah. "Customer Value Co-Creation Behavior in the Online Platform Economy." *Journal of Self-Governance and Management Economics* 7, no. 1 (2019): 19–24. https://doi.org/10.22381/JSME7120193.

Miller, Grace. "Uber's Marketing Strategy in 7 Steps, Without the Bad Press." *Annex Cloud*, online, 2016. Accessed 19 December, 2020. https://www.annexcloud.com/blog/ubers-marketing-strategy-in-7-steps/.

Annex Cloud. "Uber's Marketing Strategy in 7 Steps, Without the Bad Press." Accessed 19 December, 2020. https://www.annexcloud.com/blog/ubers-marketing-strategy-in-7-steps/.

Mircica, Nela. "Restoring Public Trust in Digital Platform Operations: Machine Learning Algorithmic Structuring of Social Media Content." *Review of Contemporary Philosophy* 19, (2020): 85–91. doi:10.22381/RCP1920209

Mollick, Ethan. "The Dynamics of Crowdfunding: An Exploratory Study." *Journal of Business Venturing* 29, no. 1 (2014): 1–16. https://doi.org/10.1016/j.jbusvent.2013.06.005.

Moysidou, Krystallia, and J. Piet Hausberg. "In crowdfunding we trust: A trust-building model in lending crowdfunding." *Journal of Small Business Management* 58, no. 3 (2019): 511–543. https://doi.org/10.1080/00472778.2019.1661682.

Nadanyiova, Margareta, Lubica Gajanova, Jana Majerova, and Lenka Lizbetinova. "Influencer marketing and its impact on consumer lifestyles." *Forum Scientiae Oeconomia* 8, no. 2 (2020): 109–120. https://doi.org/10.23762/FSO_VOL8_NO2_7.

O'Connell, Bryan. "History of PayPal: Timeline and Facts." Accessed December 19, 2020. https://www.thestreet.com/technology/history-of-paypal-15062744.

Oláh, Judit, Nemer Aburumman, József Popp, Muhammad Asif Khan, Hossam Haddad, and Nicodemus Kitukutha. "Impact of Industry 4.0 on Environmental *Sustainability*." Sustainability 12, no. 11 (2020): 4674, 1–21. https://doi.org/10.3390/su12114674.

Ordanini, Andrea, Lucia Miceli, Marta Pizzetti, and A. Parasuraman. "Crowd-funding: transforming customers into investors through innovative service platforms." *Journal of Service Management* 22, no. 4 (2011): 443–470. https://doi.org/10.1108/09564231111155079.

Parker, Geoffrey G., Marshall W. Van Alstyne, and Sangeet Paul Choudary. "Platform Revolution: How Networked Markets Are Transforming the Economy and How to Make Them Work for You." New York: W. W. Norton & Company, 2016.

Patkar, Mihir. "How to Get Free Trello Gold & What You Can Do with It." *Make Use Of*, online, 2015. Accessed December 19, 2020. https://www.makeuseof.com/tag/get-free-trello-gold-can/.

Make Use Of. "How to Get Free Trello Gold & What You Can Do with It." Accessed December 19, 2020. https://www.makeuseof.com/tag/get-free-trello-gold-can/.

Petrovic, Otto. "A Digital Platform for Marketing Communications in the Mobile and Social Media Space." In: Wojciech Cellary, and Elsa Estevez (eds.). "Software Services for e-World. I3E 2010. IFIP "Advances in Information and Communication Technology". vol 341. Springer, Berlin, Heidelberg. 182–192. https://doi.org/10.1007/978-3-642-16283-1_22.

Petrovic, Otto. "A Digital Platform for Marketing Communications in the Mobile and Social Media Space." In 10th IFIP WG 6.11 Conference on e-Business, e-Services, and e-Society, I3E 2010, 182–192. Berlin, Heidelberg: Springer, 2010. https://doi.org/10.1007/978-3-642-16283-1_22.

Posthumus, Tobias. "Platform Strategy, How to monetize a market with network effects." Accessed December 19, 2020. https://doi.org/10.13140/RG.2.2.14802.04809.

Pedregosa, Carlos Sanchís, Emma Berenguer, Gema Albort-Morant, and Jorge Antón Sanz. "Guaranteed Crowdlending Loans: A Tool for Entrepreneurial Finance Ecosystem Sustainability." *Amfiteatru Economic* 22, no. 55(2020):775–791. https://doi.org/10.24818/EA/2020/55/775.

Pereira, Tony. "Posten | Feed | *LinkedIn*." LinkedIn, online, 2020. Accessed December 12, 2020. https://www.linkedin.com/feed/update/urn:li:activity:6706317651834208256/

LinkedIn. "Posten | Feed | LinkedIn." Accessed December 12, 2020. https://www.linkedin.com/feed/update/urn:li:activity:6706317651834208256/

Pop, Rebeka, Zsuzsa Săplăcan, Dan-Cristian Dabija, and Anetta Monika Alt. "The Impact of Social Media Influencers on Travel Decisions: The Role of Trust in Consumer Decision Journey." *Current Issues in Tourism*, 2021. https://doi.org/10.1080/13683500.2021.1895729

PricewaterhouseCoopers. "Global Top 100 companies - June 2020 update. Global Top 100 companies - June 2020 update." Accessed December 12, 2020. https://www.pwc.com/gx/en/services/audit-assurance/publications/global-top-100-companies.html.

Sayedi, Amin, and Marjan Baghaie. "Crowdfunding as a Marketing Tool." *Available at SSRN 2938183*, 2017. Accessed December 11, 2020. https://dx.doi.org/10.2139/ssrn.2938183.

Sayedi, Amin, and Marjan Baghaie. “Crowdfunding as a Marketing Tool.” Accessed December 11, 2020. https://dx.doi.org/10.2139/ssrn.2938183.

Sârbu, Roxana, Felician Alecu, and Răzvan Dina. “Social Media Advertising Trends in Tourism.” Amfiteatru Economic 20, Special no. 12 (2018):1016-1028. https://doi.org/10.24818/EA/2018/S12/1016.

Ślusarczyk, Beata, Manueala Tvaronavičienė, Adnan Ul Haque, and Oláh Judit. “Predictors of Industry 4.0 technologies affecting logistic enterprises’ performance: international perspective from economic lens.”, Technological and Economic Development of Economy 26 no. 6 (2020): 1263–1283. https://doi.org/10.3846/tede.2020.13376.

Sridhar, Shrihari, Murali K. Mantrala, Prasad A. Naik, and Esther Thorson. “Dynamic Marketing Budgeting for Platform Firms: Theory, Evidence, and Application.” *Journal of Marketing Research* 48, no. 6 (2011): 929–943. https://doi.org/10.1509/jmr.10.0035.

Metka, Stare, and Andreja Jaklič. “Sources of Value Creation in Service Global Value Chains.” *Amfiteatru Economic* 22, no. 55 (2020): 846–866. https://doi.org/10.24818/EA/2020/55/846.

Sion, Grațiela. “Social Media-based Self-Expression: Narcissistic Performance, Public Adoration, and the Commodification of Reified Persona.” *Contemporary Readings in Law and Social Justice* 11, no. 2 (2019): 70–75. doi:10.22381/CRLSJ112201910

Sternberg, Robert J., and Todd I. Lubart. “Defying the crowd: Cultivating creativity in a culture of conformity.” New York: Free Press, 1995.

Suzuki, Takayuki, Kiminori Gemba, and Atsushi Aoyama. “Changes in product benefits contributing to customer satisfaction the case of the digital camera.” *Forum Scientiae Oeconomia* 7 no. 4 (2019): 41–51. https://doi.org/10.23762/FSO_VOL7_NO4_3

Tassi, Paul. “Why It’s Perfectly Fine If Microsoft Has Lost Money on Xbox One [Updated].” *Forbes*, online, 2014. Accessed December 21, 2020. https://www.forbes.com/sites/insertcoin/2014/08/11/why-its-perfectly-fine-if-microsoft-has-lost-400m-on-xbox-one/.

Forbes. “Why It’s Perfectly Fine If Microsoft Has Lost Money on Xbox One [Updated].” Accessed December 21, 2020 https://www.forbes.com/sites/insertcoin/2014/08/11/why-its-perfectly-fine-if-microsoft-has-lost-400m-on-xbox-one/.

Tie, Ylona Chun, Melanie Birks, and Karen Francis. “Grounded theory research: A design framework for novice researchers.”. *SAGE Open Medicine* 7, no. 3 (2019): 1–8. 205031211882292. https://doi.org/10.1177/2050312118822927.

TikTok. “Terms of Service | TikTok.” Accessed December 21, 2020 https://www.tiktok.com/legal/terms-of-use.

Toker-Yildiz, Kamer, Minakshi Trivedi, Jeonghye Choi, and Sue Ryung Chang. “Social Interactions and Monetary Incentives in Driving Consumer Repeat Behavior.” *Journal of Marketing Research* 54, no. 3 (2017): 364-380. https://doi.org/10.1509/jmr.13.0482.

WeAre8. “WeAre8.” Accessed December 19, 2020. https://www.weare8.com/.

Wirtz, Jochen, and Patricia Chew. “The effects of incentives, deal proneness, satisfaction and tie strength on word-of-mouth behaviour.” *International Journal of Service Industry Management* 13, no. 2 (2002): 141–162. https://doi.org/10.1108/09564230210425340.

Zheng, Haichao, Jui-Long Hung, Zihao Qi, and Bo Xu. “The role of trust management in reward-based crowdfunding.” *Online Information Review* 40, no. 1 (2016): 97–118. http://dx.doi.org/10.1108/OIR-04-2015-0099.

Zhu, Feng, and Nathan Furr. “Products to platforms: Making the Leap.” *Harvard Business Review 94*, no. 4 (2016): 72–78. Boston, Mass: Harvard Business School Publ. Corp.

AI and the resurrection of Technological Determinism

This paper elaborates on the connection between the AI regulation fever and the generic concept of Social Control of Technology. According to this analysis, the amplitude of the regulatory efforts may reflect the lock-in potential of the technology in question. Technological lock-in refers to the ability of a limited set of actors to force subsequent generations onto a certain technological trajectory, hence evoking a new interpretation of Technological Determinism. The nature of digital machines amplifies their lock-in potential as the multiplication and reuse of such technology is typically almost cost-free. I sketch out how AI takes this to a new level because it can be software and an autonomous agent simultaneously.

Keywords: *artificial intelligence; technological determinism; social control of technology; AI ethics.*

Acknowledgements

The research was supported by the János Bolyai Research Scholarship of the Hungarian Academy of Sciences

Author Information

Mihály Héder, Budapest University of Technology and Economics; SZTAKI Institute for Computer Science and Control
https://orcid.org/0000-0002-9979-9101

How to cite this article:

Héder, Mihály**.** "AI and the resurrection of Technological Determinism."
Információs Társadalom XXI, no. 2 (2021): 119–130.
https://dx.doi.org/10.22503/inftars.XXI.2021.2.8

Introduction

In this paper I argue that the current wave of Artificial Intelligence Ethics Guidelines can be understood as desperate attempts to achieve social control over a technology that appears to be as autonomous as no other. While efforts at the social control of technology are nothing new, AI with its unique nature may very well be the most resistant to such control, which validates the amount of attention the question receives.

However, should regulatory attempts fail, future society may be determined by the nature of this technology, dread many thinkers. There is an attitude/historiographic methodology called "technological determinism", which has been widely criticized and almost completely dissected since the second half of the 20th century. This attitude is recurrent again in the case of AI, and perhaps has found a more solid footing there.

One pillar of technological determinism is a perceived inevitability about the direction of technological progress, which, like gravity, tends towards ever higher efficiency, and trying to resist it for long is a fool's gambit. The other pillar is that this predetermined nature of technological evolution acts as an exogenous force on society and causes it to change. In other words, technology progresses following its own internal logic and society is restructured as a side effect of this. Consequently, humanity trades its potential for being a Being for an Iron Cage, where only mass-produced Whipped Cream is available but not the real thing.[1]

Social scientists and critically minded philosophers ever since the sociological turn – sometime around the sixties – came up with one case study after another that all showed the surprising causal powers of persons or groups of people on the trajectory of technology. These investigations indicated the reverse of the deterministic view. It appeared that the idiosyncratic decisions of some humans – rooted in their culture, world view, office politics and other factors of this kind, but not in technological reasoning – acted as an exogenous force on technology, rather than the other way around.

For the sake of understanding the relationship between AI and society, this article reconstructs the technological determinist position and investigates the aftermath of the technological determinism debate. On the surface, it might appear that the case is closed and social constructionism has won; at least that the stronger formulations of technological determinism cannot be maintained against the decisive evidence from several case studies. Yet, the general attitude of the deterministic view seems to be resurfacing in discussions around climate change, the effects of social media, and so on. Maybe this is only because of the lack of awareness of the determinism debate and its outcome. But could it also be that some of these technological trajectories – AI being one – are different?

After reviewing the technological determinism landscape, we venture further to examine the notions of technological lock-in, the irreversibility of technology and the existential risk of technology. These seem to suggest that while technology may be indeed socially constructed at a given point in time, later generations have limited freedom in re-interpreting it or phasing it out. In this way, technology may be-

[1] The author apologizes for the conflation of references to Martin Heidegger (1952), Max Weber (1904) and Albert Borgmann (2003).

come an inter-generational tool of power by which earlier generations determine some important aspects of – and even limit the boundaries for – later societies: a possible new type of technological determinism.

The fundamentals of Technological Determinism

Technological determinism refers to the notion that technology shapes society and culture. There is no canon definition for this, rather there are several versions that share a family resemblance with each other. Arguably, the most extreme, hard form of technological determinism, in which there is no place for social control, is quite difficult to defend and as a result it would be hard to find even a handful of serious proponents for it. But the non-existence of the phenomenon is equally implausible. Therefore, technological determinism concepts must be distributed on a scale between these two extremes of full determinism and full indeterminism.

As we look for common features among the several formulations of technological determinism, we will find a claim about *causation* and another about *imbalance*. The first claim considers how a given technology – or sometimes the technologically modern state of affairs in general, i.e. a technological milieu (Ellul 1964) – can be a cause and a feature of an aspect of society (usually a negative one, like the loss of freedom) that arises as an effect of that cause.

The notion of determinism should evoke certain metaphysical concepts. Indeed, some metaphysical theories argue for a completely deterministic view of the universe and therefore every feature of it. Most variants of these theories, and certainly the technological one, are causal determinisms.

If we attempted to interpret technological determinism in this wider framework, perhaps we could point to some causal chains of such a world, in which the role of a technological artefact precedes a change in a feature of society, and hence the cause-and-consequence sequence would be established. The problem of the deterministic world is often discussed in the debate around free will.

But this is usually not how a technological deterministic debate is structured. Instead, such debate tends to sidestep the metaphysical question, or at least assume a world where freedom is at least a theoretical possibility, not dive into the question of whether this means an indeterminate world or a compatibilist (where determinism and freedom can coexist) universe.

While not engaging with the free will problem, the critiques of technological determinism tend to rely on another concept of modern philosophy, namely underdetermination, and the Duhem–Quine thesis of it (Bloor 1991). However, for the purposes of this chapter, it is important to note that underdetermination is primarily considered an epistemic concept.

Besides the freedom of people, as a value that is worth being worried about, debate around technological determinism also often includes a notion of the *autonomy of technology*. This notion of autonomy is even less contrasted with the fundamental question about the deterministic or indeterministic nature of the world as much as the freedom of people. Instead, the autonomy of technology is discussed at the level

of history and of society, as a relative term, as in technology is free from human control, even in Latour's actor-network theory (Latour 2013), where he introduces elements of science and technology as non-human agents.

The second claim, the one about imbalance, grants stronger, more dominant causal powers for technology in comparison with society or culture. This clause is necessary, because it is evident that consumer behaviour, the inventor's and corporate decisions, technology regulation and other venues of human agency – and therefore social control – do have a causal effect on technological artefacts and technological development. Since that is hard to deny, a technological determinist position needs to claim that the role of technology is still more dominant. In spite of all the factors above, this tends to be the decisive factor. This is why, in a technological determinist view, the nature or essence of technology has such big importance: that its essence will eventually manifest itself in the character of society.

However, the case studies from STS and other historic accounts serve as convincing arguments that there is not much point in talking about this issue in very broad, generic terms. Except from some extreme forms of the technological determinist position, the determining powers of technology should differ from case to case, place to place and perhaps even between different historical ages. So a well-formulated technologically deterministic position should state which particular technology has a causal effect on which particular feature of society or culture, instead of making categorical claims about the supremacy of technology in general.

This does not mean that general claims cannot be found. Ellul's (1964) technological milieu concept discusses technologically advanced societies, while occasionally showing some concrete examples of the stated problems. Feenberg (2009) also operates with a concept of technological hegemony, which serves as an ambient background that is beneficial for the causal powers of technology. In this respect his position is similar to Borgmann (2003), who also discusses general tendencies, albeit in a very nuanced manner.

Also, the versions of the theory vary around the role of different groups of people. It is possible to construct theories in which technological, political or economical elites escape being determined by technology, or they may even determine the life of others through technology. Based on this differentiation, the elements of a technologically deterministic theory are often found in political philosophies, like Marxism and its successors, that partially reject and partially elaborate it, like the Frankfurt school, just as well as in other technocratic views of the world.

The stakes of this question are very high since the answer obviously is an input for social organization. A view of the possibilities of taming technology can reinforce our approach to AI among other high-potential technologies.

Cases of social control

There are several supposed examples of technological determinism throughout history. One that is widely stated is the effect of the invention of the printing press on the politics of organized religion on the European continent, or simply put, how

book printing led to reformation. Another example has to do with the invention of the stirrup and feudalism. Yet another concerns gunpowder technology and the colonization of the world by European empires. These accounts of course do not withstand the scrutiny of a more detailed economic–sociological analysis. They usually neglect the possibility of the sociological context having an equally large causal effect on the invention than the invention has on the society. As pointed out in the previous section, in technological determinism it is implied that technology is the dominant force, not the co-evolution of science and technology as equals.

Another issue with these historic examples is that they do not report on negative cases equally, hence violating the well-known principle of symmetry. Bijker, Hughes and Pinch (1992) and other social constructionists of technology use this methodological maxim, inherited and adapted from the strong programme, to describe the necessity of treating technological failures and successes with equal attention. In our context, this would mean contrasting those cases in which a technological breakthrough apparently led to social change with those other cases where a similar technological advancement did not have the same effect.

The use of gunpowder is a very good example for highlighting the need to consider social factors. Pioneered in medieval China, and later adapted by Japan, the Ottoman Empire, the Russian Empire and many others, it did not lead to the same social transformation in those regions as in Europe (Hoffman 2012). So other factors must have been at play. This evidences the necessity for the presence of certain social conditions for change to happen.

If that is the case, we cannot think about these issues with a monocausalistic model anymore. That is, we cannot further maintain that technology is the sole cause of change in these cases, and if it has to share this role with several social factors; thereby the dominance of technology, as encapsulated in the technological determinism concept, is again lost.

And in truth, the social factors are plentiful. Religion and ideology is one obvious candidate to enhance or hinder the acceptance of technological change. Wage levels are often seen as a necessary condition for labour-saving capital expense. War is often cited as a catalyst of technological breakthroughs; albeit if all for the wrong reasons. Also technological determinism not only has to share its influence with social factors but possibly with other forms of determinisms too. For instance, geographical determinism suggests that being on a certain spot on the planet may be decisive. The inhabitants of Easter Island had to face challenges because of the nature of their habitat, just as the Europeans who were denied access to Middle-Eastern trading routes, or the British with access to coal but with the necessity to pump out the water from the shafts; for which, steam power proved to be handy.

Most thinkers when confronted with the implausibility of the extreme positions around technological determinism tend to seek a middle ground. Some thinkers consider their position more in line a form of soft technological determinism (Dusek 2006, Heilbroner 1967).

Another way to find the middle ground is through considering the concept of underdetermination, as Andrew Feenberg does. This solution is especially interesting as it focuses on the co-causal powers of technology and human agency. This view

allows for a theoretician to appreciate the difference between a passive and an techno-politically conscious society.

Feenberg acknowledges that technology, if left alone, has inherently anti-democratic tendencies. He further claims that as more and more social activities become mediated by technology, those tendencies will gain more room to flourish. Therefore, if technology is left alone – instead of actively developing a critical view about it – our freedom will indeed diminish. This is why Feenberg argues for actively injecting democracy into technology and into the technologically mediated areas of life (which are more and more as time progresses); even in areas that were previously thought off-limits for democratic decision-making, like in a factory.

However, Feenberg argues, this really needs to be actively pursued, in order to avoid a natural tendency of society towards becoming ever-more technocratic, and hence less democratic. This means that in his model of the world, change will still happen, but without an active, conscious agency of humans, but also, that without the timely, active participation, our window of opportunity may be lost for ensuring control of that change. Based on this framing, Stump (2006) categorizes Feenberg's view as one that still involves the essentialism of technology.

While Feenberg never uses the following particular terminology from the philosophy of technology, the possibility he explores depends on the co-causation model of social change. In this, there is room for humans to work as a causal component to counterbalance the anti-democratic causal component that technology represents.

The Social Construction of Technology

By describing society as a co-causative factor, we can overcome another deterministic concept, the supposed "trade-off" situation of technology adoption. This considers that society has to make a tough decision about technology: it either uses the technology and suffers its side effects, or it does not adopt it and may be harmed by missing out on the potential advantages and economic growth the technology could bring. This description of the technology adoption problem makes society look external to the technological change; in effect, a mere bystander that needs to make up its mind about a new situation it may find itself in.

The contrary of the trade-off view is Constructivism, or the Social Construction of Technology (SCOT). This position sees the direction of technological change as being underdetermined by mathematics, the laws of nature, or other non-negotiable features of our universe. And if this is undetermined, it means there is room for society to manoeuvre. It has to be noted, that instead of the concepts of underdetermination and co-causation, as commonly used in the philosophy of science, the sociologists who explore this situation tend to rely on expressions from the philosophy of language. As a result, technology is subject to "interpretation" in this terminology. The outcome of this interpretation, of course, is to a great extent up to the users of the language. Translating the analogy for the question at hand yields that the outcome of technological change is up to the makers and users of the technology.

Yet another linguistic concept is the hermeneutics of technology: an iterative interpretation process that society exercises when adopting a new technology. In the context of this process, technical objects have two hermeneutic dimensions. Their social meaning, which is established in a manner that may even be called argumentative, and this defines what kind of role an object may play in the lifestyles of its users. This approach counters the second dimension related functionalism, which considers objects in an inherently de-contextualized manner and views technical objects as neutral means to achieving certain ends that are fundamentally external to the object. Feenberg, by re-contextualizing technology in its social environment, breaks down a hidden assumption behind technological determinism: that rationality is culture-independent.

And if that is not the case, it is also impossible that the trajectory of technological development, which supposedly is about always picking the most rational means to achieve ever-increasing efficiency, is set in the stars. Such rationality will now depend on the social context, and at this point, democratic rationalization is a straightforward possibility and just requires a cultural preference for technological development.

The important conclusion of the above arguments for the possibility of social control over technology is that when it comes to the ontology of technical artefacts, we cannot maintain that one kind of aspect – like functionality or rationality in reaching a goal – is inherent, while other kinds – like social meaning or preferences – are assigned just in the observer's mind. Instead, we must conclude that these aspects are equally essential to the given artefact. This is the "double aspect theory". If there is an unreflexive interpretation process – that is, an interpretation that does not acknowledge that it analyzes the objects with completely idiosyncratic preconceptions of effectiveness – technology will indeed appear as an external force on society.

This raises the question on whether anything has actually fundamentally changed our "modern world", meaning our digital, virtual, and industrialized worlds. This question is crucial, since humanity has always been technical – in fact, elaborated tool usage is a common milestone in the historical accounts of the evolution of our species.

Increasingly powerful technologies

In the first waves of the technological determinism debate, the *autonomy* of technology meant an abstract situation in which the nature of technology keeps relentlessly manifesting itself though the rationalization efforts of humans. We saw how the double aspect theory questioned whether this is inevitable. However, what if a technology – several instances of it to be more precise – is more *literally* autonomous, like AI? My argument is that in this case we have to deal with technology of a different nature, rendering most of the constructionist arguments irrelevant. But before we get there, it is best to build up a picture through considering other, equally recent, technological achievements.

Take social media. In the 2020s, it is a common argument that the nature of political campaigning has drastically changed thanks to this technology and that its users

grip on reality may already be incorrigibly broken. This is a picture in which the medium dominates the discourse, which has fuelled renewed interest in the works of McLuhan, somebody who is usually categorized as a technological determinist, but who later in his career softened his stance somewhat.

On the internet, and particularly social media, there is support for seemingly any claim no matter how far-fetched with reports on evidence that would seemingly corroborate it. This is just another form of increase in control – paradoxically, we seem to be able to control and entrench our own beliefs even by tendentiously selecting the content we consume. But the control is not evenly distributed, it is affected by AI and somewhat determined by our biology. In criticisms of social media, references about how we are being manipulated through targeting our dopamine centres are common.

The time dimension of being dominated this way is especially interesting. In the case of social media, it has been stated that it takes sustained conditioning over some amount of time to arrive at a drastically polarized society, in which the camps are not capable of having discourse anymore due to their incompatible perceived realities and semantics. At this point, the positions become so entrenched that it seems there is no way back anymore.

The time horizon appears to be even more important as we arrive at technologies that are able to change the environment, cause climate warming and environmental pollution in general. Here the urgency for action is derived from the predictions that the window of opportunity is closing – in fact, for the most positive scenarios it has already slammed shut. The immediate importance of this topic was already evident and the situation seemed already dire in 1999 when Feenberg's *Questioning Technology* (2009) dedicated a chapter to environmentalism and the surrounding politics. Andrew Light (2006) added interesting further thoughts to the debate that the chapter analyzes.

I think that the very urgency that everybody exhibits around the issue of the environment and global warming illustrates a realization of the possibility of irreversible negative change. But that irreversibility in turn means that the world is changing in a way that the arguments against technological determinism and on behalf of social control become less and less convincing.

This evokes a concept that is reasonably present in the management and history of technology but curiously underrepresented in science and technology studies.

The concept in question is technological lock-in.

Based on the terminology of the previous sections, we can summarize technological lock-in as a process where the possibility of change technology is gradually lost as the window for modifications becomes closed.

There are multiple reasons for this. David (1985) identifies technological co-dependence, economies of scale and irredeemable investments as key reasons. This is expanded by Cowan and Hultén (1996) with several new factors, like the necessity of a crisis, regulation, and technological breakthroughs for changing an incumbent technology, while the lack of these means the status quo being sustained. Foxon (2014) further elaborates the role of institutions and the epistemic aspect in general.

It seems logical that a concept of irreversibility is necessary for explaining how

the unfortunate situation of technological lock-in may occur. In the next section, we take the case of irreversibility to the extreme.

AI enters the scene

As the presence of new technology in our everyday life increases, sometimes the general public may suddenly become alarmed by its towering presence. It is not clear when exactly this happens. For instance, in the last years of the 2010s several regulation efforts all around the world were launched to handle the ethics of Artificial Intelligence. Global institutions like UNESCO, professional bodies like the IEEE (2019) and the European Union (2020), and several other organizations and companies made declarations in this area in or close to 2019 (Héder 2020). Their urgency appeared quite similar to what we are experiencing around global warming, and I argue that the reason is the same: AI has a tremendous lock-in potential.[2]

There are several factors that make AI especially prone to being locked in.

First, AI is *software*. Like with any software, the cost of "manufacturing" – producing more instances of the same design, which is copying in this case – is ridiculously low; indeed in most cases, completely negligible. And yet, a profit may be realized on each "unit" or licence, meaning that creating well-received software can be extremely lucrative – write once, derive profits over and over again. Indeed, the most successful track for social mobility seems to be creating and owning software and related IT. Many of the wealthiest people in the world, unless inheritance was in play, rose up by developing some successful software – think of Gates, Bezos, Musk, Zuckerberg, etc. And, of course, thanks to the internet, not only "manufacturing" but also "delivery" (downloading) is basically costless with software.

This means that if in the future a problem class is quite successfully solved by a piece of AI software that is also reasonably available – free or cheap, considering the value – then there will not be much incentive to develop alternatives. In reality, this rarely means a complete monopoly over a problem class and there is always a small number of commercial and some open-source competitors, but if we think about it, many categories of software today are covered by extremely few options. Think about the number of pdf readers or web browsers you use. There is more than one, but the list ends surprisingly quickly, and there is also the fact that some of these are really the same under the hood, but with different interfaces.

If the software in question is also free and open source with a licence that is compatible with most interests, the dominance of one single solution can become extreme. A case in point is the Linux kernel that is present in any android-powered device and that serves the overwhelming majority of web pages and can be found in billions of smart appliances. It really is like a stick-and-carrot situation – writing your own operating system is insurmountable except for the largest institutions, while on the other hand, reusing what is already there is free. Now, this only means

[2] There are, of course, other theories as to explain the sudden surge in AI ethics, like the extension of Politics to regulation (Gyulai and Ujlaki 2021), or simple „ethics washing" (Vică et al. 2021).

that the people that have a say in the development of the software in question have an oversized control over an entire industry, so they need to be engaged on various platforms in order to achieve the social control of technology. However, the situation is worse than that. In fact, many of these projects are inter-generational and the current shepherds of any single technology may have limited control over the trajectory of software, especially if the software is already ubiquitous and any significant rewrite would require more effort than the current generation can offer.

On top of the digital nature of software – which I argue, is enhancing its lock-in propensity – there is now the phenomenon of Software-as-a-Service (SaaS, or, vaguely, the "cloud"), which mobilizes economies of scale, in this case for data. This elevates the lock-in potential of software to an entirely new level. By aggregating several users and use cases, companies offering SaaS can leverage the network effect between those users for their own benefit. While copying and delivering software is negligibly cheap, there is still a cost to using non-SaaS software, mainly installation and maintenance costs. With SaaS, these costs, too, are greatly reduced. This creates situations akin to natural monopolies: the author of this paper surveys his Ethics of AI students each semester, and always finds that there is a 100% penetration of Gmail among the students surveyed. This is despite the fact that really absolutely nothing prevents anyone from running a similar service.

The already unusually ample lock-in potential of the combination of software with the internet (SaaS) is further enhanced by a particular feature of AI: the need for data for machine learning. Artificial Intelligence delivered as SaaS has a unique potential that no other distribution method can match. Therefore, we can expect, with some confidence, that whenever a SaaS AI becomes sufficiently good in the targeted problem space – e.g. a translator or proofreader solution – then it will become simply uneconomical to compete against it.

Finally, this picture would be completed with the possibility of a self-enhancing, ever-more autonomous SaaS AI, which is really one of the promises of machine learning. This would enable over time opening a gap between any new contenders and an established solution in a problem space – for the benefit of the incumbent.

An autonomous – which in this case only means self-driven, proactive intelligent behaviour – AI agent present entirely different problems for social control. Regardless of what phenomenological state we ascribe to such an agent, the interactive nature of such machines will make them actors rather than mere objects. Suddenly, in the debate around technological determinism, these agents may appear on the other side of the equation, the one that has so far been reserved for humans only. And this truly counts as the resurrection of the technological determinism debate.

Discussion

This article summarized some of the positions around technological determinism for understanding the reception of contemporary AI. To analyze the various shades of technological indeterminism and social control, we used – sometimes inspired by the STS literature itself – the terminology from the philosophy of science, namely the

epistemic concept of underdetermination, the criticism of monocausation and the arguments for co-causation in place of it.

However, the Quineian arguments for the undefeatable underdetermination of theories by empiria are not arguments for the underdetermination of change by all the factors that we cannot control. In fact, there is no guarantee that all the relevant processes we care about – for instance, change in society – will always be controllable as well.

In this article, I explored a dynamic view of the balance between the primacy of technological and social factors. Specifically, I posited the question of whether this may shift over time, and not to the advantage of society. This idea of course is nothing new: irreversible environmental change and technological lock-in have both been commonly discussed for several decades. It is a fair and existential question then whether the means of technological power and social control are in such an imbalance.

The nature of scientific knowledge and engineering knowledge – that it is easier to reuse than to discover, easier to copy than to design – suggests that it is easier to increase the general level of technological prowess than to decrease it. In other words, the margin cost of reusing knowledge is diminishingly low. Extrapolating this thought to digital technology, we found that AI is especially interesting, since the multiplication and reuse of such technology is typically almost cost-free.

I sketched out how AI could be such a technology, by the virtue of it being software, but more specifically Software-as-a-Service. This enhances the lock-in potential of AI as all the necessary conditions of technological lock-in are present: fast dissemination and an uncommonly strong economical factor for reuse instead of re-creation, turbo-charged with the economies of centralized data collection for the sake of machine learning.

Finally, I touched on the question of whether technology can have actual agency, instead of the metaphorical agency the proponents of technological determinism have suggested before. This would mean AI agents appearing as relevant social groups in the shaping of their own trajectory, and thereby completely re-framing the debate of technological determinism.

References

Bijker, Wiebe, Thomas P. Hughes, and Trevor Pinch. “The Social Construction of Technological Systems.” In *Shaping Technology/Building Society: Studies in Sociotechnical Change*, edited by Wiebe Bijker and John Law, Cambridge: MIT Press, 1992.

Bloor, David. *Knowledge and Social Imagery*. Chicago: University of Chicago Press, 1991.

Borgmann, Albert. *Power failure: Christianity in the culture of technology*. Baker Books, 2003.

Cowan, Robin, and Staffan Hultén. “Escaping lock-in: the case of the electric vehicle.” *Technological forecasting and social change* 53, no. 1 (1996): 61–79.

David, Paul A. “Clio and the Economics of QWERTY.” *The American economic review* 75, no. 2 (1985): 332–337.
Dusek, Val. *Philosophy of technology: An introduction*. Blackwell, 2006.
Ellul, Jacques. *The Technological Society*, trans. J. Wilkinson, New York: Vintage, 1964.
European Commission. *On Artificial Intelligence – A European approach to excellence and trust*. COM 65. Accessed: June 30, 2020. https://ec.europa.eu/info/sites/info/files/commission-white-paper-artificial-intelligence-feb2020_en.pdf
Feenberg, Andrew. *Questioning Technology*. Routledge, 1999.
Feenberg, Andrew. “Democratic rationalization: Technology, power, and freedom.” In *Readings in the philosophy of technology*, edited by David M. Kaplan, 139-155., Lanham, MD: Rowman and Littlefield, 2010.
Foxon, Timothy J. “Technological lock-in and the role of innovation.” In: *Handbook of sustainable development*, edited by Giles Atkinson, Simon Dietz, Eric Neumayer, and Matthew Agarwala, Edward Elgar Publishing, 2014.
Gyulai, Attila and Anna Ujlaki. “The political AI: a realist account of AI regulation.” *Információs Társadalom* 21, no. 2 (2021). https://doi.org/10.22503/inftars.XXI.2021.2.3.
Heidegger, Martin. “The Question Concerning Technology.” (QCT), in *The Question Concerning Technology and Other Essays*, New York: Harper Collins, 1952.
Heilbroner, Robert. L. “Do machines make history?” *Technology and Culture* 8, no. 38 (1967): 335–45 (also in Scharff and Dusek, pp. 398–404).
Héder, Mihály. “A Criticism of AI Ethics Guidelines.” *Információs Társadalom* 20, no. 4 (December 31, 2020): 57–73. https://doi.org/10.22503/inftars.XX.2020.4.5.
Héder, Mihály. “The Epistemic Opacity of Autonomous Systems and the Ethical Consequences.” AI & SOCIETY, (July 30, 2020b). https://doi.org/10.1007/s00146-020-01024-9.
Hoffman, Phillip T. “Why was it Europeans who conquered the world?” *The Journal of Economic History* 72, no. 3 (2012): 601–633.
IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems. *Ethically Aligned Design: A Vision for Prioritizing Human Well-being with Autonomous and Intelligent Systems,* First Edition. Last Accessed: Dec 20, 2019. https://standards.ieee.org/content/ieee-standards/en/industry-connections/ec/ autonomous-systems.html
Latour, Bruno. “Reassembling the social. An introduction to actor-network-theory.” *Journal of Economic Sociology* 14, no. 2 (2013): 73–87.
Light, Andrew. “Democratic technology, population, and environmental change.” In *Democratizing technology: Andrew Feenberg’s critical theory of technology, edited by* Tyler J. Veak, SUNY press, 2006.
Stump, David. “Rethinking modernity as the construction of technological systems.” In *Democratizing technology: Andrew Feenberg’s critical theory of technology, edited by* Tyler J. Veak, SUNY press, 2006.
Constantin Vică, Cristina Voinea, and Radu Uszkai. “The emperor is naked: moral diplomacies and the ethics of AI.” *Információs Társadalom* 21, no. 2 (2021): 83–
https://doi.org/10.22503/inftars.XXI.2021.2.6
Weber, Max. *The Protestant Ethic and the Spirit of Capitalism*, T. Parsons (trans.), A. Giddens (intro), London: Routledge, [1904] 1992.