Johan Eddebo & Anna-Sara Lind

Artificial Intelligence and Imperceptible Governance via Opinion Formation: Reflections on Power and Transparency from a Cross-Disciplinary Encounter[1]

1 Introduction

In a society characterized by a bourgeoning and increasingly ubiquitous digital infrastructure, hardly any field is left untouched by the increasing reliance on artificial intelligence (AI). This is also true for governance, that is, how governments and organizations steer and control behavior. One aspect of the impact of AI on governance that has received much attention (see for example in this volume the contribution by Markku Suksi) is the incipient use of automated decision-making (ADM) in the public sector. Caution is increasingly recommended due to the risks and issues relating to the introduction AI within present frameworks of governance and decision-making.[2] Examples of risks associated with the use of ADM in the public sector are lack of transparency and effective accountability, which can be partly due to technological and organizational issues, and partly to a lack of clear regulatory provisions on the legislative side.

[1] The authors are grateful for valuable inputs and comments by Ass. Professor Sandra Friberg, PhD Oliver Li and Assistant Prof. Katja de Vries.

[2] See *Dagens Nyheter*. "Myndighetsbeslut måste alltid vara rättssäkra – oavsett om de fattas av människor eller maskiner". Online: https://www.dn.se/ledare/myndighetsbeslut-maste-alltid-vara-rattssakra-oavsett-om-de-fattas-av-manniskor-eller-maskiner/.

All the same, it must be stated that algorithmic systems already exert a significant influence over decision-making processes, both directly and indirectly. Notwithstanding their formal integration in explicit governance, AI is already a substantial factor in terms of the formation of public opinion and political consent,[3] particularly in the conduct of communications and flows of information in digital contexts.[4] Such influence is arguably a factor in all flows of information in digital contexts where algorithms influence the perceptions of the recipient in any way. As a general remark, we distinguish between three levels of AI-systems where mainly two of these are central to the discussion. First, we have the simple algorithms used in such circumstances as data-filtering processes, such as a bit of code that picks out posts with a certain frequency of listed keywords. Then we have the more advanced systems based in evolving or self-learning algorithms basically capable of processing large amounts of data, and then "extrapolating" factors relevant for future decisions. These systems can be characterized by an inherent unpredictability even from the programmer's perspective. Thirdly, there is the hypothetical category of strong AI exhibiting behaviour indistinguishable from that of human agents, in principle able to perform any type of decision-making task. This category will not be addressed in this chapter, however.

This chapter will focus on the implicit governance exercised through AI which arguably already is in place and expanding, and give special attention to a form of complex or layered opacity peculiar to the phenomenon. That is, indirect algorithmic governance effected by e.g. private tech-platforms firstly employ AI systems which are proprietary and inaccessible to external review. In our contribution, we assume a general definition of governance as processes of policy creation involving different actors and networks, which impact upon social formation and the reproduction or establishment of institutions. Secondly, it is very difficult to get a sense of the actual effects of this automated discourse manage-

[3] See SOU 2014:75 pp. 31–32; Young Mie Kim, "Algorithmic Opportunity: Digital Advertising and Inequality in Political Involvement", *The Forum*, 14 (4), 2016.

[4] Cf. Samuel C. Woolley, Philip N. Howard (eds.), *Computational Propaganda*, New York: OUP 2019; Ujué Agudo, Helena Matute, "The influence of algorithms on political and dating decisions". *PLoS ONE* 16(4), 2021. Online: https://doi.org/10.1371/journal.pone.0249454; Riksrevisionen, Automatiserat beslutsfattande i statsförvaltningen – effektivt, men kontroll och uppföljning brister (RiR 2020:22). The shaping of consent via mediatic processes is a contentious topic all by itself, beset by significant issues relating to the principles of rational, unguided and uncompelled democratic deliberation.

ment due to the unpredictability of the underlying algorithms and the lack of access to the platforms' private data on traffic and information flow. We will argue for the importance of mitigation strategies, present a set of viable options, and discuss legislative possibilities. The chapter's outline is as follows. First, in section 2, a brief description of the current situation will be presented. Here we aim to sketch the frames of the issues that we would like to study closer and we explain the arguments as to how AI is in effect exerting de facto governing functions. Thereafter, in section 3, our attention turns to the concept of *double intransparency* and how it is manifested in decision-making processes. In section 4, possible mitigation strategies, such as for example the Artificial Intelligence Act, are discussed but also other legislative options are presented in light of our results (section 5). We then conclude our contribution (section 6) with observations relating to our discussions and findings.

Furthermore, it should be addressed that this chapter has been written jointly by a philosopher of religion and a legal scholar.[5] The dialogue between the authors has been accordingly done across disciplinary borders and is worth reflecting upon. As the attentive reader will see, the disciplinary backgrounds of these two authors do have an impact on how our questions are asked and on how the discussion is framed. In the chapter's concluding part, we will accordingly come back to this and reflect upon our encounter.

Methodologically speaking, philosophy's general approach is to scrutinize the meaning of abstractions, in terms of everything from pure concepts to established social institutions. When the discussion regards a complex social situation like the present one, a useful way to proceed is to then explore the possible theoretical and structural implications of the relevant abstractions. Here one attempts to discern which consequences, applications or developments are likely or even inevitable in principle (or vice versa) when these abstractions are taken to regulate or guide social processes. This discernment is also preferably anchored in supporting empirical data or material expressing the intentions of the institutions and agents involved.

This type of general theoretical overview may come off as naïve from the point of view of jurisprudence or the social sciences, which are more

[5] The authors are part of the national research program WASP-HS for more information see wasp-hs.org) and collaborates in the project Artificial intelligence, democracy and human dignity.

familiar with the details of the complex limitations and possibilities of the social structures involved. The advantage is that this perspective may also afford novel solutions difficult to discern from within these inevitably entrenched specialized disciplines.

In working with this chapter, we have attempted to proceed by first establishing philosophy's more unfiltered speculative suggestions, and then relating them to actual legislation and juridical practice. All through the working process, we have had a continuous dialogue and several meetings exchanging views and learning from each other's field. This is also shown in the text as it to a large extent mirrors the ongoing dialogue between the authors.[6] Our experience is that this multidisciplinary dialogue can add new insights to the respective disciplines involved but also contribute to developing new questions.

2 The current situation – points of departure

In February 2021, Facebook announced that 97% of all "hate speech" was pre-emptively detected and removed by their automated systems before any human had flagged it.[7] Their proactive removal was said to rely upon a complex contextual analysis of language and the communicative setting. The interrelation of text, comments and images was ostensibly taken into account, which was said to enable a high accuracy of the automated decisions. Hate speech is an increasingly prominent concept in the contemporary political discourse, a conceptual construct characterized by a certain ambiguity, which in the case of targeted suppression or censorship efforts adds another level of transparency issues.[8] Correspondingly, Facebook defines hate speech as any type of communication which attacks people in relation to their "protected characteristics", while adding that there is no consensus in terms of exactly what amounts to a transgression in this sense.[9]

[6] Compare with Lind, A-S, *Den offentliga rätten i mångvetenskaplig forskning*, pp. 207 ff.
[7] Schropfer, Mike, "Update on Our Progress on AI and Hate Speech Detection", *Facebook* 2021. Online: https://about.fb.com/news/2021/02/update-on-our-progress-on-ai-and-hate-speech-detection/.
[8] Cf. Brudholm, Thomas, Johansen, Schepelern Brigitte (eds.) *Hate Politics Law*, New York: OUP 2018, p. 5–11.
[9] Richard Allan, "Hard Questions: Who Should Decide What Is Hate Speech in an Online Global Community?", *Facebook* 2017. Online: https://about.fb.com/news/2017/06/hard-questions-hate-speech/.

These kinds of assurances of successful and effective detection are quite optimistic, particularly in relation to the complexity added by the formal ambiguity of hate speech as a concept, and one ought to be sceptical of the validity of such purported automatic assessments. This level of complex estimation is namely difficult even for human persons. It arguably necessitates a thorough familiarity with the cultures, languages and social settings involved, which renders doubtful the accuracy of such automated flagging which at best can mark patterns of symbolic association from predetermined directives such as lists. Whereas a human can rationally recognize a certain kind of contextually relative speech act and even empathize with the intention of its originator so as to actually understand it, the AI will at best only flag possible and probable associations. Moreover, the human need not rigidly adhere to a fixed set of symbols or list of connections, but can make a reasoned assessment of social interactions and possible breaches of trust and etiquette in a fluid environment.

All this is to say that the AI is likely to err in ways human persons would not, while sometimes also being prone to reinforce human error, such as the aggravation of biases.[10] To this we must add that the proliferation of automated AI review and censorship will literally impact trillions of interactions. We are here in a sense dealing with something akin to the butterfly effect, where a small change in the initial conditions of the algorithm, especially if self-learning, will likely produce immense and unforeseeable effects upon interaction patterns and the flow of information.[11] The potential ramifications are varied and far-reaching, and makes the admittedly massive influence of traditional mass media seem rather primitive and superficial in comparison.

That interference at this scale is likely to be characterized by structurally proliferated errors in judgment is severely problematic. Of greater importance, however, are the potential political effects of this type of automated interventions. Remaining with the example of hate speech suppression, one or the first remarks often made by scholars is that the very definition of the concept is rife with ambiguity and contradiction.[12] The accounts

[10] Julia Angwin, Jeff Larson, Surya Mattu & Lauren Kirchner, "Machine Bias", *Propublica* 2016. Online: https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing.

[11] Robert M. Entman, Nikki Usher, "Framing in a Fractured Democracy: Impacts of Digital Technology on Ideology, Power and Cascading Network Activation", *Communication Theory* vol. 68, no. 2 2018.

[12] Alexander Brown, *Hate Speech Law*, New York: Routledge 2015, p. 4–5.

of hate speech outside the field of law are quite varied, in accordance with the diversity of the fields researching the concept, such as sociology, psychology, linguistics and political science, while already jurisprudence as such exhibits a broad range of competing descriptions.[13] Such ambiguity coupled with large-scale suppressive interventions by private operators of the digital infrastructure renders the politicization of hate speech countermeasures a distinct possibility. As a kind of speech act which must be defined contextually, in relation to its purpose and effects, hate speech as such can often not be unambiguously specified,[14] and the actual stipulations thereof arguably give a great leeway for politicized interpretations.[15]

A related example is found in Facebook's recent push towards stifling that which is designated "political extremism". In practice, algorithms filter out posts and comments according to undisclosed criteria that the user cannot access, and issue warnings to the poster's contacts as well as recipients of the information. The poster is on the other hand not notified of the warnings, which either amount to implying that an acquaintance of the recipient might be "becoming an extremist", or informing the recipient that he or she "may have been exposed to harmful extremist content recently".[16] The obvious consequence that friends and acquaintances of the originator of content designated as politically problematic will be targeted, ostensibly with the purpose of creating an indirect social pressure to counteract such perspectives deemed as politically problematic.

"Extremism" as a concept is characterized by ambiguities far more significant than those regarding hate speech. As an unqualified noun, its function is entirely relative to some implicit norm, and can in practice refer to any kind of political position whatsoever. The extremism decried by Justin Trudeau will very likely differ from the one opposed by Viktor Orbán. However, Facebook reassuringly lets us know that it cooperates with NGOs and academic experts in developing these algorithmic countermeasures against politically undesirable content, so we should be confident that the system will not be abused.[17]

[13] Cf. e.g. Glasser 1994; Vasquez and de las Fuentes 2000; Fraleigh and Tuman 2011.

[14] Stavros Assimakoupoulos, Fabienne H. Baider, Sharon Millar, *Online Hate Speech in the European Union*, Cham: Springer 2017, p. 3–4.

[15] Cf. Nadine Strossen, *Hate*, New York: OUP 2018, Chs. 4 & 7.

[16] BBC 2021.

[17] Nawab Osman, Adam Burke, "New Resources to counter Hate and Extremism Online". *Facebook* 2021. Online: https://about.fb.com/news/2021/06/new-resources-to-counter-hate-and-extremism-online/.

All the same, the formal and structural problems remain in place, while private authorities independently and without much oversight providing the key definitions for a far-reaching automated discourse formation. This kind of interference, and at the massive scale the interrelated social media corporations operate, arguably amounts to a form of de facto governance in practice, and holds an obvious potential to exert political influence in myriad ways.

The establishment of an AI-based "counter-disinformation" framework by institutions such as Facebook is one particularly clear example.[18] The approach here is to simply suppress politically undesirable content by either removing the posts as such, or demoting them so that they are unlikely to appear in various feeds and will spread little when shared. These processes will in turn affect the visibility of the primary media outlets which to some extent depend upon the tech platforms as infrastructure, and strongly encourage an adaptation of their content in line with that which is selected as acceptable by the custodians of the platform.

This is inevitably going to shape discourses and opinion formation in the public sphere, with potential ramifications for any kind of decision-making that can be influenced by prominent media narratives. Interference of this kind must therefore be considered a type of governance even in accordance with more stringent definitions of the concept, since it actively delineates acceptable public policy and reproduces public consent for preferred positions.[19]

Another example which touches upon the broad range of further possibilities in terms of governance is the tech platforms' proactive "nudging" of users who have simply interacted with politically objectionable content. The concept, coined by Nobel Laureate Richard Thaler, entails the preemptive shaping of an informational environment so as to control the outcome of subsequent decisions.[20] In practice, certain users who have "liked" or otherwise interacted with undesirable content are then targeted with information promoting divergent perspectives or simply negating the content in question. The actual efficacy of these interven-

[18] Linda Slapakova, "Towards an AI-based Counter-Disinformation Framework", *The Hague Diplomacy Blog* 2021; Techcrunch 2021 https://techcrunch.com/2020/05/12/facebook-upgrades-its-ai-to-better-tackle-covid-19-misinformation-and-hate-speech/.

[19] Michael Barnett, Raymond Duvall, *Power in Global Governance*, New York: Cambridge University Press 2005, p. 15.

[20] Richard Thaler, *Nudge: Improving Decisions About Health, Wealth and Happiness*, New Haven: Yale University Press 2008.

tions are supported by research indicating that, at least in certain situations, these types of interference in the dissemination of marginalized narratives successfully discredit the targeted stories about half the time. In other words, after being confronted with a seemingly authoritative correction, half of the users rescinded belief in the addressed narratives.[21] Even minimal nudging has been shown to change the consensus perspective towards a preferred position, and merely simple filtering algorithms rather than complex self-learning algorithmic systems seem sufficient for an effective regulation of opinion formation.[22]

3 The double intransparency of the imperceptible influence of AI over decision-making processes and opinion formation

Transparency is an important aspect of the exercise of power in open societies for several reasons. Most obviously, it is an important part of the rule of law. It safeguards accountability for errors and abuses, which depends on a visible chain of decisions and a clearly established causation. Transparency is naturally also key with regard to a functional democratic influence over the decision-making processes in society, without which there can be no proper open deliberation over governance and the distribution of power. In Swedish constitutional law, transparency is expressed in several ways. It is embraced and promoted in the Freedom of the Press Act where the right to access to official documents is stated.[23] Transparency is also achieved through the constitutional demand that the work of courts and legislative bodies should be done publicly.[24] In the Administrative Procedure Act documentation and motivation of decisions are expressed

[21] Avaaz, "White Paper: Correcting the Record", *Avaaz.com*, 2021. Online: https://secure.avaaz.org/campaign/en/correct_the_record_study/.

[22] Nicola Perra, Luis E. C. Rocha, "Modelling opinion dynamics in the age of algorithmic personalization", *Scientific Reports 9*, 7261, 2019. Online: https://doi.org/10.1038/s41598-019-43830-2.

[23] Chapter 2 Section 1 Freedom of the Press Act. See also the contributions made by Johan Hirschfeldt and Anna-Sara Lind respectively in the anthology Transparency in the future – Swedish Openness 250 years (Lind, Reichel and Österdahl, Eds.), 2017. See also Axberger, pp. 43–44 and Lind (2015).

[24] See for example Chapter 2 Section 11 para. 2 Instrument of Government and Chapter 5 Section 42 the Local Government Act (2017:725).

as a rule and a general demand for all public bodies and those who have been given the right to take administrative decisions.[25]

Indeed, even in an authoritarian society, a minimal level of transparency with regard to the exercise of power is arguably necessary to enable regulatory oversight and to avoid abuse and the entrenchment of corruption, if for no other reason than to maintain basic safety, the efficiency of production, and social cohesion and stability.[26]

And given what is mentioned in the previous section pertaining to the current situation, there's an obvious argument that the influence exercised through algorithms within the framework of digital communication platforms and contemporary media technologies exhibits a complex form of intransparency. This is more specifically what we referred to as "double" intransparency in the introduction, since it pertains to both our inability to access the algorithms as such, as well as the difficulties of reproducing and examining their actual effects.

To begin with, there is no clear and reliable way for the end-user in a social media framework to ascertain whether he or she is being subtly "nudged" by having his information feed or search results altered in relation to an algorithmic assessment of what information is deemed appropriate, or even if the data is being tailored in accordance with the user's traced prior activity. The information being presented to us may have been algorithmically prioritized to the detriment of our access to other information, or with the purpose of discrediting certain narratives and perspectives, and there is no obvious way to determine whether or not this is the case. Neither is it in most instances possible for the user to find out if any of the information he or she has passed on is in any way suppressed or diverted, as in the case of "shadow banning" (although irregular patterns in others' interactions can possibly be a sign). Shadow banning is the practice whereby a user's communication is suppressed, e.g. by downranking his or her content to that it is less visible or almost invisible in others' news feeds. This is naturally also much less conspicuous than the outright banning of a user. Marked and sudden reductions in interactions is the only obvious indication that something like this may have taken place.

[25] Sections 27, 28, 31 and 32 Administrative Procedure Act.

[26] In *Transparency and Authoritarian Rule in Southeast Asia* (London: Routledge 2004), Garry Rodan for instance argues that transparency measures beneficial from an efficiency perspective have been implemented in Southeast Asia while indirectly supporting rather than challenging authoritarian rule.

If we then add to the equation the self-learning aspect of advanced AI, we also have an inherent unpredictability in the situation which essentially precludes full transparency, even from the perspective of those providing the initial programming. In other words, a self-learning decision-making algorithm may effect changes in relation to a fluid information environment which are in principle unpredictable at the outset.

When AI systems influence opinion formation at a very large scale (Google for instance serves almost 4 billion search queries per day),[27] we are faced with a situation of an almost imperceptible influence over popular opinion, over media narratives, and indirectly parliamentary processes. We have no real access to many of the filtering algorithms since they are proprietary (or at least trade secrets) and difficult to reverse engineer, and any user data from which researchers could empirically infer interaction patterns and the presence of bias or active influence is likewise privately held.

Moreover, the influential interactions between the user and something like politically modified search results are transient and short-lived. They generally cannot be recreated after the fact due to the changing information environment, nor registered in any straightforward way, which renders them nearly impossible to audit.

When we then establish the active influencing of popular opinion and political processes using these kinds of tools, a qualitatively new, and quite undetectable form of governance, has essentially been set up. To be sure, an entity like Facebook has only utilized these technologies of control in relation to contentious issues such as hate speech or purportedly false information, yet they also refuse to explicitly repudiate a broader usage:

> Facebook declined to answer a question from Recode about whether it will apply its warnings to other types of misinformation in the future.
>
> For companies like Facebook, it's a lot easier to draw a line in the sand on misinformation about coronavirus topics than around more politically contentious ones, like gun rights, abortion, immigration, or even the 2020 US elections.[28]

[27] Internet Live Stats 2021. Online: https://www.internetlivestats.com/google-search-statistics/.

[28] Shin Ghaffary, "Facebook will start nudging users who have 'liked' coronavirus hoaxes", *Recode* 2020. Online: https://www.vox.com/recode/2020/4/16/21223972/facebook-coronavirus-hoaxes-warning-misinformation-avaaz.

As for Google, the active engineering of search results for political ends has been a debated issue for several years.[29] But notwithstanding the actual scope or character of these kinds of influence, their "benevolent" implementation will normalize structures of intervention and institutionalize these new forms of governance. Interventions targeting fake news and perspectives associated with universally derided groups may well be especially prone to catalysing such a development due to the assent it can plausibly engender. For instance, popular opinion might be quite accepting of the suppression of political views associated with movements such as the radical right, whether or not their connection is accidental.

4 Reflections relating to possible solutions and mitigation strategies

So far, we can conclude that the presence and influence of AI also in private settings is getting more complex. This complexity should however not be understood as impossible to handle. Possible solutions need however to take into account that when it comes to the transnational nature of AI, several jurisdictions interplay and are applicable at the same time. This also means that different traditions, administrative settings, legal cultures and a variety of bodies need to interact with the legislative measures and forms chosen.

One possible approach towards addressing issues of influence and opacity, which at the same time circumvents some of this institutional complexity, could be that we embark upon a major quest to surveil the algorithmic impact of all these private digital media and platforms.[30] This would necessitate impartial surveillance bodies, using methods and tests for gathering information (anonymous of course) in order to achieve a transparent comparison of how these private systems generate different

[29] Cf. John D. McKinnon, Douglas MacMillan, "Google Workers Discussed Tweaking Search Function to Counter Travel Ban", *Wall Street Journal* 2018. Online: https://www.wsj.com/articles/google-workers-discussed-tweaking-search-function-to-counter-travel-ban-1537488472; Kirsten Grind, Sam Schechner et al., "How Google Interferes With Its Search Algorithms and Changes Your Results", *Wall Street Journal* 2019. Online: https://www.wsj.com/articles/how-google-interferes-with-its-search-algorithms-and-changes-your-results-11573823753.

[30] Cf. the Digital Services Act as an example. See also Bernard Rieder: https://policyreview.info/articles/analysis/towards-platform-observability.

political nudges depending upon the users' profiles and online behaviour. This suggestion is also in principle possible to realise using existing legislation, and does not necessitate any newer intrusive or controversial legal measures as long as it links to and respect the realisation of fundamental rights, such as the rights to privacy and/or data protection rules.[31] Its effectiveness, however, seems to be contingent upon the quality of a complex network of supervision, as well as of the effective broadcasting and political reception of this network's reviews and critique.

Another theoretically possible option could be to actually expropriate and make public all data the tech platforms gather and use in order to realise the different search results. In this way, the data resources could eventually become a type of universal big data in the form of a "public commons" accessible to all.[32] Accordingly, the monopolistic situation of a few multinational companies would be undermined and other private entities as well as states would have new opportunities to e.g. create their own search engines open to comparisons focusing on unwanted bias patterns. In this situation, open research could additionally in principle render transparent such "informational influence" that today is exerted with the use of proprietary data banks and digital platforms. A related option would be to create a parallel system of a big data commons that would not expropriate current private entities per se, but that over time would be able to mirror their information resources and thereby both challenge and scrutinize their political and economic influence.

Especially the second option is however not easily realised due to the pluralistic legal landscape of contemporary society. The two models need to be adapted to national and European law (both the European Union as well as the Council of Europe) and respect fundamental and human rights relating to amongst other things, the right to privacy and the right to property. This becomes more difficult as our discussion involves private

[31] Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation), the European Charter in Fundamental Rights, the Treaty on the European Union, the Treaty on the Functioning of the European Union as well as the relevant legal sources from the Council of Europe.

[32] See for example the proposed Data Governance Act: https://www.consilium.europa.eu/en/press/press-releases/2021/10/01/eu-looks-to-make-data-sharing-easier-council-agrees-position-on-data-governance-act/. See also http://infolegproject.net/call-for-papers-for-workshop-data-and-the-common/.

multinational companies, having rights in their own capacity according to both private and public law. Traces of this can also be seen in the work done by the European Commission. Moreover, the political influence of large multinational corporations would likely hamper any initiatives towards actual expropriation of what in many cases is their key resource and source of income. This could in turn trigger further consolidation from the corporate side, possibly exacerbating present issues of regulatory capture and corporate political influence, prompting these already very prominent private entities to strengthen their ties to legislative processes.

5 Legislative options – current initiatives

There are currently several approaches being discussed pertaining to handling and steering the development of AI both for private and public use. The legal initiatives taken come from several actors of different sorts. Private companies, often international and dominant, have tried to create their own communication strategy for legitimizing the use of algorithms, as we have seen above with the example of Facebook. There are also international organisations drafting soft law documents elaborated by states in dialogue with different expert groups. Online public consultation on these matters attires great attention.[33]

The latest initiative, however, is taken by the European Commission in its Proposal for an Artificial Intelligence Act (AI Act).[34] This proposal followed after intense discussions since the General Data Protection Regulation was enacted in 2016. In the EU, a high-level expert group on AI (HLEG), comprised of 52 experts was established and so was an AI Alliance with 4000 stakeholders, holding a yearly AI Assembly. The European Parliament and the Council explicitly requested a process of regulating AI in 2017 and in the political guidelines 2019–2024 entitled "A Union that strives for more", this was underlined.

[33] See for example the work with the AI Act within the European Union, Proposal for a Regulation laying down harmonised rules on artificial intelligence (artificial intelligence act) and amending certain union legislative acts (COM(2021) 206. See also how the process of enacting the Data Protection Regulation followed a similar procedure before the EU Commission started its work as stated in the EU Treaties. Reichel, Jane & Lind, Anna-Sara, Regulating Data Protection in the EU, I: Perspectives on Privacy, Dörr, Dieter & Weaver, Russell L. (ed.), de Gruyters publisher, 2014, pp. 22–45.

[34] Proposal for a Regulation laying down harmonised rules on artificial intelligence (artificial intelligence act) and amending certain union legislative acts (COM(2021) 206.

The two options suggested in the previous section would not have great success within the framework of the AI Act. In the Proposal, focus is on risk management and a risk-based assessment of the different AI systems is key. The stronger the risk, the more safeguards are demanded. Transparency becomes a core value, according to the Proposal, when we deal with a high-risk AI system. The AI Act is construed in such a way that it imposes legal rules for high risk AI systems while all those delivering AI that is not of high risk need to voluntarily obey to codes of conduct. Further clarification is accordingly needed as to how this should be handled. High risk is decided upon the level of negative effects a certain system has. If it is unacceptably high, the system is forbidden. Some high risk systems might be allowed, however, if there are public security aims legitimising the use of the system. In those circumstances, certain conditions must be met. For the high risk systems, documentation, data governance, transparency and information to users are core conditions that need to be fulfilled. It is also expressly stated that the system needs to have human oversight (Article 14 AI Act).

The AI Act explicitly states that it will not contravene the mechanisms put in place through the GDPR, nor consumer protection, non-discrimination and gender equality expressed in different legislative documents. The risk of overlapping legislative acts is clear. In the AI Act the European Commission tries to handle this through references of giving priority to the GDPR as soon as privacy issues relating to individuals are at stake.[35] In the new administrative structure suggested by the Commission, the European AI Board needs to consult and respect the opinion of the European Data Protection Board.

The complex issue of transparency, legal certainty and surveillance can be illustrated with the following. In the AI Act, market surveillance mechanisms are suggested. As the Act mostly has as its legal basis the internal market rules in the Treaty on the Functioning of the European Union (art. 114), it mainly focuses, as we have seen, on the realisation of free market rules and rights. The market surveillance authorities are suggested to be public bodies. The idea is to have a system of notifications in place so that users inform providers if risks have appeared or if some-

[35] See for example para. 23 in the preamble.

thing is not working properly.[36] The providers must inform the market surveillance authorities if their post-marketing monitoring identifies risks or cases of non-compliance. This mechanism is only designed for the AI systems showing to be of higher risk. Unfortunately, the individual is not given the opportunity to sue a provider or user for not respecting the AI Act. The right to lodge complaints for the individual has not been included in the AI Act and it does not contain any such right for groups either. These basic dimensions would have been a simple and obvious way to strengthen transparency and foster good development in future realisation of AI.

6 Concluding observations

Current and proposed legislative and regulatory approaches request a stringent assessment of the potential harm of AI systems, precautionary measures ensuring the transparency of "high risk systems", as well as effective human oversight. In light of the arguments and analyses presented in this chapter, two main questions follow in relation to these approaches.

1. How can a reasonable distinction of high vs. low-risk systems be made in relation the AI systems in question?

It seems difficult to designate any type of AI system as low-risk when even the most basic types of AI, such as simple filtering algorithms, can have complex and opaque effects in the sense that they have a certain unpredictability in terms of their actual consequences – especially when they are integrated into a fluid informational environment where large numbers of institutions, groups and individuals interact. Moreover, any self-learning algorithm whose operations may impact human individuals in a chaotic environment will by definition have uncertain and possibly far-reaching consequences and would seem to preclude anything akin to a low-risk assessment. To actually gauge the risk of such systems in any meaningful way seems to require an evaluation of their actual operations

[36] Regulation (EU) 2019/1020 of the European Parliament and of the Council of 20 June 2019 on market surveillance and compliance of products and amending Directive 2004/42/EC and Regulations (EC) No 765/2008 and (EU) No 305/2011.

and interaction with their intended context, i.e. much cannot really be said until after the fact of their implementation.

2. What types of precautionary measures can in practice ensure transparency within the current legal frameworks and tools of enforcement?

Ensuring transparency in the face of the difficulties we have discussed is a daunting task. Fines or similar sanctions are rarely effective measures against very large corporate entities, especially if they possess some level of influence over legislative bodies. Relying on voluntary obedience in relation to codes of conduct is also hardly a perdurable solution due to the nature of the modern corporation and the competitive environment in which it operates.

Our suggestions, which to some extent are amenable to the current legislative framework, promise to address these issues by building or promoting new structures for review and surveillance, where also in principle any potentially influential AI system can be placed under scrutiny, enabling a more comprehensive risk assessment before more intrusive measures are taken. Importantly, they at least open the door to meaningfully ask whether these incipient forms of technology and certain applications of them are really desirable from a cost-benefit perspective at an early stage before their full entrenchment in society. In relation to question 1 we would like to state that the AI Act does something resembling an contextual assessment by differentiating between low-risk and high-risk uses. That is a good thing, but it is a problematic contextual assessment in the sense that high-risk and low-risk is a problematic distinction. It is difficult, even impossible, to state that AI systems or application contexts are by definition low-risk. This trend to create "free-zones", where you only have voluntary regulation, might therefore not be desirable. At the same time, this is done in the context of the General Data Protection and in a broader sense one could see the GDPR as a sort of a regulatory safety net also for AI related legal issues. This is also underlined in the AI Act and in the Explanations presented by the commission, for example through the supremacy of the European Data Protection Board.

The AI Act also promotes new precautionary structures for review and surveillance of AI systems which leads us to our question 2. This is done as reliance on voluntary codes of conduct does not perdurable solution due to the nature of the modern corporation and the competitive environment in which it operates. Once again, this means that the regulatory free-zone for low-risk AI systems is not very desirable.

A final reflection in this chapter is a methodological one. As mentioned above in section 1, this text is a methodological encounter between two authors who have different disciplinary backgrounds, philosophy of religion and law respectively. What have been the gains? We believe that the different sources used in this text widen our perspectives and deepen our knowledge. It has also been rewarding to seek fruitful interactions between our different methodological approaches, which although related bring out quite different aspects of the problems under scrutiny. It should however be stressed that the authors do know each other professionally, as they are part of the same research group and have spent time talking to each other cross-disciplines before, which likely has expedited the process of interdisciplinary collaboration as there already is an established heuristic framework in place.

And what can we, respectively, take with us in the process of continued work on AI? Law as it is communicated and construed in different jurisdictions, and at different levels, has become more fragmented and pluralistic than was the case before. This has led to a stronger presence of the constitutional dimensions of law, that in turn embrace the dimensions of power and legitimacy relevant at all levels and a common denominator of all jurisdictions. This is true also for European law and the interplay between European, international and national law. To understand and explain these mechanisms and phenomena in relation to AI is however not the sole task of the legal domain. Law needs to interact and communicate with society in many ways and is mirrored in the changes that occur in society. To explain, visualise and test future challenges is however an endeavour that in itself must by definition be multidisciplinary. This deliberation must invite all parts of civil society as far as possible, in the interest of consolidating the foundations of liberal democracy in this transformative period characterised by complex and diverse challenges thereof.

Bibliography

Preparatory works, Sweden

SOU 2014:75 Automatiserade beslut – färre regler ger tydligare reglering

Riksrevisionen, Automatiserat beslutsfattande i statsförvaltningen – effektivt, men kontroll och uppföljning brister (RiR 2020:22)

European Union

Proposal for a Regulation laying down harmonised rules on artificial intelligence (artificial intelligence act) and amending certain union legislative acts (COM(2021) 206

Literature

Axberger, Hans-Gunnar, Constitutional Responsibility for the Free Flow of Information and Ideas in the Internet Age, in: Lind, Reichel & Österdahl (Eds.), Information and Law in Transition – Freedom of Speech, the Internet, Privacy and Democracy in the 21st Century, Martinus Nijhoff Publicher and Liber, 2015, pp. 43–54

Barnett, Michael, Duvall, Raymond, *Power in Global Governance*, New York: Cambridge University Press 2005.

Brown, Alexander, *Hate Speech Law: A Philosophical Investigation*, New York: Routledge 2015.

Brudholm, Thomas, Johansen, Schepelern Brigitte (eds.) *Hate Politics Law*, New York: OUP 2018.

Fraleigh, D.M. and Tuman, J.S., *Freedom of Expression in the Market-place of Ideas*, Thousand Oaks, CA: Sage 2011.

Glasser, I. "Introduction", in H.L. Gates Jr. et al. (eds.) *Speaking of Race, Speaking of Sex: Hate Speech, Civil Rights, and Civil Liberties*, New York: New York University Press 1994

Hirschfeldt, Johan, "Free access to public documents – a heritage from 1766", in Lind, Reichel and Österdahl, (eds.) *Transparency in the future – Swedish Openness 250 years*, Ragulka Press, 2017.

Lind, Anna-Sara, Sweden: "Free press as a first fundamental right", in Suksi, M, et al. (Eds.) *First fundamental rights documents in Europe*, Intersentia, 2015

Lind, Anna-Sara, "Freedom of the Press Act – from then to now" in Lind, Reichel and Österdahl, (eds.) *Transparency in the future – Swedish Openness 250 years*, Ragulka Press, 2017.

Lind, Anna-Sara, "Den offentliga rätten i mångvetenskaplig forskning", i Arvidsson, et al. (eds.), *Festskrift till Wiweka Warnling Conradsson*, Jure förlag 2019.

Rodan, Garry, *Transparency and Authoritarian Rule in Southeast Asia*, London: Routledge 2004.

Thaler, Richard, *Nudge: Improving Decisions About Health, Wealth and Happiness*, New Haven: Yale University Press 2008.

Vasquez, M. and de las Fuentes, C., "Hate Speech or Freedom of Expression? Balancing Autonomy and Feminist Ethics in a Pluralistic Society", in M. Brabeck (ed.) *Practicing Feminist Ethics in Psychology*. Washington: American Psychological Association 2000.

Woolley, Samuel C., Howard, Philip N. (eds.), *Computational Propaganda*, New York: OUP 2019.

Other sources

Angwin, Julia, Larson, Jeff, Mattu, Surya & Kirchner, Lauren, "Machine Bias", *Propublica* 2016. Online: https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

Agudo, Ujué, Matute, Helena, "The influence of algorithms on political and dating decisions". *PLoS ONE* 16(4), 2021. Online: https://doi.org/10.1371/journal.pone.0249454

Assimakoupoulos, Stavros, Baider, Fabienne H., Millar, Sharon, *Online Hate Speech in the European Union*, Cham: Springer 2017.

Avaaz, "White Paper: Correcting the Record", *Avaaz.com*, 2021. Online: https://secure.avaaz.org/campaign/en/correct_the_record_study/

BBC, "Facebook tests extremist content warning messages", *BBC News* 2021. Online: https://www.bbc.com/news/technology-57697779

Dagens Nyheter, "Myndighetsbeslut måste alltid vara rättsäkra". Online: https://www.dn.se/ledare/myndighetsbeslut-maste-alltid-vara-rattssakra-oavsett-om-de-fattas-av-manniskor-eller-maskiner/, *Dagens Nyheter* 2021.

Entman, Robert M., Usher, Nikki, "Framing in a Fractured Democracy: Impacts of Digital Technology on Ideology, Power and Cascading Network Activation", *Communication Theory* vol. 68, no. 2 2018.

Ghafary, Shin, "Facebook will start nudging users who have 'liked' coronavirus hoaxes", *Recode* 2020. Online: https://www.vox.com/recode/2020/4/16/21223972/facebook-coronavirus-hoaxes-warning-misinformation-avaaz

Grind, Kirsten, Schechner, Sam et al., "How Google Interferes With Its Search Algorithms and Changes Your Results", *Wall Street Journal* 2019. Online: https://www.wsj.com/articles/how-google-interferes-with-its-search-algorithms-and-changes-your-results-11573823753

Internet Live Stats 2021. Online: https://www.internetlivestats.com/google-search-statistics/

Kim, Young Mie, "Algorithmic Opportunity: Digital Advertising and Inequality in Political Involvement", *The Forum*, 14 (4), 2016.

MacMillan, Douglas, McKinnon, John D., "Google Workers Discussed Tweaking Search Function to Counter Travel Ban", *Wall Street Journal* 2018. Online: https://www.wsj.com/articles/google-workers-discussed-tweaking-search-function-to-counter-travel-ban-1537488472

Osman, Nawab Burke, Adam, "New Resources to counter Hate and Extremism Online". *Facebook* 2021. Online: https://about.fb.com/news/2021/06/new-resources-to-counter-hate-and-extremism-online/

Perra, Nicola, Rocha, Luis E. C., "Modelling opinion dynamics in the age of algorithmic personalization", *Scientific Reports* 9, 7261, 2019. Online: https://doi.org/10.1038/s41598-019-43830-2

Schropfer, Mike, "Update on Our Progress on AI and Hate Speech Detection", *Facebook* 2021. Online: https://about.fb.com/news/2021/02/update-on-our-progress-on-ai-and-hate-speech-detection/

Slapakova, Linda, "Towards an AI-based Counter-Disinformation Framework", *The Hague Diplomacy Blog* 2021. Online: https://www.universiteitleiden.nl/hjd/news/2021/blog-post---towards-an-ai-based-counter-disinformation-framework